Classification multi-étiquettes pour l'alignement multiple de séquences protéiques

Lina Fahed*, Gabriel Frey*, Julie Dawn Thompson**, Nicolas Lachiche*

*LSIIT, Université de Strasbourg, Pôle API, Bd Brant, 67400 Illkirch
{linafahed@gmail.com, g.frey@unistra.fr, nicolas.lachiche@unistra.fr}
**IGBMC, Université de Strasbourg, 1 rue Laurent Fries, 67404 Illkirch
{julie@igbmc.fr}

Résumé. Cet article présente une application de classification multi-étiquettes permettant de déterminer le programme à utiliser pour construire un alignement multiple d'un ensemble de séquences protéiques donné. Dans un premier temps, nous avons réussi à améliorer le système existant, Alexsys en ajoutant des attributs. Dans un second temps, nous déterminons pour un ensemble de séquences protéiques donné le ou les aligneurs capable de produire les alignements de meilleur score, à epsilon près. Les mesures de performances propres à la classification multi-étiquette nous permettent d'analyser l'influence de epsilon et de choisir une valeur assez petite pour distinguer les meilleurs aligneurs des autres.

1 Introduction

De nombreux programmes de construction d'alignements multiples à partir d'un ensemble de séquences protéiques (appelés aligneurs) ont été développés. Cependant, il n'existe pas un aligneur capable de bien aligner tous les types de séquences. C'est la raison pour laquelle le laboratoire de bioinformatique de l'IGBMC de Strasbourg a développé un système expert pour l'alignement multiple de séquences protéiques appelé Alexsys (Aniba et al., 2009). Le système Alexsys propose d'identifier, pour un ensemble de séquences protéiques données, les aligneurs permettant d'obtenir de bons alignements en se basant sur des techniques de fouilles de données (classification supervisée) exploitant des caractéristiques des séquences.

2 Le contexte

Un grand nombre de travaux en apprentissage supervisé concerne la prédiction d'une étiquette unique, où une seule étiquette parmi plusieurs étiquettes disjointes disponibles est associée à chaque instance. Cependant, il existe des applications dans lesquelles un ensemble d'étiquettes est associé à chaque instance. Ces données sont appelées données multi-étiquettes. La classification multi-étiquettes concerne l'apprentissage d'une fonction qui associe à chaque instance le sous-ensemble des étiquettes approprié. C'est une forme d'apprentissage supervisé où plusieurs classes booléennes sont associées à chaque instance. Des mesures de performance

spécifiques ont été définies (Tsoumakas et al., 2010; Fürnkranz et al., 2008). Nous les utiliserons dans le cadre de l'alignement multiple de séquences protéiques.

La séquence protéique est une suite d'acides aminés (aussi appelés résidus). Chaque acide aminé est représenté par une lettre. Pendant l'évolution des espèces, des changements appelés mutations peuvent se produire. En comparant les séquences entre elles et en cherchant les résidus ou les suites de résidus qui sont conservés dans une même famille de protéines, nous pouvons beaucoup apprendre sur les résidus essentiels pour certaines fonctions Les alignements multiples de séquences sont un outil important de la biologie moderne.

Le nombre de programmes d'alignement multiple disponible est sans-cesse croissant. Cependant, aucun programme actuel n'est capable de construire un alignement multiple de haute qualité pour tous les cas possibles. Les principaux critères utilisés pour décider du programme d'alignement à utiliser sont : la qualité de l'alignement, le temps d'exécution et l'utilisation de mémoire (Thompson et al., 2011). La qualité de l'alignement est généralement le critère le plus important. Dans notre étude, nous avons utilisé les aligneurs suivants, connus pour leurs performances et leurs fiabilités : ClustalW , Dialign, Mafft, Muscle, Probcons, T-coffee et Kalign. Une description des programmes d'alignement ainsi qu'un résumé de leurs avantages et de leurs inconvénients est présenté dans (Edgar et Batzoglou, 2006).

L'estimation de la qualité de l'alignement est un point critique. La fonction de score la plus utilisée est le " Sum of Pairs ". Actuellement, la qualité d'un algorithme est généralement estimée en comparant les résultats obtenus avec des alignements de référence prédéfinis (Benchmarks). Par conséquent, les données de référence doivent être de haute qualité. Un des premiers benchmarks à grande échelle construit pour l'alignement multiple des séquences se nomme BaliBase (Bahr et al., 2001). Les séquences utilisées dans la base de données de BaliBase ont toute une structure 3D connue. Les alignements sont construits à partir de ces structures 3D et raffinées à la main par des experts pour garantir l'alignement correct des résidus conservés. Les alignements sont organisés sous la forme de plusieurs ensembles de références, représentant des problématiques réelles de l'alignement multiple. L'autre benchmark utilisé lors de cette étude, OXBench (Raghava et al., 2003), contient des alignements multiples de séquences protéiques construits automatiquement par des aligneurs.

3 Amélioration de l'existant : Alexsys

Les divers algorithmes d'alignement multiple peuvent produire des alignements différents pour un même ensemble de séquences à aligner. Les qualités de ces différents alignements dépendent des caractéristiques des séquences à aligner. Les biologistes aimeraient expliciter les relations entre les caractériques des séquences protéiques à aligner et la force/faiblesse de différents algorithmes d'alignement, afin de pouvoir choisir le meilleur aligneur dans chaque cas. C'est la raison pour laquelle le système Alexsys "Alignment Expert System" a été développé (Aniba et al., 2009). Dans cette section, nous décrivons l'existant puis les améliorations apportées.

3.1 Alexsys : un système expert pour l'alignement multiple

Dans Alexsys, les connaissances ou règles du système expert ne sont pas recueillies auprès des experts, mais obtenues par apprentissage artificiel à partir de jeux de données. Les données

d'apprentissage proviennent de deux jeux de données de référence (BaliBase (Bahr et al., 2001) et OxBench (Raghava et al., 2003)) qui représentent des problèmes réels et des cas difficiles d'alignement multiples. Les données consistent en un ensemble de 890 alignements choisis par les biologistes. Les attributs utilisés par Alexsys pour représenter un ensemble de séquences à aligner appartiennent à quatre catégories : physiques, structurels, fonctionnels et physico-chimiques.

La classe de chaque aligneur est définie comme "fort" ou "faible" sur la base du score défini comme le nombre de paires de résidus alignés de la même façon dans l'alignement produit par l'aligneur et dans l'alignement de référence. Au-dessus d'un seuil de 0,5, un aligneur est considéré comme "fort", et en-dessous de cette valeur, un aligneur est considéré comme "faible". Cette valeur de seuil, bien que choisie par les biologistes, reste arbitraire. Alexsys construit alors un modèle par aligneur sous la forme d'une forêt aléatoire.

3.2 Améliorations

Nous avons poursuivi les travaux d'Alexys. Tout d'abord, nous avons considéré une version enrichie de la base de données contenant 1058 instances et gardant les même attributs. L'attribut de pourcentage d'identité (PCID) est un des attributs les plus importants utilisés par Alexsys. Nous avons défini une autre méthode pour le calculer. Cette méthode se base sur l'algorithme d'alignement global Needleman-Wunsch (Needleman et Wunsch, 1970) dans lequel l'identité entre deux séquences est calculée pour chaque couple de séquences de l'ensemble de séquences. Puis le maximum, le minimum, la moyenne et l'écart type d'identités obtenues sont calculés. Ainsi les valeurs de quatre attributs liés au PCID sont générées pour chaque ensemble de séquences utilisé dans la base de données.

La génération de l'attribut de PCID consomme beaucoup de temps et de mémoire Nous avons défini et évalué une autre méthode moins exigeante en ressources nous permettant d'estimer la similarité entre les séquences. C'est la méthode de comptage des n-gram. Nous avons créé plusieurs attributs de n-gram en changeant la longueur des mots d'acides aminés. Les attributs 1-gram, 2-gram, 3-gram, et 4-gram ont été testés.

Le taux de bonne prédiction macro est évalué en utilisant le PCID ou le 1-gram séparément, ou les deux conjointement, en plus des attributs usuels d'Alexsys. Nous ne considérons pas seulement le seuil de 0,5 pour étiqueter les bons et mauvais aligneurs, mais nous avons essayé de faire varier le seuil de score. Plusieurs seuils ont été testés de 0 à 1, avec un pas de 0,1. L'évaluation est faite en moyennant 5 répétitions d'une validation croisée en 5 plis. Nous utilisons les forêts aléatoires implémentées dans Weka (Witten et al., 2011), comme dans la version originale d'Alexsys. En comparant le comportement des nouveaux attributs (Figure 1), nous remarquons qu'en utilisant les attributs de 1-gram avec l'attribut de PCID en même temps, le taux de bonne prédiction macro est meilleur qu'en cas d'utilisation de l'attribut de PCID uniquement, et bien sûr en utilisant toujours tous les autres attributs préalablement définis. Par ailleurs, les résultats ne sont pas présentés ici, mais nous avons observé que les attributs de 1-gram obtiennent de meilleurs performances que ceux de 2-gram, 3-gram et 4-gram. Dans le cas du seuil à 0,5, 1-gram et PCID ensemble donnent un taux de bonne prédiction macro de 90% pour les modèles créés pour les sept aligneurs.

4 Choix du meilleur aligneur

Pour chaque aligneur, nous lui attribuons l'étiquette "meilleur" si son score d'alignement est le score maximum à epsilon près. Nous autorisons une marge de epsilon car il peut y avoir plusieurs alignements intéressants pour les biologistes Un objectif de ce travail est d'identifier un seuil raisonnable pour epsilon. Puisque les scores varient de 0 à 1, il est clair qu'une valeur de 1 pour epsilon conduirait à considérer tous les aligneurs comme "meilleurs". Par la suite, lors des différents test, epsilon variera entre 0 à 0,2.

Notre problème consiste à trouver les aligneurs étiquetés "meilleur", autrement dit l'ensemble des aligneurs capables de produire des alignements multiples de score maximum ou presque. C'est de la classification multi-étiquettes. En calculant la cardinalité, qui est le nombre moyen des étiquettes "meilleur" sur tous les exemples (Figure 2), nous remarquons que deux aligneurs (en moyenne) sont "meilleurs" pour epsilon égal à zéro (où seulement le score maximum est choisi pour être "meilleur"). Ce résultat est surprenant. Nous nous attendions à avoir seulement un meilleur aligneur pour epsilon égal à 0. En vérifiant les données en cas de epsilon égal à 0, nous remarquons que 35% des exemples ont plus que deux aligneurs acceptables comme "meilleurs". La figure 2 montre aussi la vitesse à laquelle le nombre d'aligneur acceptables augmente en fonction de epsilon. Le nombre d'aligneurs acceptables tend vers 7 pour les grandes valeurs de epsilon.

L'implémentation de Weka des forêts aléatoires est encore utilisée, en faisant 5 validations croisées en 5-plis. Nous calculons la moyenne et l'écart-type, sur ces 5 itérations de la validation croisée, des différentes mesures de performance de la classification multi-étiquettes afin de déterminer la valeur de epsilon la plus pertinente.

En observant les mesures de performance basées sur l'exemple en fonction de la cardinalité (Figure 3), nous remarquons que beaucoup de mesures n'apportent pas d'information autre que : en augmentant epsilon donc la cardinalité, c'est-à-dire quand plus d'aligneurs deviennent acceptables, ces mesures augmentent. En particulier, la précision et le rappel, donc la mesure-F1, et le taux de bonnes prédictions et le rappel pour la classe négative croissent quasi-linéairement avec la cardinalité. Le taux de bonne prédiction de l'ensemble et 1-hammingLoss croissent également mais avec une pente moins régulière. Seule l'AUC a un comportement différent et atteint rapidement un maximum, reste constante puis diminue quand epsilon et la cardinalité augmentent.

Nous observons un comportement similaire des mesures de performance basées sur l'étiquette, micro et macro, en fonction de epsilon à la Figure 4. De plus, epsilon=0,01 apparaît comme un point singulier. Tout d'abord, c'est le maximum pour la macro AUC. Rappelons que nous cherchons le plus petit epsilon (et cardinalité) ayant de "bonnes" performances pour prédire le (ou les) meilleur aligneur. À cette valeur 0,01 de epsilon, le taux de prédiction de l'ensemble est à 25%, c'est-à-dire qu'une fois sur quatre on obtient le sous-ensemble des bons aligneurs et seulement eux, parmi $2^7 = 128$ sous-ensembles possibles. Enfin, la plupart des mesures macro et micro sont à 0,75, en particulier 3 fois sur 4 pour la précision et pour le rappel, c'est-à-dire que 3/4 des aligneurs prédits acceptables font réellement parti des meilleurs et que 3/4 des meilleurs aligneurs sont proposés.

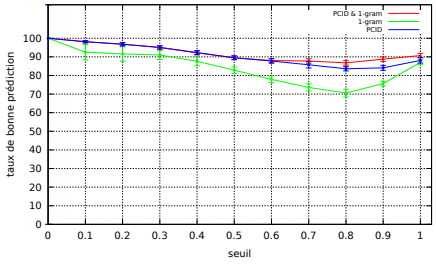

FIG. 1 – *Comparaison 1-gram et PCID pour plusieurs seuils de score.*

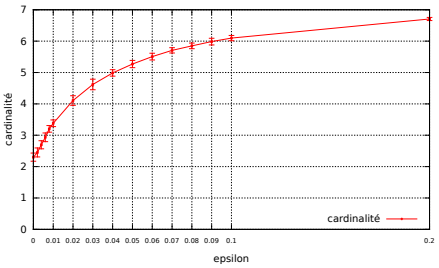

FIG. 2 – *Cardinalité (moyenne et écart-type) en fonction de epsilon.*

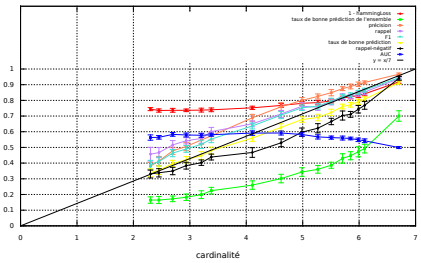

FIG. 3 – *Mesures de performance basée sur l'exemple en fonction de la cardinalité.*

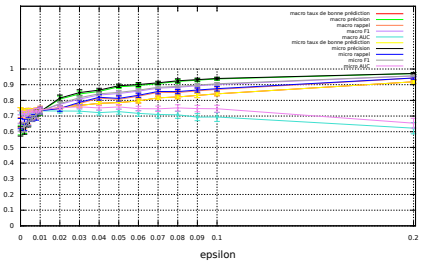

FIG. 4 – *Mesures de performance en fonction de epsilon.*

5 Conclusion

Le système Alexsys (Aniba et al., 2009) sert à prédire le programme à utiliser pour construire un bon alignement multiple d'un ensemble de séquences protéiques. Nous avons amélioré l'approche d'Alexsys en reprenant le calcul du PCID et en lui ajoutant des attributs à partir du 1-gram plus rapides à calculer et complémentaires, c'est-à-dire que la combinaison de ces attributs et du PCID améliore le taux de bonne prédiction par rapport au PCID utilisé seul dans Alexsys.

La seconde contribution de ce travail a consisté à reformuler le problème en considérant qu'un aligneur est acceptable si son score est maximum à epsilon près. L'analyse des différentes mesures de performance spécifiques à la classification multi-étiquette a été nécessaire pour déterminer la plus petite valeur de epsilon adéquate. En effet la plupart des mesures augmentent quand epsilon augmente et que tous les aligneurs deviennent acceptables. Cependant l'AUC nous a permis d'identifier un premier pic pour epsilon égal à 0,01. À cette valeur, le taux de bonne prédiction de l'ensemble, c'est-à-dire la prédiction d'exactement tous les aligneurs acceptables et seulement eux, est de 25%, c'est-à-dire nettement au dessus de l'aléatoire à $1/128 = 0,8\%$. Cette mesure est particulièrement exigeante. La plupart des autres mesures de performance sont à 75%, en particulier la précision et le rappel micro et macro.

Les perspectives de ce travail sont nombreuses. Une première perspective concerne l'apprentissage de tri (ranking) multi-étiquettes. Une autre perspective consiste à tenir compte de

la qualité des données, car les informations sur les séquences sont très bruitées. Une dernière perspective envisageable est l'utilisation algorithmes dédiés à la prédiction d'évènements rares ou de classes très déséquilibrées, les problèmes considérés en bioinformatique concernant souvent des cas rares.

Références

Aniba, M. R., S. Siguenza, A. Friedrich, F. Plewniak, O. Poch, A. Marchler-Bauer, et J. D. Thompson (2009). Knowledge-based expert systems and a proof-of-concept case study for multiple sequence alignment construction and analysis. *Briefings in Bioinformatics 10*(1), 11–23.

Bahr, A., J. D. Thompson, J.-C. Thierry, et O. Poch (2001). Balibase (benchmark alignment database) : enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research 29*(1), 323–326.

Edgar, R. C. et S. Batzoglou (2006). Multiple sequence alignement. *Current Opinion In Structural Biology 16*, 368–373.

Fürnkranz, J., E. Hüllermeier, E. L. Mencía, et K. Brinker (2008). Multilabel classification via calibrated label ranking. *Machine Learning 73*(2), 133–153.

Needleman, S. B. et C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*(3), 443 – 453.

Raghava, G. P. S., S. M. J. Searle, P. C. Audley, J. D. Barber, et G. J. Barton (2003). Oxbench : A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics 4*, 47.

Thompson, J. D., B. Linard, O. Lecompte, et O. Poch (2011). A comprehensive benchmark study of multiple sequence alignment methods : Current challenges and future perspectives. *PLoS ONE 6*(3), e18093.

Tsoumakas, G., I. Katakis, et I. P. Vlahavas (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pp. 667–685. Library of Congress Control Number.

Witten, I. H., E. Frank, et M. A. Hall (2011). *Data Mining : Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann.

Summary

This article introduces the use of multi-label classification to find the best aligner to use to build a multiple alignment of given protein sequences. First, we have succeeded in improving an existing work, Alexsys considering supplementary attributes. Secondly, we predict which aligners are the best to align a given set of sequences, that is to say the aligners that produce alignments whose scores are less than epsilon away from the maximum score recorded among all the aligners considered on this study. Performance measures specific to multi-label classification allow to study the influence of epsilon and to choose a value small enough to distinguish the best aligners from the others.