Agrégation des Propriétés Physico-Chimiques des Acides Aminés

Jean-Luc Falcone
Département d'Informatique
Université de Genève, Suisse
jean-luc.falcone@cui.unige.ch

Paul Albuquerque
Institut d'Ingénierie en Informatique
Ecole d'Ingénieurs de Genève, Suisse
albuquer@eig.unige.ch

Résumé

Les données issues du séquençage des organismes nécessitent la mise au point de méthodes de prédiction de la fonction et de la structure des protéines. Pour que ces méthodes soient efficaces, elles doivent se baser sur les propriétés physico-chimiques des 20 acides aminés (masse, charge, ...). On en recense actuellement 484. Or beaucoup de ces propriétés sont très fortement corrélées. Pour les regrouper, nous proposons une métrique issue de la corrélation. Nous avons classifié les propriétés physico-chimiques selon cette distance à l'aide des méthodes k-means et global k-means. Un indice de qualité a permis de déterminer des nombres "naturels" de groupes. Les résultats des agrégations sont conformes aux attentes biochimiques. Les centres des groupes ainsi produits constituent un nouvel ensemble de propriétés.

Mots clés : *Agrégation de données, mesure de similarité, corrélation, bioinformatique, acides aminés.*

1. INTRODUCTION

Les protéines sont les constituants essentiels des cellules vivantes. Ces macro-molécules sont les unités fonctionnelles contrôlant et catalysant les réactions chimiques de la vie. Elles sont composées de chaînes linéaires, repliées dans l'espace, de 20 molécules de base : les acides aminés. L'essentiel de la biochimie et de la biologie moléculaire consiste à analyser ces protéines et comprendre leur rôle.

L'engouement récent pour le séquençage des organismes produit une grande quantité de séquences de protéines inconnues. Ceci entraîne un besoin pour des méthodes de prédiction de leur fonction, structure tridimensionnelle et localisation cellulaire. La plupart de ces méthodes dérivent d'algorithmes de classification et d'agrégation de données (*data clustering*), comme les réseaux de neurones ou les arbres de décisions. Une des difficultés réside dans le codage des entrées. Les méthodes habituelles représentent chaque acide aminé par un symbole. Malheureusement, elles ne tiennent pas compte des affinités entre les différents acides aminés. Or, les acides aminés ont de nombreuses caractéristiques physico-chimiques (comme la masse, la charge, l'hydrophobicité, etc.) qui font que pour une propriété donnée certains de ces acides aminés sont plus proches que d'autres. De plus, les machineries protéiques sont sensibles aux interactions physico-chimiques et non à des symboles abstraits.

Il serait intéressant de représenter chaque acide aminé par un vecteur formé de ses propriétés, actuellement au nombre de 484 [1]. Ce nombre élevé, impliquant une charge de calcul rédhibitoire, rend nécessaire un regroupement des propriétés proches. Cette réduction est justifiée par le fait qu'il est difficile d'imaginer l'existence de 484 facteurs physico-chimiques différents. La multitude observée serait due aux nombeuses manières de mesurer expérimentalement ces propriétés.

Dans cet article, nous regroupons les propriétés physico-chimiques des acides aminés. Celles-ci sont répertoriées dans la base de données AAindex v6.0, disponible en ligne [2]. Chaque propriété y est définie par 20 valeurs correspondant aux 20 acides aminés. Pour mettre en évidence la proximité entre les propriétés, les auteurs ont représenté leurs données sous la forme d'un *minimal spanning tree* construit à partir des corrélations. Ils observent l'apparition de plusieurs groupes naturels. Leurs résultats confirment les redondances entre les propriétés.

2. K-MEANS ET GLOBAL K-MEANS

Pour agréger ces données, nous avons utilisé les algorithmes k-means et global k-means. Ces deux méthodes

CCECE 2004 – CCGEI 2004, Niagara Falls, May/mai 2004
0-7803-8253-6/04/$17.00

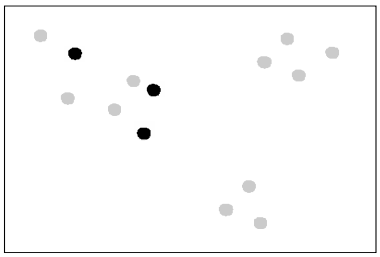

FIG. 1 – Tirage au sort des centres initiaux. Les K centres sont en gras.

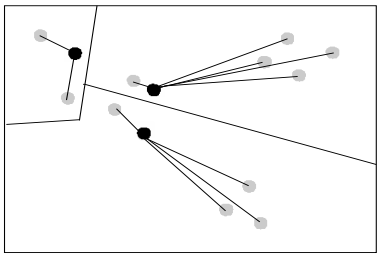

FIG. 2 – Partitionnement de l'espace en fonction des K-centres. Chaque point de donnée représenté est relié au centre le plus proche.

nécessitent une métrique à laquelle une notion de centre de masse est associée. Elles imposent de fixer le nombre K de groupes voulus.

L'algorithme k-means [3] est décrit ci-après.

0. On tire au sort K points de donnée comme centres initiaux. Leurs valeurs sont stockées dans une liste de centres (cf. fig. 1).

1. On partitionne l'espace en K groupes en associant chaque point de donnée au centre le plus proche (cf. fig 2).

2. On calcule le centre de masse de chaque partition. Cette nouvelle liste de centres remplace la précédente (cf. fig 3).

3. On répète l'algorithme depuis le point **1** jusqu'à ce qu'un critère d'arrêt soit rencontré.

Le critère d'arrêt que nous avons employé est la stabilisation de l'erreur d'agrégation, définie par l'éq. (1).

Cette méthode est cependant sensible au choix initial des centres. La méthode global k-means résout ce problème par une heuristique et rend k-means déterministe [4]. Le global k-means avec K groupes se déroule comme suit :

– On recherche le centre de masse de notre ensemble initial de données. Ce centre est le résultat du global k-means avec $k = 1$.

– On construit les global k-means successifs en incrémentant k jusqu'à atteindre $k = K$. A chaque étape $k = n$, on effectue un k-means en choisissant

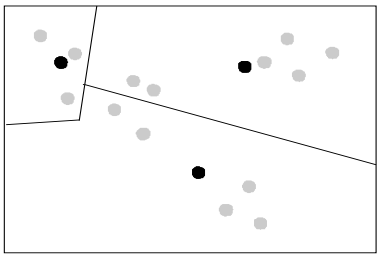

FIG. 3 – Les K centres de masse sont recalculés pour chaque partition.

n centres de la manière suivante :

– On initialise $n - 1$ centres avec les résultats du global k-means avec $k = n - 1$.

– Le dernier centre est choisi parmi les points de donnée en les testant tous successivement. On accepte celui dont l'erreur d'agrégation est la plus faible.

L'algorithme global k-means nécessite donc l'exécution de $(K - 1)N$ k-means, où N est le nombre de points de donnée à agréger.

L'erreur d'agrégation (*clustering error*) est donnée par

$$E(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K) = \sum_{j=1}^{N} \sum_{l=1}^{K} I_{\{\boldsymbol{x}_j \in C_l\}} \left(d(\boldsymbol{x}_j, \boldsymbol{g}_l)\right)^2 \qquad (1)$$

où $\boldsymbol{g}_l$ désigne le l-ème centre, $I_{\{\boldsymbol{x}_j \in C_l\}}$ est une fonction qui retourne 1 si $\boldsymbol{x}_j$ appartient au groupe l et 0 sinon, et $d(\boldsymbol{x}_j, \boldsymbol{g}_l)$ est la distance entre $\boldsymbol{x}_j$ et $\boldsymbol{g}_l$.

3. DISTANCE BASEE SUR LA CORRELATION

Les deux algorithmes décrits plus haut nécessitent une distance et une notion de centre de masse. Comme mentionné précédemment, la corrélation est une bonne mesure de similarité entre deux propriétés physico-chimiques. On associe à chaque propriété un vecteur dont les 20 composantes sont les valeurs de celle-ci pour chaque acide aminé.

Soit $\boldsymbol{x} = (x_1, \ldots, \boldsymbol{x}_D)$ un tel vecteur dans un espace de dimension $D = 20$. On dénote par $\bar{x}$ la moyenne des composantes de $\boldsymbol{x}$ et σ_x leur écart type. On pose encore $\bar{\boldsymbol{x}} = (\bar{x}, \ldots, \bar{x})$. On centre et réduit les variables via la réduction d'échelle

$$\boldsymbol{x}^* = \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{\sigma_x}$$

afin de rendre les données comparables. Cette transformation implique que nos points de donnée sont pro-

jetés sur une hypersphère à D dimensions de rayon $\sqrt{D}$ centrée à l'origine.

La corrélation entre deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ à D composantes est définie par

$$corr(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{D}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{D}(x_i-\bar{x})^2\sum_{i=1}^{D}(y_i-\bar{y})^2}}$$

ce qui peut aussi s'exprimer

$$corr(\boldsymbol{x},\boldsymbol{y}) = \frac{(\boldsymbol{x}-\bar{\boldsymbol{x}})\cdot(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|\,\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|} = \frac{1}{D}(\boldsymbol{x}^*\cdot\boldsymbol{y}^*)$$

où $\cdot$ désigne le produit scalaire et $\|\ \|$ la norme.

On propose la distance

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{1-(corr(\boldsymbol{x},\boldsymbol{y}))^2} \qquad (2)$$

laquelle implique que la distance entre deux points de donnée est nulle si et seulement si $y_i = ax_i + b$, où $a, b \in \mathbb{R}, a \neq 0$.

Le centre de masse $\boldsymbol{g}$ sur la sphère de rayon $\sqrt{D}$ d'un ensemble de point $C = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$, par rapport à cette distance (éq. 2), est le point qui minimise la fonction

$$F_C(\boldsymbol{g}) = \frac{1}{N}\sum_{j}^{N}(d(\boldsymbol{x}_j,\boldsymbol{g}))^2 \qquad (3)$$

sous la contrainte

$$H(\boldsymbol{g}) = 1 - \frac{1}{D}\boldsymbol{g}\cdot\boldsymbol{g} = 0 \qquad (4)$$

En utilisant la méthode des multiplicateurs de Lagrange, on se ramène à chercher le minimum parmi les solutions de

$$\nabla F_C(\boldsymbol{g}) = \lambda\nabla G(\boldsymbol{g})$$

Ceci correspond en fait à calculer les vecteurs propres et valeurs propres de la matrice symétrique $\boldsymbol{M} = (m_{pq})$ de dimension $D \times D$ définie par

$$m_{pq} = \frac{1}{ND^2}\sum_{j=1}^{N} x_{jp}^* x_{jq}^* \quad (p,q = 1,\ldots,D)$$

Le centre de masse est le vecteur propre, normalisé selon la contrainte H (éq. 4), pour lequel la fonction F_C (éq. 3) est minimale. Le détail des calculs est accessible dans [5].

4. RESULTATS

Le nombre d'itérations nécessaires à la convergence de l'algorithme k-means sur nos données avec la distance définie précédemment est faible (cf. tab. 1).

Les inconvénients de la méthode k-means résident d'une part dans le choix du nombre de groupes voulus, d'autre part dans le choix des centres initiaux. Si le deuxième problème est résolu en partie avec l'algorithme global k-means, la question du choix du nombre "naturel" de groupes est habituellement adressée par le biais d'indices de qualité. Parmi les nombreux indices existants [6], nous avons utilisé l'indice SD, adapté à notre contexte. Il est composé (i) d'une valeur *scat* reflétant la dispersion moyenne d'un ensemble de groupes $\{C_1,\ldots C_K\}$,

K	5	10	20	40
moyenne	12.3	13.7	11.2	9.1
écart-type	5.5	4.9	3.4	2.3

TAB. 1 – Nombres moyens d'itérations et écart-type pour la convergence de k-means à K centres, effectués sur 100 mesures.

$$scat(K) = \frac{1}{K}\sum_{i=1}^{K}\frac{F_{C_i}(g_i)}{F_C(g)}$$

où $C = C_1 \cup \cdots \cup C_K$, g_i étant le centre de masse de C_i et g le centre de masse de C ; (ii) d'une valeur *disc* reflétant la séparation entre les centres,

$$disc(K) = \frac{D_{\max}}{D_{\min}}\sum_{i=1}^{K}\left(\sum_{j=1}^{K} d(g_i,g_j)\right)^{-1}$$

où $D_{\max} = \max\{d(g_i,g_j)\,|\,i,j = 1,\ldots,K\}$ et $D_{\min} = \min\{d(g_i,g_j)\,|\,i,j = 1,\ldots,K\,,\ i \neq j\}$. On considère l'indice de qualité défini par

$$SD(K) = disc(K_{\max})\,scat(K) + disc(K) \qquad (5)$$

où $K_{\max}$ est le nombre maximum de groupes considérés. Les groupements "naturels" de données sont à chercher parmi les minima de cette fonction. Nos mesures indiquent que les nombres de groupes "naturels" sont $8, 12, 17$ et 28 (cf. fig. 4). Les résultats de ces groupements sont disponibles en ligne [7].

Nous donnons une interprétation biochimique succinte des groupes issus de l'agrégation avec $k = 17$ (cf. tab. 2). La signification de certains groupes n'a pas pu être établie avec certitude. Cependant, les propriétés dont on attend qu'elles soient regroupées, appartiennent effectivement au même groupe. Citons comme exemples le volume des résidus et leur masse moléculaire (groupe no. 7) et les propriétés indiquant une propension à la formation de feuillets-β (groupe no. 6). Par ailleurs, si les propriétés issues de l'abondance sont bien regroupées, l'algorithme différencie bien les abondances des protéines membranaires (groupe no.

12), des autres abondances (groupe no. 9). Pour des explications relatives à ces propriétés nous renvoyons à [8].

5. CONCLUSION

L'algorithme global k-means donne des résultats satisfaisants par rapport à notre problème. Il existe peu de littérature dans ce domaine permettant d'effectuer des comparaisons. La méthode d'agrégation utilisée est simple à implémenter et rapide à l'exécution. Cependant, on peut penser que des méthodes plus évoluées permettraient de mieux faire ressortir les groupes sous-jacents, ceux-ci ayant probablement une structure peu compacte. Or, les algorithmes k-means et global k-means sont plutôt adaptés à la détection de structures convexes. Nous proposons de répéter cette étude avec des méthodes basées sur la densité spatiale des données (DBSCAN [9]) ou des méthodes de graphe (coupes minimales). Toutefois, il faut relever que cette première approche est encourageante car les résultats ont un sens biochimique. Finalement, il intéressant de noter que la distance proposée dans cette étude peut être utilisée pour tous les problèmes dans lesquels la corrélation est une mesure de similarité.

No.	Taille	Signification
0	48	Propension hélice-α
1	19	Hydration, solvation
2	24	Propension hélice-α inversée, propension coude
3	49	Propension coude
4	36	Propension hélice, propension coude
5	38	Hydrophobicité ?
6	39	Propension feuillet-β
7	33	Volume, masse
8	11	Charge négative, propension hélice
9	30	Abondance
10	12	Point isoélectrique, charge positive, énergie d'activation de dépliement
11	19	?
12	16	Abondance dans les protéines membranaires
13	43	Hydrophobicité
14	57	Flexibilité, hydrophobicité
15	9	?
16	1	Point de fusion

TAB. 2 – Interprétation biochimique d'une agrégation en 17 groupes réalisée avec global k-means. Les points d'interrogation indiquent une incertitude quant au sens du groupe.

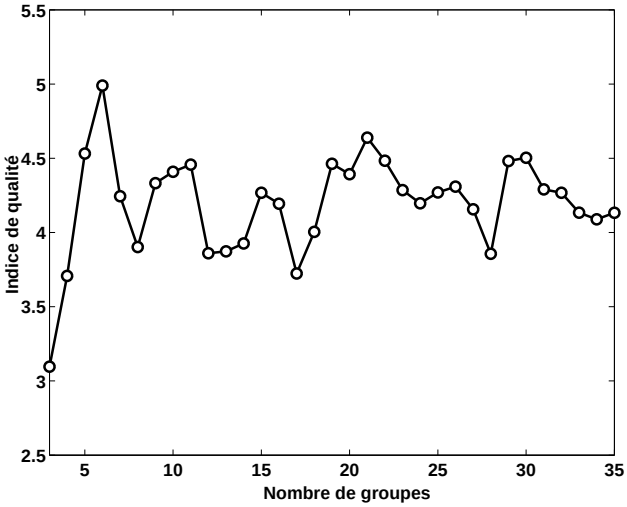

FIG. 4 – Indice de qualité SD (éq. 5) en fonction du nombre de groupes.

Références

[1] Kentaro Tomii and Minoru Kanehisa, "Analysis of amino indices and mutation matrices for sequence comparison and structure prediction of proteins", *Protein Engineering*, vol. 9, n. 1, pp. 27–36, 1996.

[2] "Amino acid indices and similarity matrices", `http://www.genome.ad.jp/dbget/aaindex.html`.

[3] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering : a review", *ACM Computing Surveys*, vol. 31, n. 3, pp. 264–323, 1999.

[4] A. Likas, N. Vlassis and J. J. Verbeek, "The global k-means clustering algorithm", Technical report, The Netherlands, 2001.

[5] J.-L. Falcone and P. Albuquerque, "A Correlation-Based Distance", `http://www.arxiv.org/abs/cs.IR/0402061`, February 2004.

[6] M. Halkidi, Y. Batistakis and M. Vazirgianni, "Cluster Validity Methods", *SIGMOD Record*, June-September 2002.

[7] "Amino-Acids Index Clustering", `http://cui.unige.ch/~falcone/AAIndex/`.

[8] Lubert Stryer, *Biochemistry*, W. H. Freeman and Company, fourth Edition, 1995.

[9] M. Ester et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in *KDD96*, pp. 226–331, 1996.