Recherche d'une signature proteomique du cancer du poumon.

Cécile Amblard, Sylvie Michelland, Florence De Fraipont, Denis Moro-Sibilot, François Godard, Marie-Christine Favrot, Michel Seve

Recherche d'une signature protéomique du cancer du poumon.

Cécile Amblard[1]& Sylvie Michelland[2] & Florence de Fraipont[2] & Denis Moro-Sibilot[3] & Francois Godard[2]& Marie Christine Favrot[2] & Michel Seve[2].

1: TIMC-IMAG, Faculté de Médecine, 38730 La Tronche,
2: Centre d'Innovation en Biologie, Pav B, CHU Grenoble BP217, 38043 Grenoble,
3: PMAC Clinique de Pneumologie, UF Oncologie thoracique, CHU Grenoble.

Résumé: Ce travail est consacré à la recherche d'une signature protéomique du cancer du poumon. A partir de spectres acquis avec un spectromètre de masses SELDI, nous recherchons un ensemble de protéines permettant de discriminer les patients atteints d'un cancer du poumon des sujets sains. Le problème est difficile, en particulier parce qu'il y a plus de protéines que de patients et parce qu'il y a une grande variabilité dans les spectres d'une même population. Nous cherchons aussi à nous affranchir des effets de l'âge et du sexe. Nous avons choisi d'utiliser Random Forest. Cette méthode de classification naturellement multivariée, permettant d'identifier les variables discriminantes et étant capable de travailler avec plus de variables que d'échantillons nous semble bien adaptée au problème posé. Nous avons ainsi trouvé un ensemble de protéines discriminantes dont certaines ont pu être identifiées.

Abstract: The aim of this study is to identify signature of protemic peaks that can be used to differentiate lung cancer patients from heallthy individuals. Plasma samples form 83 subjects are analyzed by Surface Enhanced Laser Desorption Ionisation (SELDI) time of light mass spetrometry. The problem is hard, especially because there are more proteins than subjects and some large variations in specta belonging to the same set. Furtemore we try to take into account the bias introduced by age or sexe of the subjects. We have choosen to use Random Forest- a multivariate classification or regression method, which can detect discriminant variables and which is able to work with more variables than subjetcs. We have found a set of discriminant proteins. Some of them have been named.

Mots clefs: Biologie -génome, Analyse des données-data Mining, Protéomique, Random-Forest.

1 Introduction

Le cancer du poumon est un cancer particulièrement agressif et est une cause majeure de mortalité dans le monde. Le diagnostique se fait par scanner-outil ne permettant pas de déceler un cancer précoce ou par prélèvement-méthode invasive et onéreuse, ne pouvant être utilisée dans le cadre d'un dépistage systématique. La mise en évidence

d'anomalies biologiques, détectables dans le sang, des mois avant l'apparition de signes cliniques de ce cancer permettrait d'effectuer une campagne de dépistage, de diagnostiquer des cancers en phase précoce et donc d'améliorer le taux de survie des sujets atteints. La présence d'une tumeur dans un tissu se traduit par une modification significative du profil protéomique du patient: présence ou absence anormale de certaines protéines. Ce travail est consacré à la recherche d'une signature protéomique du cancer du poumon, c'est à dire à l'identification de différences significatives entre les profits protéomiques de patients atteints du cancer et ceux de sujets sains. Les profils protéomiques sont acquis avec un spectromètre de masses à temps de vol SELDI, déjà utilisé en oncologie Wright (2002), Wiesner (2004). La recherche de protéines permettant de discriminer les sujets sains et les patients est un problème difficile: Il y a une grande variation entre les profils protéomiques de sujets appartenant à une même population. Le nombre de protéines ou de peptides est généralement bien supérieur au nombre de sujets. De plus, les protéines secretées par les tissus sont bien moins présentes que d'autres propres au serum très abondantes comme l'albumine, l'immunoglobuline, la transférine ou la fibrinogène, ce qui rend difficile leur identification. Enfin d'autres facteurs que la maladie peuvent avoir un effet significatif sur le profil protéomique d'un sujet, par exemple le sexe, l'âge...Nous avons choisi d'utiliser Random Forest Breiman (2001) pour réaliser l'influence du sexe et de l'age chez les sujets sains puis pour discriminer les profils protéomiques des sujets sains er des patients et identifier une signature protéomique du cancer du poumon en prenant en compte les effets d'age et de sexe. Cette méthode de classification naturellement multivariée, capable de travailler dans un contexte où il y a plus de variables que d'échantillons et enfin permettant d'identifier les variables les plus discriminantes nous a semblé très appropriée. Dans une première partie, nous présentons les données. Dans une seconde partie, nous expliquerons comment nous avons identifié l'effet du sexe, de l'age et de la maladie sur les profils protéomiques à l'aide de Random Forest. Enfin nous présenterons les résultats obtenus.

2 Description des données

L'étude porte sur un ensemble de 83 échantillons de plasma issus de 42 sujets identifiés comme cliniquement sains et de 41 patients atteints du cancer du poumon, dont 14 sont atteints d'un adénocarcinome, 26 d'un SCC (squamous cell carcinome) et 1 d'un autre type de cancer. Le profil protéomique de ces échantillons a été réalisé a l'aide de la plateforme ProtéinChip Biomarker System Cyphergen Biosystems Inc, Fremont CA). Un spectre est une collection ordonnée de couples (masse, quantité de particules ionisées détectées).
Un spectromètre de masses à temps de vol se compose schématiquement d'une entrée où l'on dépose une plaque de métal contenant une solution ionisée à analyser, d'une portion cylindrique que l'on peut soumettre à un champ electrique, d'une portion cylindrique sous vide qui aboutit à un détecteur de ions et un laser dont on peut régler l'intensité qui permet

de decrocher de leur support les particules ionisées. Une quantité de plasma mélangée à une solution ionisante: la matrice, est déposée sur un support métallique. Ce support est placé à l'entrée du spectromètre. Les particules ionisées sont détachées du support à l'aide du laser. Ces particules sont accélérées dans la partie soumise au champ électrique puis continuent à "voler" dans la partie sous vide jusqu'a atteindre le détecteur. Un spectre "brut" est donc un ensemble de couples où la première coordonnée est le temps (l'instant d'acquisition) et la seconde coordonnée est le nombre de ions d'ètectés à cet instant. Le temps de vol étant dépendant de la masse de la particule, le temps est classiquement transformé par une équation quadratique en ensemble de massess.
Avant de véritablement commencer l'analyse statistique des spectres, il est nécessaire d'effectuer un certain nombre de prétraitements sur ces données tels que le débruitage, la normalisation, l'alignement de spectres. Ces prétraitrements ont été réalisés à l'aide du logiciel Cyphergen ProteinChip Software 3.1. Le bruit additif, dû à la solution chimique ionisante se traduit par la présence d'un grand nombre de protéines en début de spectre. On s'affranchit de ce bruit en effectuant une régression sur les minima locaux et en soustrayant au signal cette tendance lisse obtenue. On s'affranchit du bruit de comptage en ne conservant sur une fenêtre de taille fixe que les pics pour lesquels le rapport signal sur bruit est supérieur à 5. L'alignement des pics se fait manuellement en alignant les centres de chaque pic diffus. Enfin, les spectres sont normalisés, en supposant que la charge totale de chaque spectre est identique. Après toutes ces étapes, nous disposons de 83 spectres constitués de 221 couples (masse, intensité) que nous appelons pic.

3 Analyse statistique

Le but de notre travail est de détecter un ensemble de protéines-signature du cancer du poumon afin d'être capable d'identifier un patient atteint de cette pathologie. Cette signature ne doit pas être biaisée par des facteurs exterieurs tels que le sexe ou l'age. C'est pourquoi cette analyse cherche à répondre aux 3 problématiques suivants:

- A : Le sexe d'un sujet a t il une influence sur le profil protéomique du sujet?
- B : L'age d'un sujet a t il une influence sur le profil protéomique du sujet?
- C : Existe t il une signature protéomique du cancer du poumon (resp. d'une histologie particulière du cancer du poumon)?

La question C est un problème d'analyse discriminante classique. La méthode utilisée doit identifier parmi tous les pics la combinaison qui discrimine au mieux les sujets sains et malades et doit nous permettre de détecter si un individu est sain ou malade avec un faible taux d'erreur. La méthode doit ètre capable de travailler avec un nombre de variables plus grand que le nombre de sujets. Random Forest (RF) permet justement

de réaliser ces 2 objectifs. C'est une méthode naturellement multivariée, capable de travailler lorsque il y a plus de variables que d'individus. De plus le pouvoir discriminant de chaque variable est réellement basé sur l'aptitude d'une variable à partitionner une population en sous groupes. Notons que cette méthode a déjà été utilisée dans le cadre de traitement de données génomiques (par exemple Lê Cao (2007) ou protéomiques (Izmirlian (2004). Nous avons utilisé la version libre sous R de Random Forest disponible à l'adresse http://cran.r-project.org. Pour répondre à la question C, nous avons utilisé RF sur l'ensemble des patients sains/malsains (resp. sains/atteints d'une histologie particulière du cancer du poumon). Afin de prendre en compte le fait que les classes sont mal équilibrées, nous avons utilisé comme paramètres un vecteur de poids égal à la proportion des individus dans chaque classe. Les pics liés à l'age ou au sexe sont retirés. Après avoir utilisé RF, nous conservons les 11 pics identifiés comme étant les plus discriminants par RF, rangés par pouvoir discriminant décroissant. Toutes les combinaisons, contenant le premier ou le second pic, de 2 à 9 pics, sont alors considérées. Pour chaque combinaison, RF est alors utilisé sur l'ensemble des spectres réduits à cet ensemble de variables. L'évaluation des résultats se fait alors à l'aide du taux d'erreur estimé à partir des éléments Out of Bag (OOB) Breiman (1996). Ce taux compte pour chaque arbre de la forêt, le nombre d'élements mal classés parmi les individus non répliqués dans l'échantillon bootstrap utilisé pour la construction de cet arbre. Les meilleures combinaisons sont donc celles ayant obtenu le meilleur taux d'erreur.
Pour répondre à la question A, nous avons utilisé RF comme suit. Nous nous sommes limité à l'ensemble des individus sains, constitué de 27 femmes et de 15 hommes.

1. Après avoir utilisé RF avec un vecteur de poids égal aux proportions de chaque classe, sur cette population, nous conservons les 11 variables les plus discriminantes, rangées par ordre de pouvoir discriminant decroissant.

2. Nous considérons toutes les combinaisons de 2 à 9 variables contenant la première variable. Si cette combinaison permet au moyen de RF de classer les individus de manière assez satisfaisante (sensibilité et spécificité évaluées avec les éléments OOB supérieures à 70%), la variable la plus discriminante est enlevée. elle est identifiée comme étant liée au sexe. On recommence alors en 1 avec les spectres précédents privés de cette variable.

3. L'algorithme s'arrête lorsqu'aucune combinaison restante ne permet d'obtenir une spécificté et une sensibilité supérieures à 70%. Les variables liées au sexe sont celle qui ont été ôtées.

Pour répondre à la question B, nous avons utilisé RF comme méthode de regression. La méthodologie est similaire à celle utilisée pour étudier l'effet du sexe; une combinaison "gagnante" à laquelle nous ôtons la première variable est une combinaison permettant d'expliquer 40% des variations de l'age.

4 Résultats

Les peptides liées au sexe sont celles situées aux masses 49201, 28115, 6633, 28327, 50167, 12850, 13793, 4588, 6436, 4452, 13289, 7443, 9350, 11728, 13051 et 9254. Les peptides liées à l'âge ont pour masses 13387, 9367, 13758 et 3102. Tous ces pics sont dorénavent ôtés des spectres. La meilleure combinaison de masses permettant de discriminer les patients sains et atteints d'un cancer du poumon est 11761, 6167, 8734, 8559, 8144, 6812 et 9160. Cette combinaison permet d'obtenir un taux d'erreur (OOB) de 91,5% avec une sensibilité de 90,2% et une spécificité de 92,8 %. 37 patients sur 41 et 39 sujets sains sur 42 sont bien classés. Il est intéressant de noter que 3 parmi les 4 patients mal classés sont atteints d'un adénocarcinome. Enfin, notons que la combinaison des trois premiers marqueurs permet d'obtenir un taux d'erreur de 87%. 6 combinaisons de peptides permettent de classer 96,1 % des patients atteints d'un SCC et 95,2 % des sujets sains. Ces combinaisons contiennent toutes les masses 11761, 6167, 14114 et 15287. Toutes contiennent de plus les masses 6609 ou 6812. On peut remarquer aussi que le marqueur 8734 précédement trouvé comme discriminant des sains/cancéreux n'apparait plus. Afin de compléter l'interprétation de ces résultats, nous avons fait une étude descriptive des différents marqueurs. Il apparait que les masses 6812, 6609 et 6418 sont sous-exprimées chez les patients atteints de SCC et très fortement positivement corrélées. Ces 3 masses contiennent probablement la même information. La masse 6167 ne fait pas parti de ce groupe. Nous n'avons pas fait l'étude des sujets sains/patients atteints d'un adénocarcinome parce qu'il y a trop peu d'adénocarcinomes (14).
Parmi toutes les protéines trouvées, trois ont pu être identifiées: la masse 11761 est la SAA: d'autres publications ont mentionné que cette protéine était élevée chez les patients atteints de cancer de al prostate ou du larynx. Les masses 13758 et 13793 sont respectivement la transtyretine et la cystatine.

Bibliographie

[1] Breiman L. (2001), Random Forest, *Machine Learning*, 45,5-22.

[2] Breiman L. (1996), Out of bag estimate, *Curr Pharm Biotechnol*

[3] Lê Cao K.-A., Gonçalves O., Besse P., Gadat S. (2007). Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm,*Statistical Applications in Genetics and Molecular Biology* , Vol. 6 : Iss. 1, article 29.

[4] Wiesner A. (2004), Detection of tumor markers with Protenichips technology,*Curr Pharm Biotechnol*, 5, 45-67.

[5] Wright GL. (2002), SELDI Protenichips MS: a platform for biomarker discovery and cancer diagnosis. *Expert Rev Mol Diagn*, 2, 549-63.