

Prédiction sur les protéines

Tous les programmes suivants sont des programmes qui calculent des "informations" sur les protéines à partir de leur séquence.

1 - Profil d'index

Les profils d'index sont définis par rapport à divers paramètres tels que la flexibilité, l'hydrophilicité, l'accessibilité, l'hydrophobicité, le potentiel d'hydratation etc... La plupart de ces paramètres sont dérivés de mesures expérimentales sur les aminoacides libres (tels que coefficient de partage entre phase aqueuse et phase organique, temps de rétention sur colonne HPLC, etc..) ou de paramètres calculés à partir des structure cristallines de protéines (flexibilité (facteur B), accessibilité ..)

Un profil d'index $\{MP(i)= f(i)\}$ est une représentation graphique faite par un programme qui calcule pour chacune des positions dans la séquence la valeur moyenne d'un paramètre (p) pour une région du polypeptide centrée autour de cette position :

$$MP(i) = \frac{1}{2m + 1} \sum_{j=i-m}^{j=i+m} p(j)$$

où p(j) est la valeur du paramètre p pour l'acide aminé en position j. On dit que c'est une moyenne centrée sur une fenêtre de longueur de (2m+1).

Nous pouvons classer les paramètres utilisés en deux catégories :

- hydrophilicité, accessibilité, flexibilité : ces paramètres sont témoins des aminoacides susceptibles d'être à la surface des protéines.

- hydrophobicité, hydrophobie : ces paramètres sont témoins des aminoacides susceptibles d'être à l'intérieur de la molécule ou dans un environnement apolaire.

La lecture de ces profils est une aide à la prédiction de déterminants antigéniques en utilisant la première catégorie de paramètres et à la prédiction de structure trans-membranaire en utilisant la deuxième catégorie de paramètres.

2 - Prédiction d'antigénicité

Les prédictions d'antigénicité les plus utilisées sont celles déduites des profils d'index et celle déduite des régions de haute mutabilité dans une famille de protéines.

2.1 - A partir des profils

Les profils les plus utilisés sont ceux d'hydrophilicité (Hoop et Woods) et les profils composés qui sont une combinaison linéaire d'un ensemble de paramètres (Parker, Jameson et Wolf). Parker, par exemple combine le temps de rétention en HPLC, l'accessibilité et la flexibilité.

2.2 - A partir des régions de haute mutabilité

Dans une série de séquences d'une famille de protéine, Frommel a observé une corrélation directe entre les régions de haute mutabilité et les déterminants antigéniques. Le paramètre de mutabilité, calculé sur une fenêtre centrée de $(2m+1)$ résidus, est la somme du nombre de mutations sur la famille de protéines en comparant ces dernières deux à deux.

3 - Prédiction de structure

Deux sortes de prédictions de structure secondaire :

- **prédiction statistique** : calcule pour une protéine les probabilités d'existence de structure (, , -turn, extended, coil) à partir de tableaux de valeurs expérimentales. Celles-ci sont calculées à partir de structures cristallines connues.

- **prédiction vectorielle** : (ou encore méthode des moments) elle permet de savoir si dans des structures de type ou , la molécule présente dans ces régions une possibilité de structure amphipatique (un coté hydrophobe et un coté hydrophile)

3.1 - Prédiction statistique

Citons trois méthodes qui utilisent la méthode de la fenêtre et ont chacune leurs tableaux de valeurs de référence.

3.1.1 Chou-Fasman

Les auteurs ont calculé les valeurs des paramètres de conformation d'un aminoacide de se trouver dans une structure α -hélice, β -sheet ou β -turn à partir de la structure cristalline de 29 protéines.

Le paramètre (ou score) pour une position donnée dans la séquence est calculée sur une fenêtre de 4 aminoacides en ne tenant compte que des trois suivants (nucléation d'une structure).

$$SC_s(i) = \frac{1}{4} \sum_{j=i}^{j=i+3} sc_s(j) \quad \text{où } s \text{ fait référence aux différentes structure } , \text{ ou } \text{-turn et } sc_s(j)$$

est la valeur du score de l'acidoacide dans le tableau de référence.

Est ajoutée à ceci un tableau de fréquence d'apparition des aminoacides participant à une structure de coude. Pour ce paramètre, la valeur est calculée ainsi :

$$F(i) = \sum_{j=i}^{j=i+3} f(j)$$

Une fois ce tableau de quatre valeurs calculées pour chaque position le programme les analyse avec les règles suivantes : il prend tout d'abord en considération la valeur la plus grande des trois structures prédites et puis :

- **α -hélice** : une région de 4 résidus consécutifs, où SC est supérieur SC et à SC_{turn} , initie une α -hélice et la région est étendue à droite et à gauche jusqu'à la rencontre de SC inférieure à 1. Cette région doit remplir les conditions suivantes :

- ° longueur au moins de six résidus
- ° pas de proline à l'intérieur de l' α -hélice ou du côté C-terminal

- **β -sheet** : une région de trois résidus, où SC est supérieur SC et à SC_{turn} , initie une structure de β -sheet et la région est étendue à droite et à gauche jusqu'à la rencontre de SC inférieure à 1.

- **β -turn** : il faut SC_{turn} supérieur à SC et SC pour une région avec quatre aminoacides avec pour le premier une valeur de F supérieure à un seuil ($0,75 \cdot 10^{-4}$)

3.1.2 Gor method

Cette méthode tient compte du fait que la probabilité d'un aminoacide d'appartenir à un type de structure secondaire dépend de la nature et de la position de ses voisins. Pour cela, les auteurs (Garnier et Robson) utilise la théorie de l'information ou probabilité conditionnelle. Soient deux événements, x et y, soit $P(x|y)$ la probabilité conditionnelle que x advienne sachant que y est advenu. On appelle l'information associé à l'événement x contraint par y :

$$I(x;y) = \log(P(x|y) / P(x)) \quad \left\{ \begin{array}{l} \text{si } P(x|y) = P(x) \text{ x indépendant de y alors } I(x;y) = 0 \\ \text{si } P(x|y) > P(x) \text{ y favorise x } \quad I(x;y) > 0 \\ \text{si } P(x|y) < P(x) \text{ y défavorise x } \quad I(x;y) < 0 \end{array} \right\}$$

Si y se décompose en deux événements y_1, y_2 , nous avons :

$$\begin{aligned} I(x;y_1, y_2) &= \log \{ P(x|y_1, y_2) / P(x) \} \\ &= \log \{ P(x|y_1, y_2) / P(x|y_1) \} + \log \{ P(x|y_1) / P(x) \} \\ &= I(x;y_2|y_1) + I(x;y_1) \end{aligned}$$

Le premier terme représente l'effet de y_2 sur l'événement x, sachant que y_1 a eu lieu.

ou encore de manière plus générale :

$$I(x;y_1, y_2, y_1, y_3 \dots y_n) = I(x;y_1) + I(x;y_2|y_1) + I(x;y_3|y_1, y_2) + \dots + I(x;y_n|y_1, \dots, y_{n-1})$$

Considérons que pour le premier événement, il n'y ait que deux résultats possibles x et son événement contraire \bar{x} , nous pouvons définir la préférence de y pour l'événement x comme

$$I(S=x: \bar{x} ; y) = I(S=x ; y) - I(S=\bar{x} ; y) \text{ ou encore :}$$

$$= \log \{ P(S=x|y) / P(S=x) \} - \log \{ P(S=\bar{x} |y) / P(S=\bar{x}) \}$$

Remarquons que $P(S=\bar{x}) = 1 - P(S=x)$ et $P(S=\bar{x} |y) = 1 - P(S=x|y)$

L'application à la prédiction de structure va se faire de la manière suivante :

S sera l'ensemble de l'état α -hélice ou non α -hélice et nous prendrons pour les événements représentés par $y_1 \dots y_n$ les positions dans la séquence (R). Bien sur il y aura les mêmes définitions pour les structures β , β -turn et coil.

La préférence pour une structure α -hélice pour un aminoacide à la position j dans la séquence de longueur n sera évaluée par le paramètre :

$$I(S_j=H: \bar{H} ; R_1 \dots R_n)$$

Nous ferons les approximations suivantes :

$$1 - I(S_j=H: \bar{H} ; R_1 \dots R_n) \approx I(S_j=H: \bar{H} ; R_{j-m} \dots R_{j+m})$$

Ceci indique que l'influence des amino acides voisins sera limitée à une fenêtre centrée de longueur $(2m + 1)$. Les auteurs ont pris $m = 8$

2 - En se référant au développement précédent, nous avons :

$$I(S_j=H: \bar{H} ; R_{j-m} \dots R_{j+m}) = I(S_j=H: \bar{H} ; R_j)$$

$$+ I(S_j=H: \bar{H} ; R_{j-1} | R_j)$$

$$+ I(S_j=H: \bar{H} ; R_{j+1} | R_j, R_{j-1})$$

$$+ \dots \dots \dots$$

$$+ I(S_j=H: \bar{H} ; R_{j+m} | R_{j-m}, \dots, R_j, \dots, R_{j+m-1})$$

Cette formule, qui contient $(2m+1)$ facteurs, est simplifiée et nous obtenons les deux méthodes "GOR" suivantes:

information directionnelle (GOR II)

$$I(S_j=H: \bar{H} ; R_{j-m} \dots R_{j+m}) \approx \frac{I(S_j=H: \bar{H} ; R_{j+m})}{m}$$

L'information du résidu à la position $(j+m)$ est la même quel que soit le résidu à la position j

et information par paire (GOR III)

$$I(S_j=H: \overline{H} ; R_{j-m} .. R_{j+m}) \simeq I(S_j=H: \overline{H} ; R_j) + \sum_{m, m=0} I(S_j=H: \overline{H} ; R_{j+m} | R_j)$$

Le deuxième paramètre informationnel ne tient compte que des paires : amino acide à la position j et amino acide à la position (j+m).

Les tableaux de référence sont calculés comme pour la méthode précédente à l'aide de la structure cristalline de 75 protéines.

Pour la méthode "information directionnelle", il comprend pour un type de structure 17 x 20 valeurs ce qui fait pour les quatre structures :

4 x 17 x 20. Elles ont été calculées à partir de la formule :

$$I(S_j=x: \overline{x} ; R_{j+m}) = \log \{ P(S_j=x | R_{j+m}) / P(S=x) \} - \log \{ P(S= \overline{x} | R_{j+m}) / P(S= \overline{x}) \}$$

où les probabilités sont des fréquences observées dans ces 75 protéines, et en tenant compte que x et \overline{x} sont des événements contraires.

3.1.3 Gascuel et Goldmard

Comme la précédente, cette méthode tient compte du fait que la probabilité d'un aminoacide d'appartenir à un type de structure secondaire dépend de la nature et de la position de ses voisins. Un score est calculé pour chacun des états possibles (S : -hélice, extended et coil) en déclarant que l'état de l'acidoacide considéré est d'autant plus influencé par un autre aminoacide que celui-ci est proche de l'acidoacide considéré :

$$CBLF(i, S) = N(S) \prod_{j=i-n}^{j=i+m} I(j, S) P(j, S)$$

N(S) est un facteur de normalisation associé à chaque état, I(j, S) mesure la préférence pour l'acidoacide en position j pour l'état S et P(j, S) est un poids qui dépend uniquement de la position relative de j par rapport à i (P(i, S)=1).

Les auteurs ont remarqué que l'influence des aminoacides n'était par forcément symétrique pour chacun des états de structure étudié. Ils ont défini leur fenêtre de calcul ainsi :

pour les structures en	-hélice la fenêtre est	n = -6	m = 11
	extended	n = -3	m = 3
	coil	n = -6	m = 3

Pour chacune des trois structures, un tableau de valeurs pour les différents paramètres a été calculé à partir de la structure cristallines de 65 protéines.

La probabilité d'un état de structure S à une position considérée est égale à :

$$P(i, S) = \frac{e^{\text{CBLF}(i, S)}}{e^{\text{CBLF}(i, H)} + e^{\text{CBLF}(i, E)} + e^{\text{CBLF}(i, C)}}$$

et l'état associé est celui de plus forte probabilité.

3.2 - Prédiction vectorielle

Appelée encore méthode des moments (Enseinberg), celle-ci calcule une grandeur vectorielle qui fera apparaître la dissymétrie de distribution des chaînes latérales hydrophobe.

Soit pour l'acide aminé i , \vec{s}_i le vecteur perpendiculaire au carbone α de l'acide aminé et pointant sur le centre de la chaîne latérale : le moment d'hydrophobicité est défini par :

$\vec{\mu}_{ih} = H_i \vec{s}_i$ où H_i est l'indice d'hydrophobicité de l'acide aminé en question (positif pour les acides aminés apolaires et négatif pour les polaires)

Pour un segment protéique, le moment structural d'hydrophobicité est :

$$\vec{\mu}_{(n,m)h} = \sum_{i=n}^{i=m} H_i \vec{s}_i, \text{ si ce moment est non nul il indiquera une dissymétrie sur la}$$

distribution des acides aminés par rapport à l'hydrophobicité. La molécule présentera dans cette région une structure amphipatique (hydrophobe d'un côté et non hydrophobe du côté opposé)

Si nous supposons que le segment protéique a une structure secondaire régulière et que pour chaque acide aminé le vecteur est normal à l'axe de symétrie et pointe vers l'extérieur, alors l'amplitude du moment structural d'hydrophobicité d'un peptide est égale à :

$$\mu_{(n,m)h} = \left[\left(\sum_{i=n}^{i=m} H_i \sin(\theta_i) \right)^2 + \left(\sum_{i=n}^{i=m} H_i \cos(\theta_i) \right)^2 \right]^{\frac{1}{2}}$$

La valeur de θ_i est égale à 100° pour la structure α -hélice et 160° pour les β -sheet et 180° pour les β -sheet plans.

Remarque : on peut aussi faire une projection d'une hélice sur un plan et entourer les résidus hydrophobe pour se représenter cette dissymétrie (voir par exemple le programme `helicalwheel` du logiciel "GCG")

Quelques références

- Chou, P.Y. and Fasman, G. D. (1978) *Ann. Rev. Biochem.* **47**, 251-276
- Enseinberg D., Weiss R. and Terwilliger T. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 140-144
- Fraga, S. (1982) *Can. J. Chem.* **60**, 2606-2610
- Frommel, C. (1988) *J. Theor. Biol.* **132**, 171-177
- Gascuel O. and Golmard J.L. (1988) *CABIOS* **4**, 357-365
- Gibrat, J.F., Garnier, J. and Robson, B. (1987) *J. Mol. Biol.* **198**, 425-443
- Hopp, T.P. and Woods, K.R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3824-3828
- Jameson and Wolf (1988) *CABIOS* **4**, 181-186
- Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.* **157**, 105-132
- Parker, J.M.R., Guo, D. and Hodges, R.S. (1986) *Biochemistry* **25**, 5425-5432