Une méthode ABC de choix de modèle pour les champs de Gibbs
Application à la prédiction de structure 3D de protéine

Aude GRELAUD[†‡◇], Jean-Michel MARIN [*◇] , Christian P. ROBERT[‡◇] , François RODOLPHE[†] , Jean-François TALY[†]

† Unité Mathématique, Informatique et Génome, INRA
* I3M, Université Montpellier 2
‡ Cérémade, Université Paris Dauphine
◇ Laboratoire de Statistique, CREST-INSEE

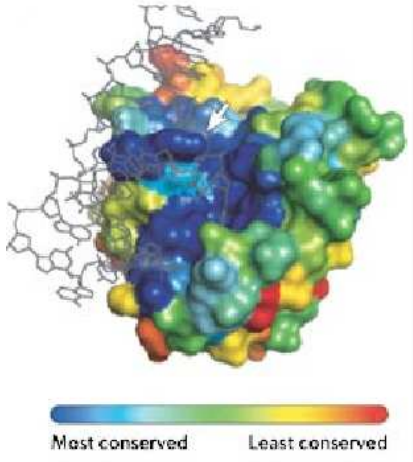
Applibugs, 4 juin 2009

Problème de départ

- Comment choisir la "meilleure" prédiction de structure 3D pour une protéine donnée ?
- **Information disponible :** séquence d'AAs + prédictions de structure.
-

Problème de départ

- Comment choisir la "meilleure" prédiction de structure 3D pour une protéine donnée ?
- **Information disponible :** séquence d'AAs + prédictions de structure.
- **Propriété à utiliser :** AAs en contact dans la structure 3D ont des propriétés simulaires.



Comment faire ?

- Problème de choix de modèle
- Quels modèles ?
  - Chaque AA appartient à une catégorie $k \in 1..K$
  - Etat d'un AA ne dépend que de l'état de ses voisins :

    $\hookrightarrow$ **Champ markovien**

Comment faire ?

- Problème de choix de modèle
- Quels modèles ?
  - Chaque AA appartient à une catégorie $k \in 1..K$
  - Etat d'un AA ne dépend que de l'état de ses voisins :

    $\rightsquigarrow$ **Champ markovien**

Mais...

- Modèles difficiles à utiliser car vraisemblance connue à une constante multiplicative près ...
- Complique d'autant plus le choix de modèle ! ! !

Définition des champs de Gibbs

Modèle statistique de vraisemblance :

$$f(\mathbf{x}|\theta) = \frac{1}{Z_\theta} \exp\{\theta^T S(\mathbf{x})\},$$

où :

- $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$,
- $S(\cdot)$ fonction de potentiel, à valeur dans $\mathbb{R}^p$,
- $\theta \in \mathbb{R}^p$ paramètre d'échelle associé au modèle,
- $Z_\theta = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{\theta^T S(\mathbf{x})\}$ constante de normalisation.

Définition des champs de Gibbs

Modèle statistique de vraisemblance :

$$f(\mathbf{x}|\theta) = \frac{1}{Z_\theta} \exp\{\theta^T S(\mathbf{x})\},$$

où :

- $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$,
- $S(\cdot)$ fonction de potentiel, à valeur dans $\mathbb{R}^p$,
- $\theta \in \mathbb{R}^p$ paramètre d'échelle associé au modèle,
- $Z_\theta = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{\theta^T S(\mathbf{x})\}$ constante de normalisation.

$\leadsto Z_\theta$ non disponible en général.

Un exemple intéressant

- Pour modéliser une dépendance spatiale entre les données (épidémiologie, image...)
- Modèle de Potts :

$$S(\mathbf{x}) = \sum_{i \sim i'} \mathbb{I}_{\{x_i = x_{i'}\}} ,$$

où $i \sim i'$ signifie que i et i' sont voisins.

Sélection de modèle

1 voisinage $\longleftrightarrow$ 1 potentiel $\longleftrightarrow$ 1 modèle

où :

- $S_m(\cdot)$ fonction de potentiel du modèle m,
- θ_m paramètre d'échelle associé au modèle m,
- Z_{m,θ_m} constante de normalisation.

Modélisation

On se place dans un **cadre bayésien** :

- Vraisemblance :
  $(\mathbf{x}|\theta_m, m) \sim f_m(\mathbf{x}|\theta_m, m)$
- Distributions a priori :
  - $m \sim \pi(m)$
  - $(\theta_m|m) \sim \pi_m(\theta_m)$

$\hookrightarrow$ **Objectif :** distribution a posteriori de m, $P(\mathcal{M} = m|\mathbf{x})$.

Critère de sélection : Bayes factor

$$BF_{m_0/m_1}(\mathbf{x}) = \frac{P(\mathcal{M} = m_0|\mathbf{x})}{P(\mathcal{M} = m_1|\mathbf{x})} \Big/ \frac{\pi(m_0)}{\pi(m_1)}$$

$$= \frac{\int f_{m_0}(\mathbf{x}|\theta_{m_0}, m_0)\pi_0(\theta_0)\mathrm{d}\theta_0}{\int f_{m_1}(\mathbf{x}|\theta_{m_1}, m_1)\pi_1(\theta_1)\mathrm{d}\theta_1}.$$

Interprétation :

- Si $BF_{m_0/m_1}(\mathbf{x}) > 1$, l'évidence est plus forte en faveur du modèle m_0,
- Si $BF_{m_0/m_1}(\mathbf{x}) < 1$, l'évidence est plus forte en faveur du modèle m_1.

Ici,

$$BF_{m_0/m_1}(\mathbf{x}) = \frac{\int \exp\{\theta_0^{\mathsf{T}} S_0(\mathbf{x})\}/Z_{\theta_0,0}\pi_0(\theta_0)\mathrm{d}\theta_0}{\int \exp\{\theta_1^{\mathsf{T}} S_1(\mathbf{x})\}/Z_{\theta_1,1}\pi_1(\theta_1)\mathrm{d}\theta_1},$$

$\rightsquigarrow$ calcul direct impossible,

$\rightsquigarrow$ méthodes MCMC non applicables,

$\Longrightarrow$ Utilisation d'une méthode **sans vraisemblance**.

Idée

- Données qui se "ressemblent" correspondent à des valeurs proches pour les paramètres.
- Ne pas savoir calculer la vraisemblance ne veut pas dire ne pas savoir simuler selon le modèle.

A définir :

- Comment choisir les valeurs des paramètres utilisés pour simuler ?
- Comment décider que 2 jeux de données se ressemblent suffisament ?

Exact rejection sampling

1. Générer θ^* suivant le prior $\pi(.)$.
2. Générer $\mathbf{x}^*$ suivant $f(\cdot|\theta^*)$.
3. Accepter θ^* si $\mathbf{x}^* = \mathbf{x}^0$.

Résultat :

- Etapes 1 et 2 : Simulation d'un couple $(\theta^*, \mathbf{x}^*)$ de densité $\pi(\theta^*) f(\mathbf{x}^*|\theta^*) \propto \pi(\theta^*|\mathbf{x}^*)$.
- On accepte si $\mathbf{x}^* = \mathbf{x}^0$: θ^* exactement de densité $\pi(\theta|\mathbf{x}^0)$.

Limite : Taux d'acceptation petit, voir nul dans le cas continu !

ε-tolerance rejection sampling

1. Générer θ^* suivant le prior $\pi(.)$.
2. Générer $\mathbf{x}^*$ suivant $f(\cdot|\theta^*)$.
3. Accepter θ^* si $\rho(\mathbf{x}^*, \mathbf{x}^0) < \varepsilon$.

Résultat : $(\theta^{1*}, \cdots, \theta^{*n})$ de densité $\pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0) < \varepsilon)$.

$\hookrightarrow$ Si ε est suffisamment petit, bonne approximation de $\pi(\theta|\mathbf{x}^0)$.
(En pratique $q_{1\%}$ sur les distances)

Limite : Comment définir ρ quand $\mathbf{x}$ de grande dimension ?

Approximate Bayesian Computation (ABC)

Soit T un vecteur de statistiques résumées :

---

1. Générer θ^* suivant le prior $\pi(.)$.
2. Générer $\mathbf{x}^*$ suivant $f(\cdot|\theta^*)$.
3. Accepter θ^* si $\rho(T(\mathbf{x}^0), T(\mathbf{x}^*)) < \varepsilon$.

Résultat : $(\theta^{1*}, \cdots, \theta^{*n})$ de densité $\pi(\theta|\rho(T(\mathbf{x}^*), T(\mathbf{x}^0)) < \varepsilon)$.

$\hookrightarrow$ Si ε est suffisamment petit, et T bien choisie, bonne approximation de $\pi(\theta|\mathbf{x}^0)$.

$\hookrightarrow$ Comment choisir T ?

Choix de T

- $dim(T) \geq 1$,
- T bon résumé de $\mathbf{x}$,
- Idéal : statistique exhaustive.

$\hookrightarrow$ Si T est exhaustive et $\varepsilon = 0$: ABC exact car

$$\pi(\theta|\rho(T(\mathbf{x}^*), T(\mathbf{x}^0)) = 0) = \pi(\theta|\mathbf{x}^* = \mathbf{x}^0)$$

Algorithme ABC-MC

Approximate Bayesian Computation Model Choice

0. Générer m^* suivant $\pi(\mathcal{M} = m)$.
1. Générer $\theta^*_{m^*}$ suivant $\pi_{m^*}(\cdot)$.
2. Générer x^*suivant $f_{m^*}(\cdot|\theta^*_{m^*})$.
3. Accepter $(\theta^*_{m^*}, m^*)$ si $\rho(T(\mathbf{x}^0), T(\mathbf{x}^*)) < \varepsilon$.

Résultat : (m^*, θ_{m^*}) de distribution $\pi\left\{(m, \theta_m)|\rho(T(\mathbf{x}^0), T(\mathbf{x}^*)) < \varepsilon\right\}$.

Estimation du Bayes factor

- Approximation de Monte Carlo de $P(\mathcal{M} = m|\mathbf{x}^0)$:

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}^0) = \sharp\{m^{i*} = m\}\big/N\,.$$

- Estimateur du BF correspondant :

$$\begin{aligned}\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M} = m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M} = m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)} \\ &= \frac{\sharp\{m^{i*} = m_0\}}{\sharp\{m^{i*} = m_1\}} \times \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)}\,.\end{aligned}$$

- En pratique, on utilise :

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1 + \sharp\{m^{i*} = m_0\}}{1 + \sharp\{m^{i*} = m_1\}} \times \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)}\,.$$

Application aux champs de Gibbs

Choix de la statistique résumée T :

- S_m : statistique exhaustive pour θ_m,
- $S = (S_1, \cdots, S_M)$: statistique exhaustive pour $(\theta_1, \cdots, \theta_M)$,

$$\begin{aligned} P(X = \mathbf{x} | S(\mathbf{x}) = s, \mathcal{M} = m) &= \frac{1}{n(S(\mathbf{x}))} \\ &= P(X = \mathbf{x} | S(\mathbf{x}) = s) \end{aligned}$$

avec $n(S(\mathbf{x})) = \sharp \{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$.

$\leadsto (S_1, \cdots, S_M)$ exhaustive pour m.

$\leadsto T = (S_1, \cdots, S_M)$.

Modèles

- **M0 : cas iid,** $Bernouilli(p)$
  - $f_0(X|\theta_0, m=0) = \frac{1}{Z_{\theta_0,0}} exp(\theta_0 \sum_i \mathbf{1}_{\{x_i=1\}})$
  - $p = \frac{exp(\theta_0)}{1+exp(\theta_0)}$
  - $S_0(\mathbf{x}) = \sum_i \mathbf{1}_{\{x_i=1\}}$
- **M1 : Chaîne de Markov de matrice de transition** P
  - $f(X|\theta, 1) = \frac{exp(2\theta_1 \ \sum_{i=1}^{n-1} \mathbf{1}_{\{x_i=x_{i+1}\}})}{2(1+exp(2\theta_1))^{n-1}}$
  - $P = \begin{pmatrix} \frac{exp(2\theta_1)}{1+exp(2\theta_1)} & \frac{1}{1+exp(2\theta_1)} \\ \frac{1}{1+exp(2\theta_1)} & \frac{exp(2\theta_1)}{1+exp(2\theta_1)} \end{pmatrix}$
  - $S_1(\mathbf{x}) = \sum_{i=1}^{n-1} \mathbf{1}_{\{x_i=x_{i+1}\}}$

En pratique

Choix des priors :

- modèles équiprobables a priori,
- uniforme pour les paramètres.

Statistique résumée :

$T = \left(\sum_i \mathbf{1}_{\{x_i=1\}}, \sum_{i=1}^{n-1} \mathbf{1}_{\{x_i=x_{i+1}\}}\right)$,

$\leadsto$ statistique exhaustive pour (m, θ_0, θ_1).

Résultats

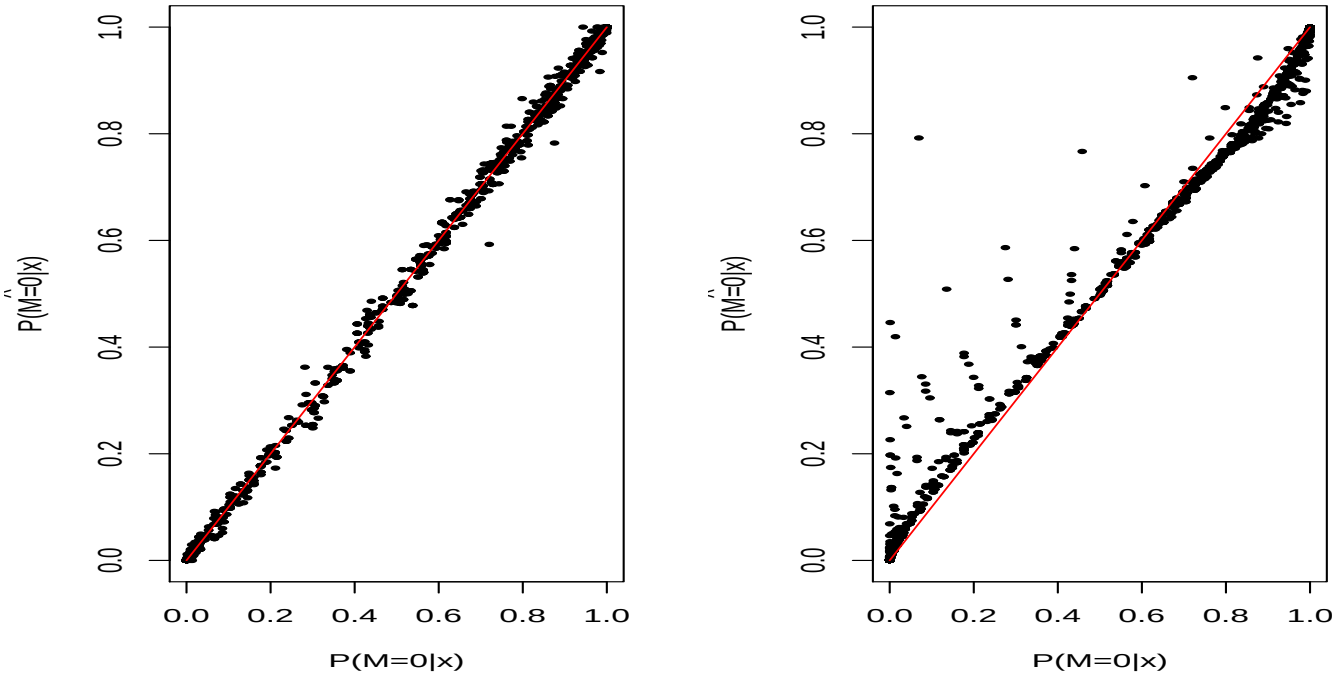


Figure: *(left)* Comparaison de la vraie valeur $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ à $\widehat{\mathbb{P}}(\mathcal{M} = 0|\mathbf{x}^0)$ sur 2 000 séquences simulées et en utilisant un échantillon de 4.10^6 propositions. La ligne rouge correspond à la diagonale. *(droite)* Même comparaison en utilisant un seuil ε correspondant au quantile à 1% des distances.

Résultats (2)

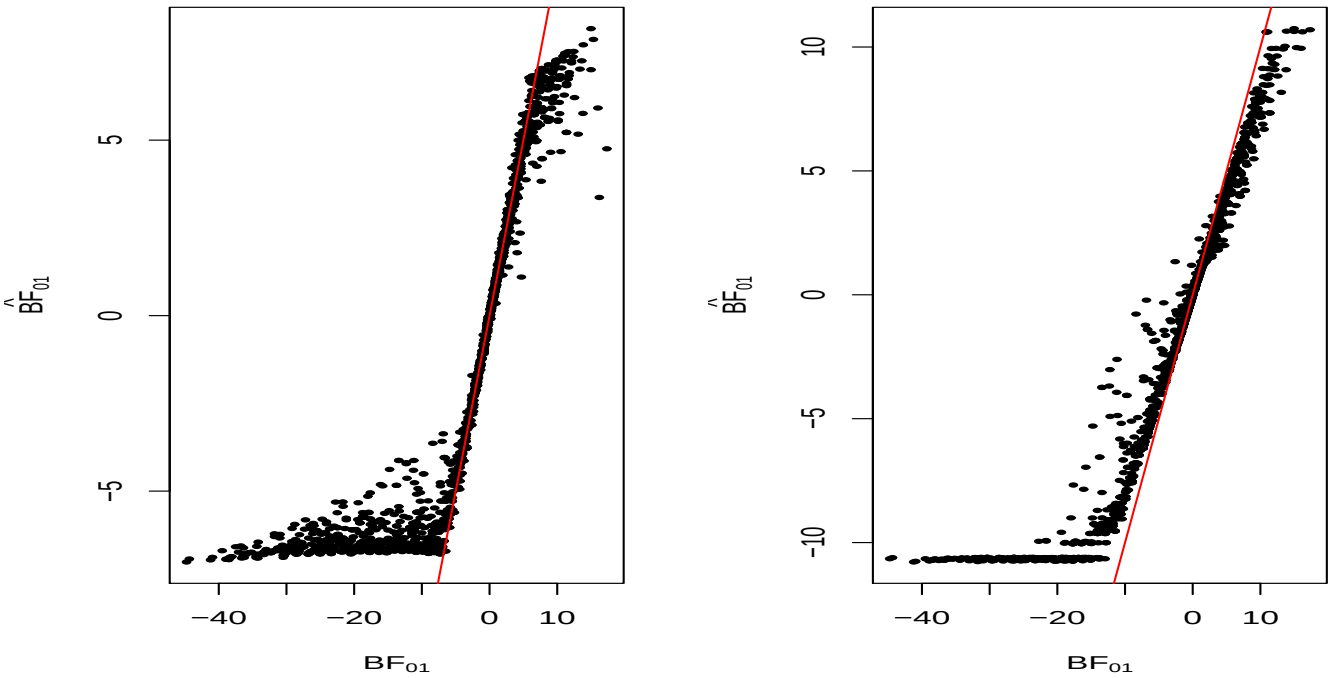


Figure: *(left)* Comparaison de la vraie valeur $BF_{0/1}$ à $\widehat{BF_{0/1}}$ sur 2 000 séquences simulées et en utilisant un échantillon de 4.10^6 propositions. La ligne rouge correspond à la diagonale. *(droite)* Même comparaison en utilisant un seuil ε correspondant au quantile à 1% des distances.

Résultats, cas exact

	$m=1$ dec.	$m=1$ fort	$m=1$ sub.	$m=1$ faible	$m=0$ faible	$m=0$ sub.	$m=0$ fort	$m=0$ dec.
$m=1$, dec.	778	9	0	0	0	0	0	0
$m=1$, fort	2	79	0	0	0	0	0	0
$m=1$, sub.	0	7	53	0	0	0	0	0
$m=1$, faible	0	0	2	63	0	7	0	0
$m=0$, faible	0	0	0	22	103	7	0	0
$m=0$, sub.	0	0	0	0	1	103	23	0
$m=0$, fort	0	0	0	0	0	5	177	6
$m=0$, dec.	0	0	0	0	0	0	13	547

Comparaison de la vraie valeur de $BF_{0/1}$ avec son estimation $\widehat{BF_{0/1}}$ sur 2 000 séquences simulées et en utilisant un échantillon de 4.10^6 propositions. Les catégories correspondent à l'échelle de Jeffrey.

Résultats , $\varepsilon = q_{1\%}$

	$m=1$ dec.	$m=1$ fort	$m=1$ sub.	$m=1$ faible	$m=0$ faible	$m=0$ sub.	$m=0$ fort	$m=0$ dec.
$m=1$, dec.	740	39	5	2	0	0	1	0
$m=1$, fort	0	64	14	2	1	0	0	0
$m=1$, sub.	0	0	39	19	2	0	0	0
$m=1$, faible	0	0	0	61	3	0	1	0
$m=0$, faible	0	0	0	2	128	2	0	0
$m=0$, sub.	0	0	0	0	2	123	1	1
$m=0$, fort	0	0	0	0	0	26	161	1
$m=0$, dec.	0	0	0	0	0	0	71	489

Comparaison de la vraie valeur de $BF_{0/1}$ avec son estimation $\widehat{BF_{0/1}}$ sur 2 000 séquences simulées et en utilisant un échantillon de 4.10^6 propositions. Les catégories correspondent à l'échelle de Jeffrey.

Comment déterminer la structure 3D d'une protéine ?

- **Expérimental :**
  - Cristallographie par rayon X,
  - Spectroscopie par résonance magnétique nucléaire,
  - Cryomicroscopie.

  ↬ Délicat et coûteux.

- **Méthodes automatiques :**
  - Basées sur les structures de protéines connues : méthodes par homologie ou par reconnaissance des repliements (*Protein Threading*),
  - Basées sur d'autres données (propriétés physico-chimiques des atomes par exemple) : méthodes *ab initio* et *de novo*.

  ↬ Besoin d'un critère de choix entre les candidats.

Idée du travail

- AAS en contact dans la structure 3D ont souvent des propriétés biochimiques similaires.
- Ici : utilisation de l'hydrophobicité (2 catégories).

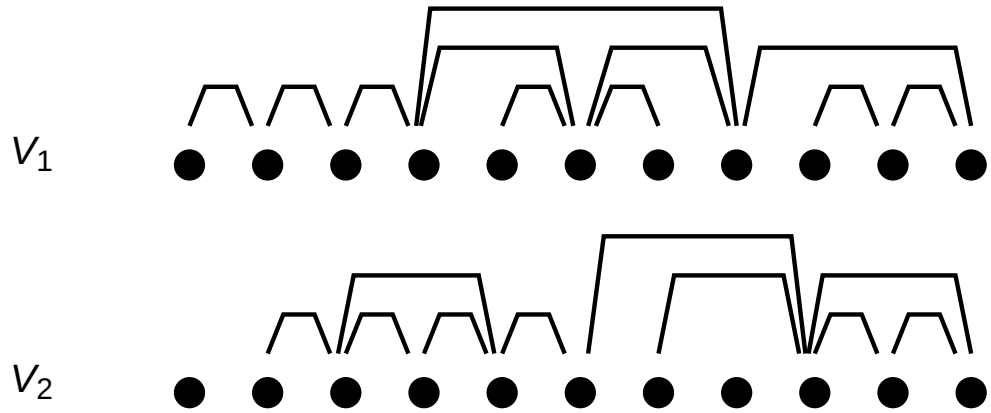
Modélisation statistique du problème

- $\mathbf{x} = (x_1, \cdots, x_n)$ avec $\forall i \in 1, \cdots, n, x_i \in 1, \cdots, K$

Modélisation statistique du problème

- $\mathbf{x} = (x_1, \cdots, x_n)$ avec $\forall i \in 1, \cdots, n, x_i \in 1, \cdots, K$

- Chaque structure définit un graphe de voisinage :



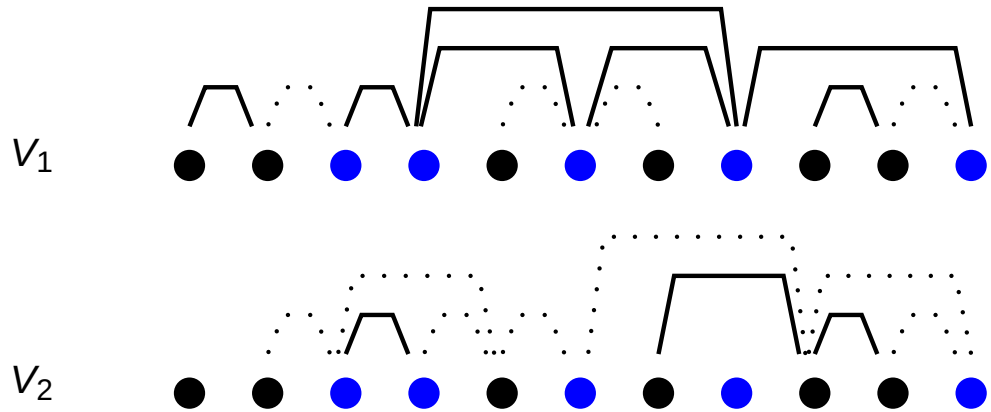
- A chaque structure on associe un champs de Gibbs :

  $\hookrightarrow$ Problème de choix de modèle

- A chaque structure on associe un champs de Gibbs :

  $\hookrightarrow$ Problème de choix de modèle

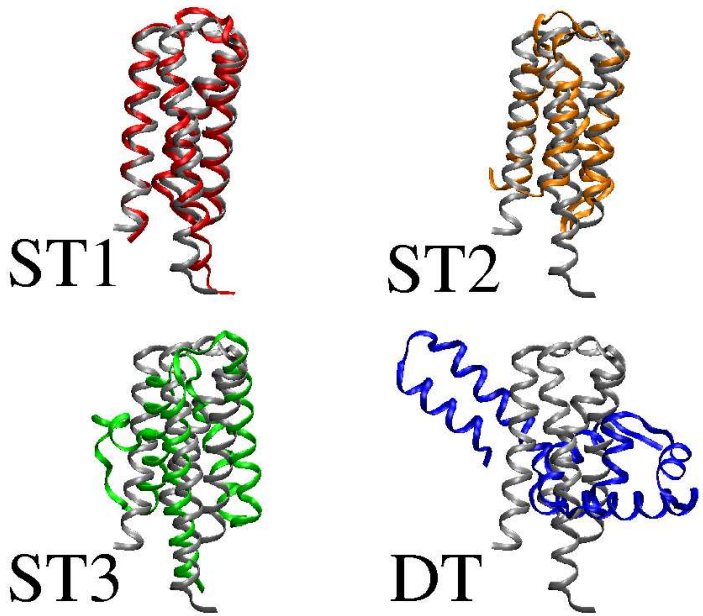
- Quel est la structure de voisinage la plus adéquate ?



Jeu de données étudié

Protéine de *Thermotoga maritima*

- Structure native (NS)
- 4 structures candidates proposées par FROST



Résultats

		FROST score	$\widehat{BF_{NS/.}}$
1i5nA	(**ST1**, Similaire)	75.3 $\longrightarrow$ Ok	1.34
1ls1A1	(**ST2**, Similaire)	8.9 $\longrightarrow$?	1.22
1jr8A	(**ST3**, ?)	8.9 $\longrightarrow$?	2.42
1s7oA	(**DT**, non similaire)	7.8 $\longrightarrow$? ?	2.76

FROST score : qualité de l'alignement de la séquence étudiée sur la structure candidate.

$>$ 9 : bon alignement, $<$ 7 mauvais.

$\hookrightarrow$ Distingue mieux les structures similaires des non similaires que FROST.

Conclusion

- ABC permet de faire du choix de modèle sans utiliser la vraisemblance.
- Champs de Gibbs : statistique exhaustive disponible, pas de perte d'efficacité.
- Application biologique : tester d'autres propriétés.
- Réseaux : autre type de champs de Gibbs à explorer...

Références ABC

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, M. W. Feldman,
Population growth of human Y chromosomes : a study of Y chromosome microsatellites,
Mol. Biol. Evol. 16 (1999) 1791–1798.

M. Beaumont, W. Zhang, D. Balding,
Approximate Bayesian Computation in population genetics,
Genetics 162 (2002) 2025–2035.

A. Grelaud, C.P. Robert, J.-M. Marin, F. Rodolphe, J.-F. Taly,
ABC methods for model choice in Gibbs random fields,
arXiv :0807.2767, (2008).

Autres références

B.P. Carlin, S. Chib,
Bayesian model choice via Markov chain Monte Carlo methods,
J. Royal Statist. Society Series B. 57 (1995) 473–484.

N. Cressie,
Statistics for Spatial Data, revised edition,
Wiley Series in Probability and Statistics, New York, 1993.