Using selective pressure to improve protein tridimensional structure prediction

Aude GRELAUD[1,2] Jean-Michel MARIN [3] , Christian P. ROBERT[1] , François RODOLPHE[2]

[1] Cérémade, Université Paris Dauphine et Laboratoire de statistique, CREST-INSEE
[2] Unite Mathématique, Informatique et Génome, INRA
[3] INRIA Saclay

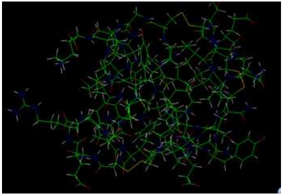
MIEP

Hameau de l'Etoile, june 2008

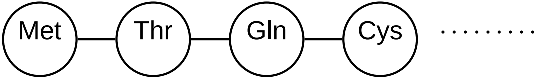
1. Context
2. Markov random fields
3. Parameter posterior distribution
4. Model choice
5. Simulations
6. Conclusion

Aim

Predict the tridimensional structure of the protein



Knowning amino acid sequence



Existing methods

- **Experimental methods :**
  - X-ray cristallography
  - Nuclear magnetic resonance spectroscopy
  - Cryomicroscopy

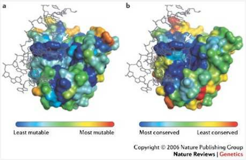
↬ Expensive and slow, but provide the exact 3D structure

- **Computational methods :**
  - Based on homologies with proteins of known structure : methods based on sequence similarity, protein threading
  - *De novo* prediction

↬ Gives several possible 3D structures, with no criterion of choice

Purpose : build a ranking method based on a phylogenetic stability criterion

- In a 3D structure, amino acids in contact frequently have similar modification tolerances



Nature Reviews | Genetics

- **Criterion :** selective pressure sequence

Data

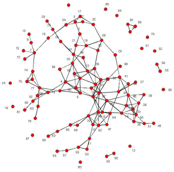
- **First step :** Estimate the selective pressure sequence $\omega_1, \ldots \omega_n$ on a multiple aligment of homologs

Seq :	caa	agg	tgc	tta
H1 :	cat	agg	tgc	gta
H2 :	cat	tgg	tgc	cta
H3 :	aat	tgg	tgc	ctg

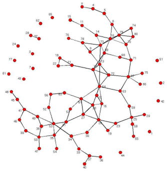
$$\downarrow$$

$$\omega_1 \quad \omega_2 \quad \omega_3 \quad \omega_4$$

- *m* folding candidates



or



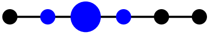
Statistical tools

- Markov random fields
- ABC (Approximate Bayesian Computation)
- Bayesian model choice

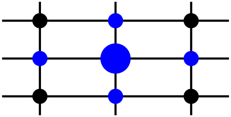
1. Context
2. Markov random fields
3. Parameter posterior distribution
4. Model choice
5. Simulations
6. Conclusion

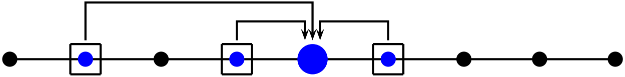
Definition

- Markov chain :



- Markov random field : Markov chain generalisation





Definition (2)

- State at a point i only depends on the state of its neighbours $n(i)$:

$$\pi(x_i = j | x_{-i}) = \pi(x_i = j | z_{n(i)})$$

- Hammersley-Clifford theorem :

$$P(X = x) = \frac{1}{Z} exp(-U(x))$$

with

- $U(x)$: potential

$$U(x) = \sum_{c \in C} V_c(x)$$

$$U(x) = -\theta \sum_{(i,j):i \overset{S}{\sim} j} \mathbf{1}_{\{x_i = x_j\}}$$

- Z : normalizing constant

$$Z = \sum_{x} exp(-U(x))$$

1. Context
2. Markov random fields
3. Parameter posterior distribution
4. Model choice
5. Simulations
6. Conclusion

Bayesian modelisation

- Prior distribution :
  - $(\theta) \sim \pi(\theta)$
- Likelihood :

$(X|\theta) \sim MRF(\theta)$

$$f(x|\theta) = \frac{1}{Z_\theta} exp(\theta \sum_{(i,j):i\sim j} \mathbf{1}_{\{x_i=x_j\}})$$

↬ **Target :** Posterior distribution of θ

Parameter posterior distribution

MCMC methods :

Hastings-Metropolis algorithm :

- Proposal : $\theta' \sim p(\theta'|\theta^{(t)})$
- $\theta^{(t+1)} = \theta'$ with probability

$$min\{1, \frac{\frac{1}{Z_{\theta'}} q_{\theta'}(X)}{\frac{1}{Z_{\theta^{(t)}}} q_{\theta^{(t)}}(X)} \frac{p(\theta^{(t)}|\theta')}{p(\theta'|\theta^{(t)})} \frac{\pi(\theta')}{\pi(\theta^{(t)})}\}$$

$\looparrowright$ Ratio involves intractable normalizing constants $Z_{\theta'}$ and $Z_{\theta^{(t)}}$

ABC : Approximate Bayesian Computation

- Bayesian inference without using likelihood
- **Idea :** Data sufficiently close provide similar parameter posterior distribution
- What we need :
    - **Simulate** data given parameter values
    - **Summary statistics** (sufficient)
    - Calculate closeness between our data (X^0) and simulated data (X^{i^*}) : **distance** between summary statistics

ABC Algorithm

- **Sufficient statistic :** $S(X) = \sum_{(i,j):i\sim j} \mathbf{1}_{\{x_i = x_j\}}$
- **Distance :** $d(S(X^0), S(X^{i*})) = (S(X^0) - S(X^{i*}))^2$
- **Algorithm :**
  - Generate $\theta^{i*} \sim \pi(\theta^{i*})$
  - Generate $(X|\theta^{i*}) \sim MRF(\theta^{i*})$
  - Calculate $d_i = d(S(X^0), S(X^{i*}))$
  - Accept θ^{i*} if $d_i < \varepsilon$
- **Result :** sample of independent draws from $f(\theta|d < \varepsilon)$
  $\looparrowright$ Good approximation of $f(\theta|X^0)$
- In practice, ε is a 1% quantile of d

1. Context
2. Markov random fields
3. Parameter posterior distribution
4. Model choice
5. Simulations
6. Conclusion

Bayesian hierarchical modelisation

1 model $\longleftrightarrow$ 1 neighborhood / 3D structure

- Prior distributions :
  - $s \sim \pi(s)$
  - $(\theta_s | s) \sim \pi_s(\theta_s)$
- Likelihood :

  $(X | \theta_s, s) \sim MRF(\theta_s, s)$

$$f_s(x|\theta_s) = \frac{1}{Z_{\theta_s,s}} exp(\theta_s \sum_{(i,j):i \overset{s}{\sim} j} \mathbf{1}_{\{x_i = x_j\}})$$

Bayes factor definition

-

$$BF_{0/1} = \frac{\frac{P(s=0|X)}{P(s=1|X)}}{\frac{P(s=0)}{P(s=1)}} = \frac{\int f_0(X|\theta_0)\pi_0(\theta_0)d\theta_0}{\int f_1(X|\theta_1)\pi_1(\theta_1)d\theta_1}$$

- **Interpretation :**
  - $BF > 1$: Model 0
  - $BF < 1$: Model 1
  - Jeffreys scale :

$< 10^{-2}$ $> 10^{2}$	$[10^{-2}, 10^{-3/2}]$ $[10^{3/2}, 10^{2}]$	$[10^{-3/2}, 10^{-1}]$ $[10^{1}, 10^{3/2}]$	$[10^{-1}, 10^{-1/2}]$ $[10^{1/2}, 10^{1}]$
decisive	very hard	hard	substantial

Another way to write the Bayes factor

$$
\begin{aligned}
P(S_i(X) = s|\theta_i) &= \sum_{X : S_i(X) = s} f_i(X|\theta_i) \\
&= \frac{1}{Z_{\theta_i, i}} exp(\theta_i \; s) \; card\{X : S_i(X) = s\}
\end{aligned}
$$

$$
BF_{0/1} = \frac{card\{X : S_1(X) = s_1\}}{card\{X : S_0(X) = s_0\}} \frac{\int P(S_0(X) = s_0|\theta_0)\pi_0(\theta_0)d\theta_0}{\int P(S_1(X) = s_1|\theta_1)\pi_1(\theta_1)d\theta_1}
$$

ABC algorithm

- **Vector of summary statistics :** $S(X) = S_1(X), ..S_m(X)$ avec $S_s(X) = \sum_{(i,j):i\overset{s}{\sim}j} \mathbf{1}_{\{x_i=x_j\}}$
- **Distance :** $d(S(X^0), S(X^{i*})) = \sum_s (S_s(X^0) - S_s(X^{i*}))^2$
- **Algorithm :**
  - Generate $s^{i*} \sim \pi(s)$
  - Generate $(\theta_{s^{i*}} | s^{i*}) \sim \pi_{s^{i*}}(\theta^*_{s^{i*}})$
  - Generate $(X | \theta^*_{s^{i*},i*}) \sim MRF(\theta^*_{s^{i*},i*})$
  - Calculate $d_i = d(S(X^0), S(X^{i*}))$
  - Accept $(s_{i*}, \theta^*_{s_{i*}})$ if $d_i < \varepsilon$

- Result : $((s_{i*}, \theta^*_{s_{i*}})_i)$

$$\looparrowright \frac{card(s_{i*} = 0)}{card(s_{i*} = 1)} \text{ estimate of } \frac{\int P(S_0(X) = s_0|\theta_0)\pi_0(\theta_0)d\theta_0}{\int P(S_1(X) = s_1|\theta_1)\pi_1(\theta_1)d\theta_1}$$

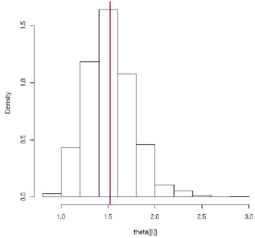
- Calculate $\dfrac{card\{X : S_1(X) = s_1\}}{card\{X : S_0(X) = s_0\}}$ to obtain $\widehat{BF}$

1. Context
2. Markov random fields
3. Parameter posterior distribution
4. Model choice
5. Simulations
6. Conclusion

Models

- **M0 : iid case,** $Bernouilli(p)$
  - $f_0(X|\theta, m=0) = \frac{1}{Z_{\theta,0}} exp(\theta \sum_i \mathbf{1}_{\{x_i=1\}})$
  - $p = \frac{exp(\theta)}{1+exp(\theta)}$
- **M1 : Markov chain with transition matrix** P
  - $f(X|\theta,1) = \frac{exp(2\theta \; \sum_{i=1}^{n-1} \mathbf{1}_{\{x_i=x_{i+1}\}})}{2(1+exp(2\theta))^{n-1}}$
  - $P = \begin{pmatrix} \frac{exp(2\theta)}{1+exp(2\theta)} & \frac{1}{1+exp(2\theta)} \\ \frac{1}{1+exp(2\theta)} & \frac{exp(2\theta)}{1+exp(2\theta)} \end{pmatrix}$

Parameter estimation

- 

- Comparison ML/ABC estimates : $|\widehat{\theta}_{abc} - \widehat{\theta}_{mv}|$

$1stQu.$	$Median$	$Mean$	$3rdQu.$	Var
$1.68e-03$	$3.28e-03$	$3.27e-03$	$5.001e-03$	$3.89e-06$

Model choice

- $$\widehat{BF} = \frac{C_{n-1}^{s_1}}{C_n^{s_0}} \frac{card(s_{i*}=0)}{card(s_{i*}=1)}$$
- $$BF = \frac{\int \frac{exp(\theta_0 S_0(X))}{(1+exp(\theta_0))^n}\pi_0(\theta_0)d\theta_0}{\int \frac{exp(2\theta_1 S_1(X))}{(1+exp(2\theta_1))^{n-1}}\pi_1(\theta_1)d\theta_1}$$
- Comparison on 10.000 simulated data :

		BF		
		$M0$	?	$M1$
$\widehat{BF}$	$M0$	4684	2	30
	?	158	339	53
	$M1$	5	261	4464

1. Context

2. Markov random fields

3. Parameter posterior distribution

4. Model choice

5. Simulations

6. Conclusion

Conclusion

- **Conclusion**
  - Parameter estimation in MRF is not too expensive using ABC
  - BF is a good way to choose between some neighborhoods
- **Perspectives :**
  - Estimate / calculate the number of configurations given a value of S
  - Apply on biological data

INRA

MIG Mathématique Informatique & Génome

UNIVERSITE PARIS DAUPHINE

SSB tatistics for ystems iology

îledeFrance