Recherche de régions génomiques conservées

Modèle mathématique et test statistique

S. Grusea, E. Pardoux[1] V. Lopez Rascol, P. Pontarotti[2]

[1]L.A.T.P., Centre de Mathématique et d'Informatique
Université de Provence

[2]Laboratoire de Phylogénomique EA 3781 Évolution Biologique
Université de Provence

ALPHY/GTGC 23-24 janvier 2006

Régions génomiques conservées

Intérêt :

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Région conservée (cluster conservé de gènes) :

deux régions chromosomiques ayant un certain nombre de gènes orthologues en commun.

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Région conservée (cluster conservé de gènes) :

deux régions chromosomiques ayant un certain nombre de gènes orthologues en commun.

Recherche de régions conservées :

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Région conservée (cluster conservé de gènes) :

deux régions chromosomiques ayant un certain nombre de gènes orthologues en commun.

Recherche de régions conservées :

- Chez une espèce A : région chromosomique fixée (région de référence)

Régions génomiques conservées

Intérêt : inférer

- des relations évolutives entre espèces
- des gènes sous une pression fonctionnelle sélective.

Région conservée (cluster conservé de gènes) :

deux régions chromosomiques ayant un certain nombre de gènes orthologues en commun.

Recherche de régions conservées :

- Chez une espèce A : région chromosomique fixée (région de référence)
- On cherche chez une autre espèce B des régions génomiques similaires.

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.
3. Tester si elles sont **significatives** (vraiment 'conservées').

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.
3. Tester si elles sont **significatives** (vraiment 'conservées').

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.
3. Tester si elles sont **significatives** (vraiment 'conservées').

Région conservée significative :

C'est très peu probable qu'elle soit apparue par hasard.

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.
3. Tester si elles sont **significatives** (vraiment 'conservées').

Région conservée significative :

C'est très peu probable qu'elle soit apparue par hasard.

Test statistique

- adapté à l'approche de type région de référence

Recherche de régions conservées

1. Trouver les orthologues des gènes de la région de référence dans l'espèce B.
2. Chercher les régions avec une grande densité de gènes orthologues.
3. Tester si elles sont **significatives** (vraiment 'conservées').

Région conservée significative :

C'est très peu probable qu'elle soit apparue par hasard.

Test statistique

- adapté à l'approche de type région de référence
- qui prend en compte l'existence de familles de gènes orthologues.

L'idée du test statistique

! Région conservée entre les deux espèces.

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

Hypothèse H_0 = hypothèse de génome aléatoire (hasard) :

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

Hypothèse H_0 = hypothèse de génome aléatoire (hasard) :

- calculer, sous cette hypothèse, la probabilité de trouver une telle région dans le génome B

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

Hypothèse H_0 = hypothèse de génome aléatoire (hasard) :

- calculer, sous cette hypothèse, la probabilité de trouver une telle région dans le génome B
- si suffisamment petite

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

Hypothèse H_0 = hypothèse de génome aléatoire (hasard) :

- calculer, sous cette hypothèse, la probabilité de trouver une telle région dans le génome B
- si suffisamment petite

L'idée du test statistique

! Région conservée entre les deux espèces.

? Est-elle significative ?

Hypothèse H_0 = hypothèse de génome aléatoire (hasard) :

- calculer, sous cette hypothèse, la probabilité de trouver une telle région dans le génome B
- si suffisamment petite

  $\Longrightarrow$ région conservée **significative** !

Formulation mathématique

Un génome = séquence ordonnée de gènes, sans séparation en chromosomes.

Formulation mathématique

Un génome = séquence ordonnée de gènes, sans séparation en chromosomes.

La longueur d'une région génomique = le nombre de gènes.

Formulation mathématique

Un génome = séquence ordonnée de gènes, sans séparation en chromosomes.

La longueur d'une région génomique = le nombre de gènes.

Les données :

- m gènes dans la région de référence du génome A qui ont des orthologues dans le génome B

Formulation mathématique

Un génome = séquence ordonnée de gènes, sans séparation en chromosomes.

La longueur d'une région génomique = le nombre de gènes.

Les données :

- m gènes dans la région de référence du génome A qui ont des orthologues dans le génome B
- ϕ_i, $i = 1, ..., m$ les tailles des familles de gènes orthologues

Formulation mathématique

Un génome = séquence ordonnée de gènes, sans séparation en chromosomes.

La longueur d'une région génomique = le nombre de gènes.

Les données :

- m gènes dans la région de référence du génome A qui ont des orthologues dans le génome B
- ϕ_i, $i = 1, ..., m$ les tailles des familles de gènes orthologues
- N la taille du génome de l'espèce B.

Prendre en compte les familles de gènes

Un dénombrement un peu différent :

Nombre d'orthologues dans une région

Dans une certaine région de B :

n_i orthologues de la i - ème famille, $i = 1, ..., m$

Prendre en compte les familles de gènes

Un dénombrement un peu différent :

Nombre d'orthologues dans une région

Dans une certaine région de B :

n_i orthologues de la i - ème famille, $i = 1, ..., m$

$\Longrightarrow$ le 'nombre' d'orthologues dans cette région :

$$h = \sum_{i=1}^{m} \frac{n_i}{\phi_i}.$$

Le test statistique (I)

Définitions :

Pour chaque h possible :

- L_h = la longueur de la plus petite région dans B contenant h orthologues.

Le test statistique (I)

Définitions :

Pour chaque h possible :

- L_h = la longueur de la plus petite région dans B contenant h orthologues.
- $r_\beta(h)$ = la longueur critique pour qu'un cluster avec h orthologues soit significatif au seuil β :

$$\mathbb{P}(L_h \leq r_\beta(h)) \simeq \beta.$$

Le test statistique (II)

On cherche un cluster significatif.

Le test statistique (II)

On cherche un cluster significatif.

On ne connaît pas a priori le nombre h d'orthologues !

Le test statistique (II)

On cherche un cluster significatif.

On ne connaît pas a priori le nombre h d'orthologues !

Le test :

1. Pour β fixé, trouver pour tous les h possibles les longueurs critiques correspondantes $r_\beta(h)$:

$$\mathbb{P}(L_h \leq r_\beta(h)) \simeq \beta.$$

Le test statistique (II)

On cherche un cluster significatif.

On ne connaît pas a priori le nombre h d'orthologues !

Le test :

1. Pour β fixé, trouver pour tous les h possibles les longueurs critiques correspondantes $r_\beta(h)$:

$$\mathbb{P}(L_h \leq r_\beta(h)) \simeq \beta.$$

2. Trouver le bon β tel que

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Estimation par Monte Carlo

La loi exacte de L_h :

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

La méthode de Monte Carlo :

On approche une moyenne théorique par la moyenne empirique.

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

La méthode de Monte Carlo :

On approche une moyenne théorique par la moyenne empirique.

On approche une probabilité par la fréquence empirique.

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

La méthode de Monte Carlo :

On approche une moyenne théorique par la moyenne empirique.

On approche une probabilité par la fréquence empirique.

- Pour chaque h, on estime la fonction de répartition de L_h :

$$F(r) = \mathbb{P}(L_h \leq r) \simeq \frac{|\{i = 1, ..., M : l_{h,i} \leq r\}|}{M}$$

$l_{h,i}, i = 1, ..., M$ simulations indép. de L_h, avec M suffisamment grand.

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

La méthode de Monte Carlo :

On approche une moyenne théorique par la moyenne empirique.

On approche une probabilité par la fréquence empirique.

- Pour chaque h, on estime la fonction de répartition de L_h :

$$F(r) = \mathbb{P}(L_h \leq r) \simeq \frac{|\{i = 1, ..., M : l_{h,i} \leq r\}|}{M}$$

  $l_{h,i}, i = 1, ..., M$ simulations indép. de L_h, avec M suffisamment grand.
- Pour β fixé, on calcule $r_\beta(h)$ pour chaque h.

Estimation par Monte Carlo

La loi exacte de L_h : difficile à obtenir.

La méthode de Monte Carlo :

On approche une moyenne théorique par la moyenne empirique.

On approche une probabilité par la fréquence empirique.

- Pour chaque h, on estime la fonction de répartition de L_h :

$$F(r) = \mathbb{P}(L_h \leq r) \simeq \frac{|\{i = 1, ..., M : l_{h,i} \leq r\}|}{M}$$

  $l_{h,i}, i = 1, ..., M$ simulations indép. de L_h, avec M suffisamment grand.
- Pour β fixé, on calcule $r_\beta(h)$ pour chaque h.
- De façon analogue, on trouve le bon β.

Modèle mathématique simplifié

- Le génome B = l'intervalle continu [0,1]

Modèle mathématique simplifié

- Le génome B = l'intervalle continu [0,1]
- Sous l'hypohèse de génome aléatoire :
  les positions dans B des gènes orthologues = v.a. indép. unif. distr. sur [0,1].

Modèle mathématique simplifié

- Le génome B = l'intervalle continu [0,1]
- Sous l'hypohèse de génome aléatoire :
  les positions dans B des gènes orthologues = v.a. indép. unif. distr. sur [0,1].
- On remplace h par $\frac{h}{m}$, donc $h \in [0, 1]$ aussi.

Le cas sans familles de gènes ($\phi_i = 1$, $\forall i$)

- Mesure de comptage renormalisée

$$\mu_m(I) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \in I\}} = \frac{|\{i : X_i \in I\}|}{m},$$

$X_i, i = 1, ..., m$ des v.a. indép. unif. distr. sur [0,1].

Le cas sans familles de gènes ($\phi_i = 1$, $\forall i$)

- Mesure de comptage renormalisée

$$\mu_m(I) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \in I\}} = \frac{|\{i : X_i \in I\}|}{m},$$

$X_i, i = 1, ..., m$ des v.a. indép. unif. distr. sur [0,1].

- Donc

$$L_h = \inf\{|I| : \mu_m(I) \geq h\}.$$

Le cas sans familles de gènes ($\phi_i = 1$, $\forall i$)

- Mesure de comptage renormalisée

$$\mu_m(I) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \in I\}} = \frac{|\{i : X_i \in I\}|}{m},$$

$X_i, i = 1, ..., m$ des v.a. indép. unif. distr. sur [0,1].

- Donc

$$L_h = \inf\{|I| : \mu_m(I) \geq h\}.$$

Le cas sans familles de gènes ($\phi_i = 1$, $\forall i$)

- Mesure de comptage renormalisée

$$\mu_m(I) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \in I\}} = \frac{|\{i : X_i \in I\}|}{m},$$

$X_i, i = 1, ..., m$ des v.a. indép. unif. distr. sur [0,1].

- Donc

$$L_h = \inf\{|I| : \mu_m(I) \geq h\}.$$

Problème

Étudier la loi de L_h.

Le cas sans familles de gènes ($\phi_i = 1$, $\forall i$)

- Mesure de comptage renormalisée

$$\mu_m(I) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \in I\}} = \frac{|\{i : X_i \in I\}|}{m},$$

$X_i, i = 1, ..., m$ des v.a. indép. unif. distr. sur [0,1].

- Donc

$$L_h = \inf\{|I| : \mu_m(I) \geq h\}.$$

Problème

Étudier la loi de L_h.

Même dans ce cas, difficile à obtenir.

Le comportement asymptotique de L_h $(m \to \infty)$

La fonction de répartition de μ_m :

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \leq t\}}.$$

Le comportement asymptotique de L_h $(m \to \infty)$

La fonction de répartition de μ_m :

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \leq t\}}.$$

- Loi des grands nombres :

$$F_m(t) \underset{m\to\infty}{\longrightarrow} t, \text{ uniformément en } t \in [0,1], \text{ p.s..}$$

Le comportement asymptotique de L_h $(m \to \infty)$

La fonction de répartition de μ_m :

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{X_i \leq t\}}.$$

- Loi des grands nombres :

$$F_m(t) \underset{m\to\infty}{\longrightarrow} t, \text{ uniformément en } t \in [0,1], \text{ p.s..}$$

- Théorème Limite Centrale :

$$\sqrt{m}(F_m(t) - t) \underset{m\to\infty}{\longrightarrow} B_t, \ t \in [0,1]$$

convergence en loi fonctionnelle vers le pont brownien.

Approximation gaussienne

On obtient :

$$\mathbb{P}(L_h \leq r) \gtrsim \mathbb{P}(\bigcup_i A_i),$$

où les A_i impliquent des v.a. i.i.d. gaussiennes.

La formule de Chung - Erdös

$$\mathbb{P}(\bigcup_i A_i) \geq \frac{(\sum_i \mathbb{P}(A_i))^2}{\sum_i \mathbb{P}(A_i) + \sum_{i \neq j} \mathbb{P}(A_i \cap A_j)}.$$

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Généralisation pour le cas avec des familles de gènes.

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Généralisation pour le cas avec des familles de gènes.

Perspectives :

- Prendre en compte l'ordre de gènes.

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Généralisation pour le cas avec des familles de gènes.

Perspectives :

- Prendre en compte l'ordre de gènes.
- Reconstruire les génomes ancestraux.

Conclusion

- Pour β fixé, on calcule $r_\beta(h)$ pour tout h.
- Algorithme stochastique (Robbins - Monro) pour trouver le β t.q.

$$\mathbb{P}(\text{il existe } h \text{ t.q. } L_h \leq r_\beta(h)) \simeq \alpha \ (0,01).$$

Généralisation pour le cas avec des familles de gènes.

Perspectives :

- Prendre en compte l'ordre de gènes.
- Reconstruire les génomes ancestraux.
- ...