ÉTUDE DE LA CONVERSION GÉNIQUE CHEZ LES FAMILLES MULTIGÉNIQUES

Frédéric Prat

Thèse soumise à
l'École des études supérieures et de la recherche
Université d'Ottawa
en vue de l'obtention de la maîtrise ès sciences à
l'Institut de Biologie d'Ottawa-Carleton

Frédéric Prat
Université d'Ottawa
1997

ISBN 0-612-20008-6

Nom Frédéric PRAT

Dissertation Abstracts International et *Masters Abstracts International* sont organisés en catégories de sujets. Veuillez s.v.p. choisir le sujet qui décrit le mieux votre thèse et inscrivez le code numérique approprié dans l'espace réservé ci-dessous.

Biologie moléculaire
SUJET

0307
CODE DE SUJET

UMI

Catégories par sujets

HUMANITÉS ET SCIENCES SOCIALES

COMMUNICATIONS ET LES ARTS
Architecture 0729
Beaux-arts 0357
Bibliothéconomie 0399
Cinéma 0900
Communication verbale 0459
Communications 0708
Danse 0378
Histoire de l'art 0377
Journalisme 0391
Musique 0413
Sciences de l'information 0723
Théâtre 0465

ÉDUCATION
Généralités 515
Administration 0514
Art 0273
Collèges communautaires 0275
Commerce 0688
Économie domestique 0278
Éducation permanente 0516
Éducation préscolaire 0518
Éducation sanitaire 0680
Enseignement agricole 0517
Enseignement bilingue et multiculturel 0282
Enseignement industriel 0521
Enseignement primaire 0524
Enseignement professionnel 0747
Enseignement religieux 0527
Enseignement secondaire 0533
Enseignement spécial 0529
Enseignement supérieur 0745
Évaluation 0288
Finances 0277
Formation des enseignants 0530
Histoire de l'éducation 0520
Langues et littérature 0279
Lecture 0535
Mathématiques 0280
Musique 0522
Orientation et consultation 0519
Philosophie de l'éducation 0998
Physique 0523
Programmes d'études et enseignement 0727
Psychologie 0525
Sciences 0714
Sciences sociales 0534
Sociologie de l'éducation 0340
Technologie 0710

LANGUE, LITTÉRATURE ET LINGUISTIQUE
Langues
- Généralités 0679
- Anciennes 0289
- Linguistique 0290
- Modernes 0291

Littérature
- Généralités 0401
- Anciennes 0294
- Comparée 0295
- Médiévale 0297
- Moderne 0298
- Africaine 0316
- Américaine 0591
- Anglaise 0593
- Asiatique 0305
- Canadienne (Anglaise) 0352
- Canadienne (Française) 0355
- Germanique 0311
- Latino-américaine 0312
- Moyen-orientale 0315
- Romane 0313
- Slave et est-européenne 0314

PHILOSOPHIE, RELIGION ET THÉOLOGIE
Philosophie 0422
Religion
- Généralités 0318
- Clergé 0319
- Études bibliques 0321
- Histoire des religions 0320
- Philosophie de la religion 0322

Théologie 0469

SCIENCES SOCIALES
Anthropologie
- Archéologie 0324
- Culturelle 0326
- Physique 0327

Droit 0398
Économie
- Généralités 0501
- Commerce-Affaires 0505
- Économie agricole 0503
- Économie du travail 0510
- Finances 0508
- Histoire 0509
- Théorie 0511

Études américaines 0323
Études canadiennes 0385
Études féministes 0453
Folklore 0358
Géographie 0366
Gérontologie 0351
Gestion des affaires
- Généralités 0310
- Administration 0454
- Banques 0770
- Comptabilité 0272
- Marketing 0338

Histoire
- Histoire générale 0578
- Ancienne 0579
- Médiévale 0581
- Moderne 0582
- Histoire des noirs 0328
- Africaine 0331
- Canadienne 0334
- États-Unis 0337
- Européenne 0335
- Moyen-orientale 0333
- Latino-américaine 0336
- Asie, Australie et Océanie 0332

Histoire des sciences 0585
Loisirs 0814
Planification urbaine et régionale 0999
Science politique
- Généralités 0615
- Administration publique 0617
- Droit et relations internationales 0616

Sociologie
- Généralités 0626
- Aide et bien-être social 0630
- Criminologie et établissements pénitentiaires 0627
- Démographie 0938
- Études de l'individu et de la famille 0628
- Études des relations interethniques et des relations raciales 0631
- Structure et développement social 0700
- Théorie et méthodes 0344
- Travail et relations industrielles 0629

Transports 0709
Travail social 0452

SCIENCES ET INGÉNIERIE

SCIENCES BIOLOGIQUES
Agriculture
- Généralités 0473
- Agronomie 0285
- Alimentation et technologie alimentaire 0359
- Culture 0479
- Élevage et alimentation 0475
- Exploitation des pâturages 0777
- Pathologie animale 0476
- Pathologie végétale 0480
- Physiologie végétale 0817
- Sylviculture et faune 0478
- Technologie du bois 0746

Biologie
- Généralités 0306
- Anatomie 0287
- Biologie (Statistiques) 0308
- Biologie moléculaire 0307
- Botanique 0309
- Cellule 0379
- Écologie 0329
- Entomologie 0353
- Génétique 0369
- Limnologie 0793
- Microbiologie 0410
- Neurologie 0317
- Océanographie 0416
- Physiologie 0433
- Radiation 0821
- Science vétérinaire 0778
- Zoologie 0472

Biophysique
- Généralités 0786
- Médicale 0760

SCIENCES DE LA TERRE
Biogéochimie 0425
Géochimie 0996
Géodésie 0370
Géographie physique 0368
Géologie 0372
Géophysique 0373
Hydrologie 0388
Minéralogie 0411
Océanographie physique 0415
Paléobotanique 0345
Paléoécologie 0426
Paléontologie 0418
Paléozoologie 0985
Palynologie 0427

SCIENCES DE LA SANTÉ ET DE L'ENVIRONNEMENT
Économie domestique 0386
Sciences de l'environnement 0768
Sciences de la santé
- Généralités 0566
- Administration des hôpitaux 0769
- Alimentation et nutrition 0570
- Audiologie 0300
- Chimiothérapie 0992
- Dentisterie 0567
- Développement humain 0758
- Enseignement 0350
- Immunologie 0982
- Loisirs 0575
- Médecine du travail et thérapie 0354
- Médecine et chirurgie 0564
- Obstétrique et gynécologie 0380
- Ophtalmologie 0381
- Orthophonie 0460
- Pathologie 0571
- Pharmacie 0572
- Pharmacologie 0419
- Physiothérapie 0382
- Radiologie 0574
- Santé mentale 0347
- Santé publique 0573
- Soins infirmiers 0569
- Toxicologie 0383

SCIENCES PHYSIQUES
Sciences Pures
Chimie
- Généralités 0485
- Biochimie 487
- Chimie agricole 0749
- Chimie analytique 0486
- Chimie minérale 0488
- Chimie nucléaire 0738
- Chimie organique 0490
- Chimie pharmaceutique 0491
- Physique 0494
- Polymères 0495
- Radiation 0754

Mathématiques 0405
Physique
- Généralités 0605
- Acoustique 0986
- Astronomie et astrophysique 0606
- Électronique et électricité 0607
- Fluides et plasma 0759
- Météorologie 0608
- Optique 0752
- Particules (Physique nucléaire) 0798
- Physique atomique 0748
- Physique de l'état solide 0611
- Physique moléculaire 0609
- Physique nucléaire 0610
- Radiation 0756

Statistiques 0463

Sciences Appliqués Et Technologie
Informatique 0984
Ingénierie
- Généralités 0537
- Agricole 0539
- Automobile 0540
- Biomédicale 0541
- Chaleur et thermodynamique 0348
- Conditionnement (Emballage) 0549
- Génie aérospatial 0538
- Génie chimique 0542
- Génie civil 0543
- Génie électronique et électrique 0544
- Génie industriel 0546
- Génie mécanique 0548
- Génie nucléaire 0552
- Ingénierie des systèmes 0790
- Mécanique navale 0547
- Métallurgie 0743
- Science des matériaux 0794
- Technique du pétrole 0765
- Technique minière 0551
- Techniques sanitaires et municipales 0554
- Technologie hydraulique 0545

Mécanique appliquée 0346
Géotechnologie 0428
Matières plastiques (Technologie) 0795
Recherche opérationnelle 0796
Textiles et tissus (Technologie) 0794

PSYCHOLOGIE
Généralités 0621
Personnalité 0625
Psychobiologie 0349
Psychologie clinique 0622
Psychologie du comportement 0384
Psychologie du développement 0620
Psychologie expérimentale 0623
Psychologie industrielle 0624
Psychologie physiologique 0989
Psychologie sociale 0451
Psychométrie 0632

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

TABLE DES MATIÈRES

Table des matières ..... p.i

Résumé ..... p.iv

Abstract ..... p.v

Liste des tableaux ..... p.vi

Liste des figures ..... p.vii

CHAPITRE I (Introduction) p.1

A - LES PHÉNOMÈNES BIOLOGIQUES p.1

1) Mécanismes de la conversion génique p.4

2) Classification des conversions géniques p.8

3) Effets de la conversion génique p.9

B - MÉTHODES DE DÉTECTION p.14

1) Méthodes de compatibilité p.15

2) Méthodes de simulation p.19

C - OBJECTIFS À ATTEINDRE p.21

CHAPITRE II MÉTHODES p.23

A - ALIGNEMENT DES SÉQUENCES p.23

B- MÉTHODES STATISTIQUES p.27

1) Principes généraux p.27

2) Méthode de Sawyer p.29

3) Méthodes des codoubles p.36

a) Méthode de Balding et al p.36

b) Méthode des taux variables p.45

4) Méthode des densités p.48

C - MÉTHODE DE COMPATIBILITÉ p.51

1) Analyse site par site p.51

a) Cas déterminés p.52

b) Cas indéterminés p.56

CHAPITRE III ÉTUDE DES DIFFÉRENTES FAMILLES MULTIGÉNIQUES p.60

A - GROUPE DES GÈNES D'ACTINES VÉGÉTALES p.60

1) Phylogénie des gènes d'actines étudiés p.62

2) Mesures statistiques p.65

B - GROUPE DES GÈNES DE GLOBINES p.73

1) Phylogénie des gènes de globines p.75

2) Mesures statistiques p.78

C - GROUPE DES GÈNES Zfx Zfy p.83

1) Rôle des gènes Zfx Zfy p.85

2) Caractéristiques des gènes de type Zfx Zfy p.87

3) Phylogénie des gènes de type Zfx Zfy p.88

4) Mesures statistiques p.90

5) Analyse site par site p.95

D - LIMITES ET FIABILITÉ DES MÉTHODES EMPLOYÉES p.115

1) Limites des méthodes employées p.115

2) Fiabilité des méthodes employées p.117

CONCLUSION p.120

RÉFÉRENCES p.123

RÉSUMÉ

Trois familles multigéniques ont été étudiées du point de vue de la conversion génique : un groupe de gènes d'actines végétales provenant de cinq espèces d'angiospermes (maïs, tomate, pomme de terre, soja et tabac), un groupe de gènes de globines (les gènes δ et β de l'humain et du tarsier, les gènes γ_1 et γ_2 de l'humain et de l'orang-outang) et un groupe de gènes de doigts à zinc (Zfx et Zfy),liés aux chromosomes sexuels de sept espèces de mammifères (rat, souris, renard, singe écureuil, babouin, orang-outang et humain). L'analyse fut effectuée en utilisant quatre méthodes de simulation et une méthode de compatibilité. La méthode de compatibilité est une variante de la méthode de Fitch et al. (1990). Les méthodes de simulation comprennent la méthode de Sawyer (1989) et la méthode des codoubles (Balding et al. 1992). Pour prendre en compte le fait que des chromosomes différents peuvent avoir des taux de mutation différents une variante de la méthode des codoubles est présentée (méthode des taux variables). Une méthode permettant de détecter les zones de conversion génique est aussi présentée (méthode des densités). Deux conversions géniques ont été détectées chez le maïs. Les résultats obtenus à partir des exons ont été confirmés par l'examen des introns correspondants. Pour les globines, les résultats obtenus confirment la présence de conversions géniques. Ces résultats sont en accord avec les analyses de compatibilité. En ce qui concerne les gènes Zfx/Zfy, une comparaison entre le rat et le singe écureuil donne des résultats significatifs. L'analyse de compatibilité indique que la conversion génique détectée s'est produite probablement dans le groupe des primates peu après la séparation des primates et des rongeurs.

ABSTRACT

Three multigene families have been studied from the point of view of gene conversion. The data sets included a group of plant actin genes from five different angiosperm species (maize, tomato, potato, soybean and tobacco), a group of primate globin genes (the δ and β genes of human and tarsier, the γ_1 and γ_2 genes of human and orangutan) and a group of zinc finger genes (Zfx and Zfy),linked to the sex chromosomes of seven mammalian species (rat, mouse, fox, squirrel monkey, baboon, orangutan, human). The analysis was done by using four simulation methods and one compatibility method. The compatibility method is a variant of Fitch method (Fitch et al.1990). The simulation methods include Sawyer's method (Sawyer 1989) and the codoubles method (Balding et al. 1992). To take into account the fact that different chromosomes may have different mutation rates, a variant of the codoubles method is introduced (the method of variable rates). All these methods are non specific and a method for specifically detecting areas of gene conversion is also introduced. Two conversion events were detected in the maize family. These results based exclusively on exons are confirmed by a high degree of sequence similarity between corresponding introns. The results obtained for the globin genes are in agreement with previous compatibility analyses and confirm the presence of gene conversion in this multigene family. In the case of the Zfx/Zfy genes, gene conversion was detected at a significant level for the rat and/or squirrel monkey lineages. Site by site analysis indicates that the conversion occured probably in the primate lineage, soon after the separation of primates and rodents.

LISTE DES TABLEAUX

Tableau 1. Liste des gènes d'actines végétales utilisés pour cette étude p.24

Tableau 2. Liste des séquences de type Zfx/Zfy p.26

Tableau 3. Résultats du test de Sawyer pour les gènes d'actines p.65

Tableau 4. Résultats du test de Balding et al. pour les gènes d'actines végétales (comparaison Pt101 Pt97/Tb66 Tb71) p.69

Tableau 5. Résultats du test de Balding et al. pour les gènes d'actines végétales (comparaison Tm52 Tm51/Tb93 Tb54) p.70

Tableau 6. Résultats du test de Balding et al. pour les gènes de globines γ_1 et γ_2 (comparaison Humain/Orang-Outang) p.79

Tableau 7. Résultats du test de Balding et al. pour les gènes de globines δ et β (Comparaison Humain/Tarsier) p.80

Tableau 8. Résultats du test des codoubles (méthode des taux variables) pour les gènes Zfx/Zfy Comparaison Singe écureuil/Rat p.93

Tableau 9. Résultats du test des codoubles (méthode des taux variables) pour les gènes Zfx/Zfy Comparaison Orang-outang/Babouin p.94

Tableau 10. Analyse site par site des gènes Zfx/Zfy p.97

LISTE DES FIGURES

Figure 1. Modèle de recombinaison de Holliday.....p.6

Figure 2. Phylogénie des gènes d'actines végétales.....p.63

Figure 3. Phylogénie des gènes Zfx/Zfy (méthode de parcimonie).....p.89

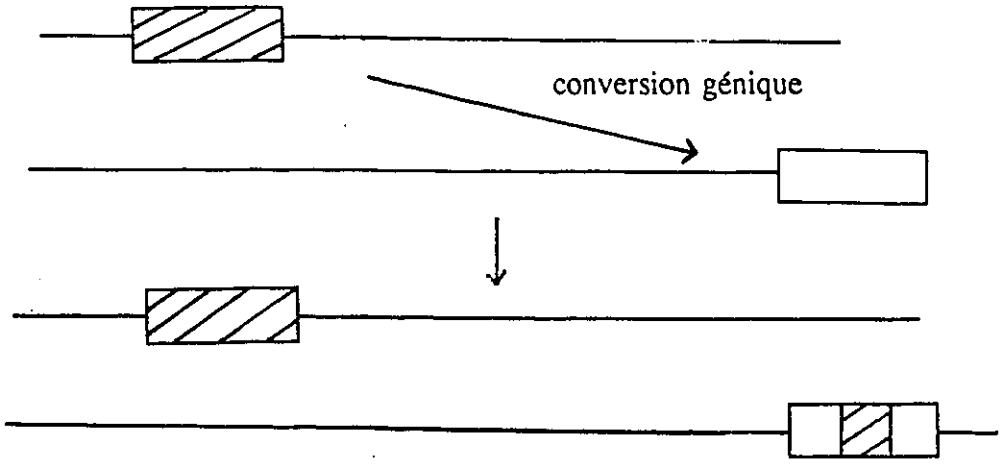
Figure 4. Arbre d'espèce utilisé pour l'analyse site par site.....p.96

CHAPITRE I (Introduction)

A - LES PHÉNOMÈNES BIOLOGIQUES

Le concept de conversion génique fut défini, à l'origine, pour décrire une situation observée chez les levures. Les levures ont une propriété intéressante : au terme de la division méiotique, elle produisent des tétrades. Cette caractéristique permet d'identifier, chez cet organisme, tous les produits de la méiose. En étudiant ces tétrades, on s'aperçut que la ségrégation de certains marqueurs génétiques se faisait de façon non mendélienne. Chez les tétrades, on observait des rapports de type 1:3 ou 3:1 (Jinks-Robertson and Petes. 1993).

Diverses expériences ont montré que la conversion génique est le résultat du transfert d'un brin d'ADN d'un gène à l'autre et de sa réparation. On définit la conversion génique comme un échange non réciproque entre deux séquences d'ADN (Kourilsky 1986). Appelons A et B les deux séquences impliquées dans la conversion et supposons que B est la séquence convertie. Au terme de la conversion, A reste inchangée alors qu'une partie de B devient identique à la partie correspondante de A (voir figure ci-dessous).



Le processus peut s'effectuer dans les deux sens. Le gène converti sera tantôt A, tantôt B. Si l'une des deux polarités est prédominante, il y a fixation de la séquence favorisée sans avoir recours à la sélection naturelle (Walsh 1986). Chez la levure, on a montré que la conversion génique pouvait aussi se produire pendant la mitose. Dans le cas des conversions méiotiques, le phénomène de polarité serait dû à la présence de sites d'initiation particuliers situés en amont d'un gradient de polarité et mis en jeu lors des recombinaisons (Malone et al. 1991).

Même si l'étude du phénomène de conversion génique s'est faite principalement chez les levures, ce phénomène fut aussi détecté chez d'autres espèces. Dans les

années 80, on observa chez les globines de l'homme des similitudes entre les séquences intercalaires IVS2 de deux gènes non homologues situés sur le même chromosome : les gènes γ^A et γ^G. On trouva plus de ressemblances entre les régions 5' de ces séquences non homologues qu'entre les séquences des régions 5' des gènes γ^A homologues. Chez les régions 3', on trouva la situation inverse. Les régions 3' des séquences IVS2 d'allèles se ressemblaient plus que les régions 3' des séquences situées sur le même chromosome. Ces observations furent interprétées comme un cas de conversion génique (Slightom et al. 1980).

On sait maintenant que la conversion génique est un phénomène assez répandu. Des cas de conversion génique ont été décrits chez de nombreuses familles multigéniques appartenant à toutes sortes d'espèces. On peut citer, entre autres, les gènes de globines (Slightom et al. 1980, 1987, 1988; Powers and Smithies 1986; Koop et al. 1989; Fitch et al., 1990; Hayasaka et al. 1992), le complexe majeur d'histocompatibilité (Ohta 1991; Hugues 1995), les immunoglobulines (Thomson 1992), le locus tardif du chorion du Bombyx (Regier et al. 1994), les enzymes P-450 chez le rat (Atchison and Adesnik 1986).

Les expériences destinées à évaluer la taille des conversions géniques on été effectuées essentiellement sur les levures. Une façon d'évaluer la dimension des régions de conversion consiste à examiner les co-conversions de certains marqueurs situés sur le même chromosome. Dans certains cas, on a observé des zones

converties de quelques centaines de nucléotides seulement. Des cas de conversions géniques recouvrant plusieurs gènes ont aussi été décrits (Judd et al. 1988). En incorporant, chez la levure, des gènes murins du complexe majeur d'histocompatibilité et en examinant les tétrades obtenues, on a observé des micro-conversions dont la taille variait entre 1 et 215 nucléotides (Wheeler et al. 1993). À partir de sites de restriction hétérozygotes, situés sur le chromosome III de la levure,on a pu montrer que dans environ 25% des cas observés, les zones de conversion pouvaient atteindre des longueurs supérieures à 5 Kb. Ces zones de conversion étaient généralement d'un seul tenant. Une étude effectuée chez les cellules de mammifères a montré que la conversion d'une région de 5380 bases n'était accompagnée d'aucune mutation. Le processus de conversion semble donc assez fidèle (Stachelek et al. 1988)

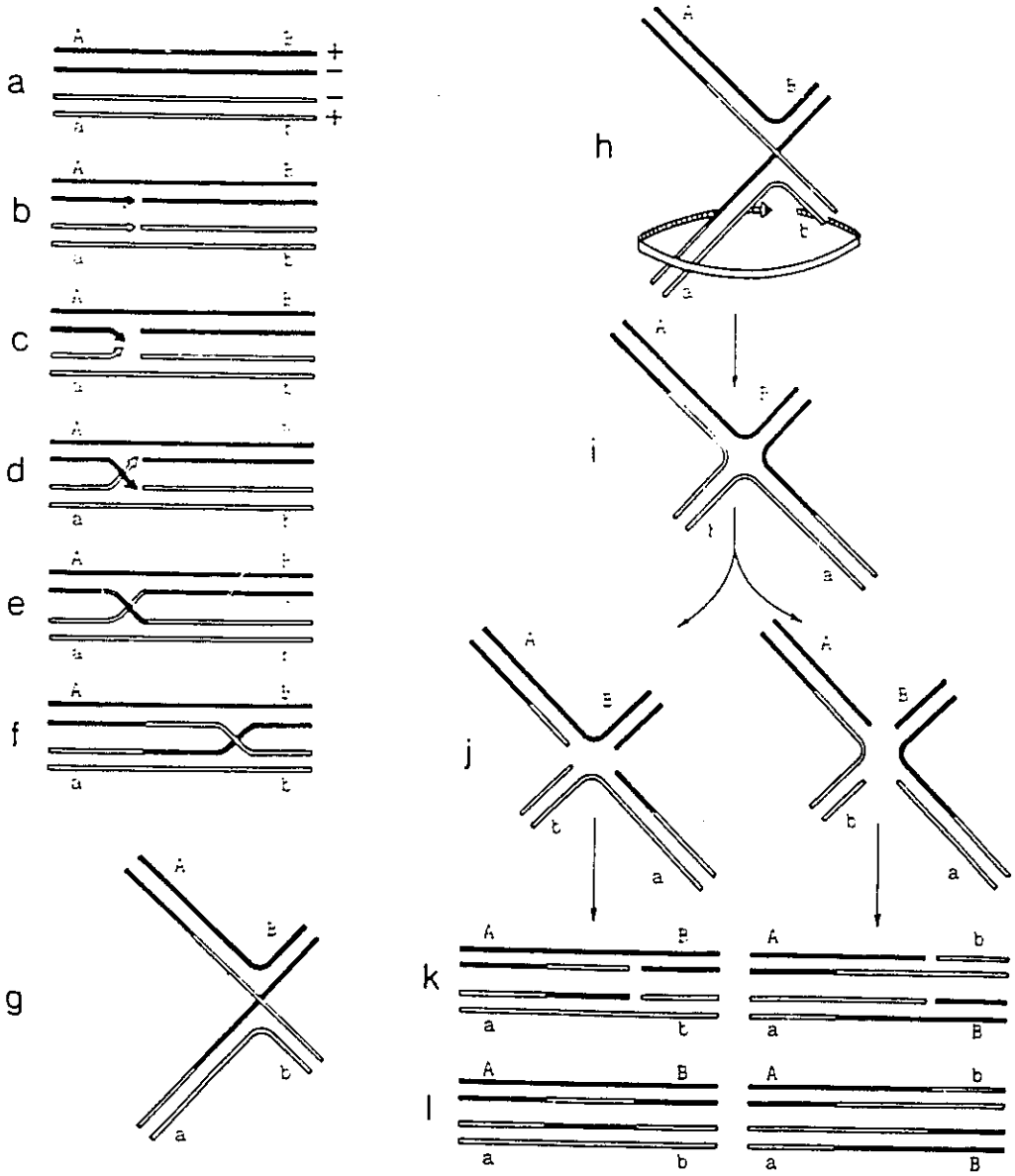
Il y a peu de données sur la fréquence des conversions géniques dans le temps. On peut citer une étude effectuée sur la chaîne β des récepteurs à antigène des lymphocytes T. En comparant plusieurs espèces de souris, on a estimé que des conversions géniques se produisaient en moyenne tous les 0.3 millions d'années (Rudikoff et al. 1992).

1) MÉCANISMES DE LA CONVERSION GÉNIQUE

La conversion génique et, d'une façon générale, les processus de

recombinaison semblent dépendre de l'action de certains systèmes enzymatiques. Pendant la prophase I de la méiose, les chromosomes homologues s'apparient, la synapse est rendue possible par la formation d'un réseau de protéines appelé complexe synaptonémal. Deux types de nodules ont été observés au niveau de ce complexe. On pense qu'il s'agit d'enzymes associés au processus de recombinaison (Engebrecht et al. 1990). Des nodules apparaissent tout d'abord au moment de la formation du complexe synaptonémal. Plus tard, lorsque le complexe est formé, on peut détecter un deuxième type de nodules. Ces nodules tardifs sont moins abondants que les premiers, leur répartition et leur fréquence semblent correspondre à celles des chiasmas, ces structures qui servent à maintenir les chromosomes homologues ensemble et qui correspondent à des recombinaisons (Engebrecht et al. 1990). Les mécanismes exacts de la conversion génique n'ont pas encore été complètement élucidés. Le modèle le plus couramment accepté est celui de Holliday (1974). Ce modèle (et ses variantes) permet d'expliquer assez bien les phénomènes observés. Il est schématisé à la figure 1.

Figure 1. Modèle de recombinaison de Holliday



L'échange se fait entre deux segments d'ADN bicaténaire. Un des brins de chacune des structures en double hélice est incisé (fig. b). Il y a appariement puis raccordement des brins d'ADN. Cette réparation peut être faite par une ligase qui catalyse la formation de liens phosphodiesters (fig. c, d, e). On a montré que la structure ainsi obtenue est stable du point de vue thermodynamique (Dressler and Potter 1985). Par propagation de l'appariement, une zone hétéroduplexe peut être formée (fig. f). On obtient une structure de Holliday. Cette structure peut donner un isomère par rotation (fig. h). Il y a deux façons de découper et de réparer la structure de Holliday (fig. j). Dans un cas, on obtient des recombinants (crossing over), dans l'autre cas (figure de droite), on obtient deux hétéroduplexes et aucun recombinant. La réparation de la zone hétéroduplexe donnera une conversion génique. On peut remarquer que la correction d'une zone hétéroduplexe peut s'effectuer en prenant pour modèle l'un ou l'autre brin. On peut donc s'attendre à obtenir, dans certains cas, des molécules mosaïques. Le modèle de Holliday permet donc d'expliquer des phénomènes comme les micro conversions (Herbomel 1993).

Même si ces modèles sont les plus couramment acceptés, ils n'expliquent pas tous les phénomènes observés. Ils semblerait qu'il existe une relation entre la conversion génique et le phénomène de crossing over. Chez la levure, 30 à 75% des conversions géniques sont associées à des crossing over (Borts and Haber 1989). On ne sait pas encore si les mécanismes moléculaires produisant des conversions associées à des crossing over sont les mêmes que ceux qui produisent seulement des

conversions. Chez Sordaria, on a observé que la région de crossing over pouvait être séparée de la zone de conversion génique par une région contenant des marqueurs dont la ségrégation se faisait de façon mendélienne (Nicolas 1985). Il semblerait que plusieurs types de conversion géniques, obtenus par des mécanismes distincts, soient possibles.

2) CLASSIFICATION DES CONVERSIONS GÉNIQUES

Lorsque les gènes impliqués dans la conversion génique ont des positions équivalentes sur des chromosomes homologues, on parle de recombinaison de type classique ou homotopique (Jinks-Robertson et al. 1993, Herbomel 1993).

Comme les gènes eucaryotes peuvent parfois exister en plusieurs exemplaires disséminés dans l'ensemble du génome, ils peuvent être impliqués dans des recombinaisons. De telles recombinaisons sont appelées ectopiques (Jinks-Robertson et al. 1993, Herbomel 1993).

La conversion génique peut impliquer des gènes situés sur le même chromosome (conversion intra homologue) ou sur des chromosomes homologues (conversion inter homologue). Elle peut aussi mettre en jeu des gènes appartenant à des chromosomes différents et non homologues.

La conversion génique peut se produire aussi pendant la mitose (Aguilera 1988). Chez la levure, la fréquence des recombinaisons est 100 à 1000 fois plus élevée pendant la méiose que pendant la mitose (Roeder et al. 1988).

La présence de conversions géniques a aussi été observée dans le génome des chloroplastes (Bowman et al. 1988, Ogihara et al. 1988, Morton et al. 1993).

On a démontré récemment que des conversions géniques pouvaient se produire entre une copie d'ADN complémentaire et de l'ADN chromosomique (Melamed et al. 1992, Derr and Strathern 1993).

3) EFFETS DE LA CONVERSION GÉNIQUE

Une fois les gènes alignés et les introns enlevés, les données se présentent sous la forme d'une matrice n x m où n représente le nombre de gènes de la famille et m le nombre de nucléotides de chaque séquence d'exons. La conversion génique exerce ses effets de plusieurs façons. Elle affecte, bien sûr, la paire de séquences impliquées dans la conversion. Elle affecte aussi les rapports qui existent entre les différentes lignes de la matrice et l'évaluation des distances phylogénétiques qui les séparent.

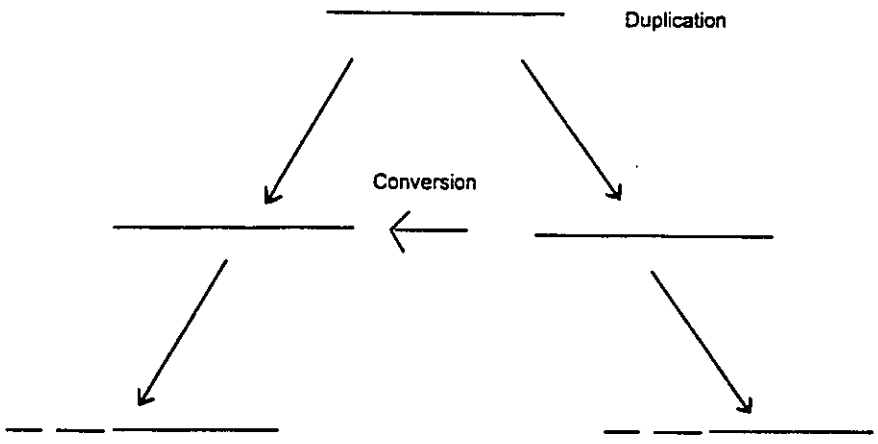
a) Effets sur la paire de séquences impliquées

Appelons densité d'une paire de séquences le rapport suivant :

ρ = nombre de différences entre les deux séquences/nombre total de sites

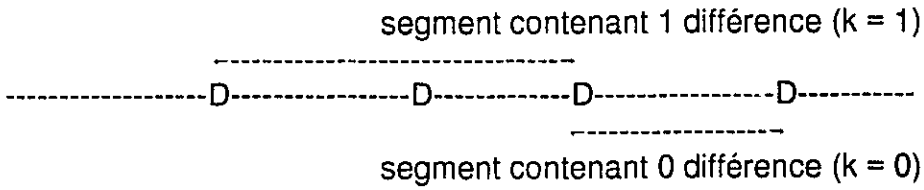
C'est la valeur p du modèle de Jukes Cantor pour les séquences non codantes (Li and Graur 1991).

Considérons deux gènes résultant de la duplication d'un gène ancestral commun. Au fur et à mesure que le temps passe, les mutations s'accumulent et les gènes divergent de plus en plus (voir la figure ci-dessous).



Si, à un moment donné, une conversion génique se produit entre les deux gènes, la zone de conversion aura, au moins pour un temps, une densité nulle. Même si, par la suite, quelques différences s'accumulent sur cette zone, elle conservera un certain retard par rapport au reste du gène. Ce retard peut, toutefois, à la longue, être rattrapé et si la conversion est très ancienne, elle peut devenir indétectable. La conversion génique aura donc pour effet l'apparition de zones à faible densité. Les zones qui auront le plus de chances d'avoir été impliquées dans une conversion génique seront celles qui ont les plus faibles densités. La recherche de ces zones est relativement simple, il suffit de considérer les zones bornées par deux différences.

Lorsque deux séquences sont comparées, les différences observées délimitent des segments de longueurs variables (voir figure ci-dessous).



On peut classer les segments en fonction du nombre de différences qu'ils contiennent. Pour chaque classe correspondant à une valeur de k, il y aura un segment de longueur maximale. On voit que pour chaque classe, le segment le plus long sera celui qui aura la plus faible densité. Pour une paire donnée, la recherche de

la zone la plus susceptible d'avoir subi une conversion génique consistera donc à repérer, pour chaque valeur de k, le segment le plus long.

Effets de la sélection naturelle

Il n'y a pas que la conversion génique qui soit à l'origine de zones à faible densité. Si les deux gènes comparés ont des fonctions analogues, il se peut que du fait de la sélection naturelle, leurs domaines fonctionnels se ressemblent beaucoup. Ces ressemblances se traduiront par des successions d'acides aminés communs. Même si la sélection naturelle n'empêche pas l'apparition de différences synonymes, ces zones auront tendance à avoir une faible densité. Si deux gènes ont à peu près la même fonction, les effets conjugués de la sélection naturelle et de la teneur en GC peuvent avoir pour résultat des zones quasi identiques pour ces deux gènes.

Pour dissocier les effets de la conversion génique de ceux de la sélection naturelle, deux approches sont possibles. Si les n gènes de la famille ont des fonctions analogues, on peut former les n (n-1) / 2 paires de séquences possibles et comparer leurs zones à faible densité. Une autre approche consiste à utiliser comme mesure de densité le rapport Ks = Ms/Ns (Li and al. 1985). Le repérage des zones qui minimisent ce rapport n'est pas aussi facile que dans le cas précédent. On peut balayer la paire avec une fenêtre de longueur variable l (l est un multiple de 3) se déplaçant de trois nucléotides à la fois. Contrairement au cas précédent, les

comparaisons ne se font plus entre régions ayant la même longueur mais entre régions ayant le même nombre de sites synonymes. La valeur Ns d'un segment dépend de sa composition en acides aminés. Cela signifie que deux segments ayant la même longueur n'auront pas forcément la même valeur de Ns et ne seront pas forcément comparables.

On peut aussi mesurer la valeur de Ms/Ns pour chaque paire de codons (cas où l = 3). Ainsi, on obtient une signature pour la paire de séquences. Elle peut être comparée aux signatures obtenues pour les autres paires. Ces signatures peuvent être analysées dans le cadre de la théorie des marches aléatoires.

b) Effets sur l'ensemble de la matrice

On part du principe que les gènes qui forment la matrice sont tous dérivés d'un ancêtre commun. Ces gènes sont organisés en phylogénie. Pour cette raison, certains types de rapports existent entre eux. En plus de créer des zones à faible densité, la conversion génique va affecter les rapports qui existent entre les lignes de la matrice.

A chaque paire de gènes, on peut associer une distance phylogénétique. Si O est l'ancêtre commun le plus récent du gène A et du gène B, la distance phylogénétique entre A et B sera le nombre de substitutions qui se sont effectivement

produites pour passer de O à A et de O à B. Pour n gènes, il y aura n (n-1) / 2 distances phylogénétiques. Si les gènes sont organisés en phylogénie, ces distances devront être cohérentes, elles devront obligatoirement satisfaire certaines relations algébriques. Les distances phylogénétiques ne sont pas connues mais elles peuvent être évaluées à partir des valeurs de densités.

A chaque matrice de données, on peut donc associer un vecteur V à n (n-1) / 2 composantes qui représente l'évaluation des distances phylogénétiques pour toutes les paires de séquences. A chaque matrice de données, on peut aussi associer un vecteur N dont les coordonnées sont les valeurs exactes des distances phylogénétiques entre les séquences de la matrice.

D'une façon générale, la conversion génique aura pour effet de perturber les rapports qui existent entre les lignes de la matrice et leur degré de cohérence. Elle accroît l'erreur d'ajustement qui existe entre le point V et le point N.

B - MÉTHODES DE DÉTECTION

Parmi les méthodes existant actuellement, on peut distinguer deux catégories d'approches très différentes : les méthodes de compatibilité et les méthodes de simulation. Les méthodes de compatibilité privilégient l'approche phylogénétique. Elles consistent essentiellement à interpréter l'information phylogénétique disponible.

Avec ces méthodes, les conversions géniques sont détectées de façon plutôt indirecte. Des conversions géniques seront postulées pour expliquer certaines anomalies, pour résoudre certaines contradictions. Comme elles dépendent des phylogénies, les méthodes de compatibilité nécessitent beaucoup d'information préalable. Les méthodes de simulation privilégient l'approche statistique et ne font pratiquement pas appel aux phylogénies. Elles consistent à simuler des situations qui auraient pu se produire dans le passé. On génère des données hypothétiques en permutant les sites silencieux des données observées. L'avantage de ces méthodes est qu'elles peuvent être utilisées même si l'on dispose de peu d'information sur la famille étudiée.

1) MÉTHODES DE COMPATIBILITÉ

Les premières méthodes proposées consistaient tout simplement à comparer les gènes et à repérer les séquences intergéniques semblables (Slightom et al. 1980). L'information était alors présentée sous forme d'histogramme, chaque barre représentant le nombre de différences observées dans un intervalle de longueur déterminée (généralement 100 nucléotides).

Bien qu'ayant été appliquée avec succès dans quelques cas très particuliers cette méthode s'avéra insuffisante car, comme on l'a déjà mentionné précédemment, des zones de similitudes peuvent être dues à l'effet de la sélection naturelle. À la fin des années 80 il fut nécessaire de mettre au point des méthodes plus sophistiquées.

La plupart des méthodes de compatibilité mettent en jeu plusieurs espèces (méthodes interspécifiques). On peut citer, entre autres, la méthode d'Andersson et al., appliquée au complexe d'histocompatibilité majeure (HMC) (Andersson et al. 1991) et la méthode de Menotti-Raymond et al., appliquée à plusieurs espèces de drosophiles (Menotti-Raymond et al. 1991). Cette méthode fut appliquée à l'étude du locus Adh de Drosophila hydei. Des comparaisons interspécifiques sont effectuées afin de déterminer si les similitudes observées au niveau de ce locus sont dues à des conversions géniques ou à une duplication récente. En prenant en compte l'information disponible quatre scénarios (phylogénies) décrivant l'évolution de D. hydei et de D. mojavensis sont proposés. Ces quatre scénarios peuvent être testés car ils impliquent chacun un degré de divergence différent entre gènes de la même espèce et gènes d'espèces différentes. La mesure des divergences est effectuée en évaluant le pourcentage de différences et les valeurs de K_s pour les gènes impliqués. Une méthode interspécifique exemplaire est celle de Fitch et al. Cette méthode nécessite beaucoup d'information préalable, la connaissance des orthologues et des phylogénies d'espèces. Elle permet dans certains cas d'identifier des conversions géniques entre gènes ancestraux (Fitch et al. 1990). Cette méthode est fondée sur le principe de parcimonie. Pour chaque site les arbres d'espèces sont reconstitués en postulant un nombre minimal de substitutions. En comparant les phylogénies ainsi obtenues on peut déceler les zones de conversion génique.

Il existe deux types d'homologies entre les gènes : les gènes orthologues sont

dérivés d'un ancêtre commun avant le processus de spéciation, les gènes paralogues sont le résultat de duplications qui se sont produites à l'intérieur d'une même espèce. Si pour un site donné, un nucléotide est commun à deux gènes paralogues on considère que ce site favorise l'hypothèse d'une conversion génique. Si d'autre part, pour un site donné, les gènes paralogues sont différents alors que les gènes orthologues sont semblables on considère que le site favorise l'hypothèse d'une absence de conversion génique. Ce type de situation suggère, en effet, une évolution indépendante des gènes provenant de la même espèce. Les zones de conversion seront celles qui coïncident avec une succession de sites favorisant l'hypothèse d'une conversion génique (voir aussi p. 51).

Cette méthode est en général assez fiable mais comme l'analyse doit être effectuée manuellement, site par site, elle peut parfois être extrêmement laborieuse. Dans le but de simplifier cette méthode Fitch et Goodman ont mis au point une méthode semi-automatique baptisée "scanning phylogénétique" (Fitch and Goodman 1991). Cette méthode exploite les informations fournies par les algorithmes de reconstitution phylogénétique. Même s'il est erroné un arbre phylogénétique obtenu de cette façon nous renseigne sur l'emplacement d'éventuelles conversions géniques.

Lorsque des gènes provenant d'une même espèce ne sont pas impliqués dans des conversions géniques et évoluent indépendamment les uns des autres les algorithmes phylogénétiques auront tendance à les regrouper avec leurs orthologues

plutôt qu'avec leurs paralogues. Dans la mesure où les conversions géniques se produisent après la spéciation, des gènes impliqués dans des conversions auront tendance à être regroupés avec leurs paralogues. Des conversions géniques peuvent se produire à tous les niveaux de l'arbre phylogénétique, aussi bien au niveau des ancêtres qu'au niveau des descendants. Toutes sortes de situations sont donc possibles et à chaque situation on peut associer un arbre phylogénétique reconstitué qui, même s'il ne correspond pas à la bonne phylogénie, décrit un résultat possible de l'application de l'algorithme phylogénétique choisi.

En tenant compte de l'information déjà disponible on peut proposer un certain nombre "d'arbres-hypothèses". Le nombre d'arbres possibles peut être assez élevé mais si on connaît déjà la bonne phylogénie la tâche peut être grandement simplifiée car un grand nombre d'arbres peuvent être éliminés d'office. Les arbres hypothétiques sont ensuite testés sur de petites régions des gènes. L'ensemble des données (les gènes alignés) est balayé par une fenêtre de longueur déterminée à l'avance (en l'occurrence 20 nuléotides). Afin d'éviter que les fenêtres soient totalement isolées de leur contexte le pas (jump) est choisi de façon qu'elles se superposent. Ainsi pour une fenêtre de 20 nucléotides le pas sera de 10 nucléotides seulement. Un algorithme phylogénétique teste les arbres hypothétiques et les classe en fonction de leur degré de concordance avec les données incluses dans la fenêtre. Lorsque le balayage est terminé on obtient pour chaque fenêtre un classement hiérarchisé des arbres hypothétiques. En examinant les arbres qui viennent en tête du classement on peut

déterminer à quels endroits des gènes et à quel niveau de la phylogénie les conversions géniques se sont produites.

Parmi les méthodes intraspécifiques, on peut citer la méthode de Drouin et Dover. Cette méthode fut appliquée aux exons des gènes d'actines de la pomme de terre. Elle suppose que l'on connaisse déjà la phylogénie exacte des gènes de la famille. Ce sera le cas si la famille n'a pas été le siège de conversions géniques importantes (macro conversions). Si tel est le cas, la phylogénie peut être obtenue par les méthodes habituelles de reconstruction. La méthode de Drouin et Dover consiste à repérer les sites incompatibles avec la phylogénie ainsi obtenue. Elle permet de détecter d'éventuelles micro conversions (Drouin and Dover 1990). On peut mentionner aussi la méthode heuristique de Jotun Hein. A partir du principe de parcimonie, cette méthode permet, dans certains cas, de reconstituer une phylogénie en tenant compte des conversions géniques (Hein 1993).

2) MÉTHODES DE SIMULATION

Une méthode statistique importante est celle de Stephens (1985). Dans le test de Stephens les séquences sont alignées et seuls les sites polymorphes (comportant au moins deux bases différentes) sont considérés. Les sites monomorphes sont exclus de l'analyse. Une partition est obtenue en regroupant les séquences ayant, à un site donné, des nucléotides identiques. À chaque site polymorphe correspond donc une

partition de l'ensemble des gènes de la famille. Les sites correspondant à la même partition sont appelés sites congruents. Pour simplifier on peut ne considérer que les sites polymorphes n'ayant que deux variantes possibles. Chacun de ces sites divise le groupe de séquences en deux sous-groupes (partition binaire). Pour m séquences il y aura $2^{m-1} - 1$ partitions binaires possibles. Une partition binaire est dite primaire lorsqu'elle a été obtenue par une substitution unique ayant affectée, à un moment donné, le site de l'ancêtre commun à toutes les séquences. Une partition binaire est dite secondaire lorsqu'elle est le résultat d'événements plus complexes tels que des substitutions multiples, des convergences, des reversions ou encore des conversions géniques. Dans le cas de conversions géniques on s'attend à ce que les sites congruents s'agrègent dans une région des séquences considérées. Il est possible de déterminer si l'agrégation de sites congruents est statistiquement significative.

Les autres méthodes sont celles qui ont été choisies pour réaliser cette étude. Il s'agit d'une méthode interspécifique, la méthode des codoubles (Balding et al. 1992) et de deux méthodes intraspécifiques, la méthode de Sawyer (Sawyer 1989) et la méthode des densités (présentée dans cette étude). Etant donné leur nature, ces méthodes ne peuvent être appliquées qu'aux exons des gènes étudiés. Elles seront présentées au chapitre II.

C - OBJECTIFS À ATTEINDRE

Le but de ce projet fut d'utiliser 5 méthodes différentes dont la méthode de Sawyer (1989) et à la méthode des codoubles (Balding et al. 1992), afin de détecter la conversion génique chez les trois familles multigéniques suivantes : la famille des gènes d'actines végétales, la famille des globines de primates et la famille des gènes mammaliens de type Zfx/Zfy. Comme les méthodes de Sawyer et de Balding et al. ne permettent pas de déterminer où se situent les zones de conversion génique, on a cherché à mettre au point une méthode permettant d'identifier les gènes impliqués et de localiser les zones de conversion (Méthode des densités).

Lorsque les données s'y prêtaient, lorsque l'information phylogénétique disponible était suffisante, on a aussi utilisé une méthode de compatibilité (analyse site par site) qui est une variante de la méthode de Fitch et al (1990).

Afin de tenir compte des caractéristiques particulières de certaines données (les gènes de type Zfx/Zfy), on a mis au point une variante de la méthode des codoubles (la méthode des taux variables).

On a donc utilisé, au total, quatre méthodes de simulation et une méthode de compatibilité. En plus de généraliser la méthode de Balding et al. au cas de l'isoleucine

il a fallut mettre au point deux méthodes statistiques (la méthode des taux variables et la méthode des densités) et une méthode de compatibilité.

Les données analysées sont constituées exclusivement de séquences d'exons. Un premier groupe de données est constitué de séquences provenant de la famille des gènes d'actines végétales. Ces gènes proviennent des cinq espèces d'angiosperme suivantes : Pomme de terre (Solanum tuberosum), Maïs (Zea maize), Tabac (Nicotiana tabacum), Soja (Glycine max) et Tomate (Lycopersicon esculemtum).

Le deuxième groupe de données comprend les séquences suivantes : les exons des globines δ et β de l'Humain (Homo sapiens) et du Tarsier (Tarsius Syrichta), les exons des globines γ1 et γ2 de l'Humain et de l'Orang-outang (Pongo pygmaeus).

Le troisième groupe de données est constitué d'exons provenant des gènes de doigts à zinc liés aux chromosomes sexuels des mammifères (gènes Zfx et Zfy). Les séquences utilisées sont les domaines codant pour les doigts à zinc. Ces séquences proviennent des sept espèces suivantes : Humain (Homo sapiens), Orang-outang (Pongo pygmaeus), Babouin (Papio cyanocephalus), Singe écureuil (Siamiri boliviensis), Renard (Dusicyon thous), Souris (Mus musculus) et Rat (Rattus norvegicus).

CHAPITRE II - MÉTHODES

A - ALIGNEMENT DES SÉQUENCES

La liste des gènes d'actine utilisés pour cette analyse est présentée dans les pages suivantes (tableau 1). Pour obtenir les différentes matrices de départ, les gènes ont été alignés manuellement. Comme les actines ont des gènes très conservés, cet alignement ne pose aucun problème.

Les gènes de globines ont déjà été étudiés du point de vue de la conversion géniques par des méthodes de compatibilité. On sait que les résultats des analyses phylogénétiques dépendent beaucoup de l'alignement utilisé. Afin de pouvoir comparer les résultats des méthodes de simulation avec ceux des méthodes de compatibilité on s'est servi des mêmes alignements pour les deux méthodes. Les alignements dont on s'est servi pour les globines et les gènes de doigts à zinc sont les mêmes que ceux qui ont été utilisés pour effectuer les analyses de compatibilité. (Fitch et al. 1990, Koop et al. 1989, Shimmin et al. 1994) La liste des gènes de doigts à zinc utilisés est présentée au Tableau 2. Les alignements utilisés pour les globines sont disponibles à la banque de données Genbank / Embl. Pour les globines γ_1 et γ_2, les numéros d'accès sont respectivement J04428 et J04429 (Koop et al. 1989). Pour l'alignement des globines δ et β, le numéro d'accès est DS5020. DAT (Fitch et al. 1990).

Tableau 1. Liste des séquences d'actines végétales utilisée pour cette étude

Espèce	Gène	N° d'accès	Référence
Pomme de terre (Solanum tuberosum)	PoAc 58	X55749	Drouin & Dover (1990)
	PoAc 71	X55750	"
	PoAc 75	X55753	"
	PoAc 97	X55751	"
	PoAc 101	X55752	"
	pot 46		Moniz de Sá and Drouin (1996)
	pot 66		"
	pot 79		"
	pot 82		"
	pot 42		"
	pot 65		"
Maïs (Zea maize)	MAc 1	J01238	Shah et al. (1982)
	maz 56		Moniz de Sá and Drouin (1996)
	maz 63		"
	maz 65		"
	maz 81		"
	maz 83		"
	maz 87		"
	maz 89		"
	maz 95		"

Tableau 1. (suite)

Espèce	Gène	N° d'accès	Référence
Tabac (Nicotiana tabacum)	TAc 25	X63603	Thangavelu et al. (1993)
	tob 54		Moniz de Sá and Drouin (1996)
	tob 71		"
	tob 93		"
	tob 103		"
	tob 104		"
Soja (Glycine max)	SAc 1	J01298	Shah et al. (1982)
	Sac 3	J12097	
	Soy 57		Moniz de Sá and Drouin (1996)
	Soy 58		"
	Soy 69		"
	Soy 70		"
	Soy 86		"
	Soy 109		"
	Soy 110		"
	Soy 115		"
	Soy 118		"
	Soy 119		"
Tomate (Lycopersicon esculemtum)	tom 32		Moniz de Sá and Drouin (1996)
	tom 41		"
	tom 51		"
	tom 52		"

Les numéros d'accès sont ceux des banques de données Genbank et Embl.

Tableau 2. Liste des séquences de type Zfx/Zfy

Espèce	Gène	N° d'accès	Référence
Orang-outang	Zfx	X75169	Shimmin et
(Pongo pygmaeus)	Zfy	X75176	al. (1994)
Babouin	Zfx	X75174	"
(Papio cyanocephalus)	Zfy	X75173	"
Singe écureuil	Zfx	X75175	"
(Siamiri boliviensis)	Zfy	X75170	"
Rat	Zfx	X75171	"
(Rattus norvegicus)	Zfy	X75172	"
Humain	ZFX	M26946	Schneider-Gädicke et al. (1989)
(Homo sapiens)	ZFY	J03134	Page et al. (1987)
Souris	Zfx	M32308	Mardon et al.(1990)
(Mus musculus)	Zfy	M24401	Mardon and Page (1989)
Renard	Zfx	M81106	Lanfear and Holland
(Dusicyon thous)	Zfy	"	(1991)

B - MÉTHODES STATISTIQUES

1) PRINCIPES GÉNÉRAUX

Les quatre méthodes statistiques utilisées pour cette étude ont toutes un certain nombre de caractéristiques communes. Elles reviennent toutes à simuler des données hypothétiques et à comparer les résultats de cette simulation avec les données observées. Dans les quatre cas, la simulation est effectuée en rearrangeant les sites silencieux. Pour illustrer le principe de base de ces méthodes, considérons l'exemple simple ci-dessous.

					Polypeptide
Ancêtre	AAT	TAT	AAA	GAA	NYKD
	↓	↓	↓	↓	↓
	AAC	TAT	AAT	GAA	NYND
	↓	↓	↓	↓	↓
Séquence observée	AAC	TAT	AAT	GAG	NYND

Cet exemple indique les substitutions qui se sont effectivement produites pour passer de l'ancêtre à la séquence connue.

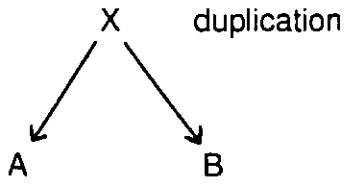
Dans cet exemple, on suppose :

- que la probabilité d'une mutation est la même tout le long du gène.
- que la sélection naturelle s'exerce au niveau du phénotype (de la chaîne polypeptidique).

Comme les substitutions synonymes n'ont pas d'effet sur le phénotype, on s'attend à ce qu'elles soient disposées de façon aléatoire. Cela signifie qu'on aurait pu avoir avec une égale probabilité la situation représentée ci-dessous.

					Polypeptide
Ancêtre	AAT	TAT	AAA	GAA	NYKD
	↓	↓	↓	↓	↓
	AAT	TAC	AAT	GAA	NYND
	↓	↓	↓	↓	↓
Séquence observée	AAT	TAC	AAT	GAG	NYND

La séquence du gène ancestral n'est généralement pas disponible mais si on dispose de la séquence d'un gène descendant du même ancêtre que le gène considéré, on a la situation représentée ci-dessous.



Si les trajets XA et XB sont aléatoires, alors le trajet AB sera aléatoire. Les différences observées chez la paire AB seront aussi disposées de façon aléatoire. En réarrangeant les différences et les similitudes observées sur les troisièmes sites des codons synonymes, on génère une population Ω de dispositions qui sont toutes équiprobables. Si une disposition est non aléatoire (par exemple si elle est due à une

conversion génique), elle n'aura pas la même probabilité que les dispositions de la population Ω. On peut remarquer que ce type de réarrangement n'affecte pas le rapport Ms/Ns.

2) MÉTHODE DE SAWYER

La méthode de Sawyer fut mise au point, à l'origine, pour remédier aux inconvénients d'une autre méthode statistique, la méthode de Stephens (Stephens 1985 et voir p.19). La méthode de Stephens fut appliquée avec succès à l'étude de la conversion génique chez des bactéries et chez des primates. Cette méthode a cependant ses limites car, comme le fait remarquer Sawyer, elle ne peut être appliquée efficacement qu'à un nombre restreint de séquences. Si les séquences comparées ont un degré de polymorphisme un tant soit peu élevé et/ou si leur nombre est élevé (plus que trois ou quatre) il devient très difficile de trouver un nombre suffisant de sites congruents. Dans le cas d'un nombre élevé de séquences le problème des comparaisons multiples peut se poser car, étant donné le nombre important de partitions binaires possibles, il peut arriver que certaines partitions aient une distribution non uniforme tout-à-fait par hasard.

Pour remédier à ces inconvénients Sawyer proposa une méthode pouvant s'appliquer même lorsque le nombre de séquences considérées est très élevé. Comme la méthode de Stephens la méthode de Sawyer consiste à tester deux

hypothèses mutuellement exclusive: l 'hypothèse H_0 est celle d'une disposition aléatoire des arrangements observés, l'hypothèse H_1 est celle d'une conversion génique ou, plus précisément, d'une disposition non aléatoire des arrangements observés. La distribution correspondant à H_1 en inconnue mais on peut échantillonner la distribution correspondant à H_0. Une variable aléatoire reflétant la présence de zones à faible densité est définie et une simulation permet d'approximer sa distribution.On peut alors estimer la valeur P.

Comme le test de Stephens, le test de Sawyer s'applique aux sites polymorphes,. comme il ne concerne que les exons il ne considère qu'un sous-ensemble de ces sites: les sites polymorphes silencieux. Pour une séquence donnée un sous-ensemble des sites polymorphes silencieux délimite deux types de fragments: des fragments internes et des fragments externes. Dans le cas des fragments internes une séquence donnée peut être découpée de plusieurs façons, dans le cas des fragments externes il n'y a qu'un seul découpage possible. Les fragments externes sont délimités par les sites polymorphes silencieux correspondant à une partition dont un des sous-groupes contient à la fois la séquence fragmentée et au moins une des autres séquences des gènes de la famille inclus dans l'analyse.

En définissant les fragments externes de cette façon, on cherche à repérer des séquences de nucléotides n'apparaissant que chez un seul des gènes du groupe considéré. L'examen des fragments externes permet de déterminer si des conversions

se sont produites entre les gènes provenant du groupe considéré (les gènes connus et séquencés) et des gènes extérieurs au groupe et qui pourraient ne pas être disponibles. Les fragments internes et externes peuvent être condensés ou non condensés suivant que l'on considère seulement les sites polymorphes silencieux ou tous les sites de la séquence. Dans cette méthode deux types de variables aléatoires sont définies. Ces variables, appelées SSCF et SSUF, s'appliquent respectivement aux fragments condensés et aux fragments non condensés. Comme les fragments peuvent être internes ou externes il y a donc quatre tests possibles. Dans chaque cas on pourra estimer la valeur P. Un résultat sera considéré comme significatif s'il est ≤ à 0.05. La probabilité d'une erreur de type I sera donc 5%.

La méthode de Sawyer fut surtout conçue pour être appliquée aux bactéries et aux virus. Le groupe de données étudié comprenait 7 souches d'E. Coli (locus gnd), 8 souches d'E. Coli pour le locus phoA et 13 souches du virus de l'influenza A de l'humain (locus NS). Même s'il n'y a pas de reproduction sexuée ni de méiose chez les bactéries et même si les comparaisons se font plutôt entre souches bactériennes, la situation observée chez les bactéries peut être mise en parallèle avec ce qui se passe chez les eucaryotes. En effet, la condition à remplir pour que le test de Sawyer puisse être appliqué est que les gènes des différentes souches bactériennes soient organisés en phylogénie. On s'attend à ce que ces gènes participent à des recombinaisons de type conversion génique. Chez les eucaryotes, les conditions d'application du test de Sawyer seront remplies si l'on considère un groupe de gènes organisés en phylogénie

(en famille multigénique) et appartenant au même organisme. La méthode de Sawyer peut donc être appliquée aux eucaryotes en tant que méthode intraspécifique. Elle ne peut fonctionner que s'il existe un certain contraste entre les zones converties et les zones épargnées par la conversion. Si toutes les séquences d'exons ont été impliquées dans une conversion génique, les deux gènes comparés seront pratiquement identiques. Si on n'examine que les exons, une conversion de cette ampleur deviendra indétectable. La méthode de Sawyer ne peut donc pas faire de distinction entre le résultat d'une duplication récente et les effets d'une conversion génique de grande ampleur recouvrant la totalité des exons.

Puissance du test

On a vu que dans la méthode de Sawyer la probabilité de rejeter H_0 alors que H_0 est vraie (erreur de type I) est 0.05. L'évaluation de la probabilité d'une erreur de type II présente certaines difficultés et, d'une façon générale, le problème de la puissance des tests se pose pour chacune des méthodes statistiques utilisées pour cette étude.

L'identification de conversions géniques va souvent de paire avec la reconstruction d'une phylogénie et si le but recherché est l'obtention de la bonne phylogénie il devient important de pouvoir apprécier la probabilité d'une erreur de type II (ne pas rejeter H_0 alors que H_0 est fausse). L'évaluation de la puissance des tests et de la probabilité d'une erreur de type II est rendue difficile par le fait que la distribution correspondant à H_1 est inconnue. On peut être tenté de déterminer la probabilité d'une erreur de type II de façon analytique en simplifiant le modèle de référence. Cette

approche est peu satisfaisante car le modèle de référence est déjà, lui même, une simplification (voir p. 115). Une approche plus prometteuse est celle de la simulation déterministe ou modélisation.

À cause des difficultés que présente l'évaluation de la puissance des tests et comme les méthodes utilisées dans le cadre de cette étude se placent plutôt du point de vue de la conversion génique que de celui de la phylogénie, le problème de la détermination d'une erreur de type II est rarement envisagé. Il n'y a guère que dans la méthode de Sawyer que le problème de la puissance des tests est abordé.

La méthode proposée par Sawyer pour évaluer la puissance de son test est, en fait, une modélisation. Pour que la modélisation soit suffisamment véridique elle fut effectuée en s'inspirant d'une situation réelle : l'évolution du locus gnd d'E. Coli. Afin de recréer une évolution possible de ce type de gène c'est la séquence consensus des 7 gènes connus qui fut utilisée comme séquence ancestrale.

À partir d'un arbre phylogénétique plausible il fut possible de reconstituer un processus évolutif produisant sept descendants différents. Pour reconstituer ce processus un certain nombre de règles furent proposées. Les mutations non synonymes furent rejetées de façon aléatoire avec une probabilité de 50%. Les mutations synonymes mettant en jeu des transitions furent toutes sélectionnées. Pour tenir compte du fait que les transversions sont généralement moins probables que les

transitions elles ne furent retenues, elles aussi de façon aléatoire, que dans 50% des cas.

Ces règles furent choisies de façon à obtenir une proportion de sites deux fois et quatre fois dégénérés comparable à ce qui fut observé dans les données. On calcula que pour obtenir 81 sites silencieux il était nécessaire d'introduire 216 mutations potentielles. Une première simulation effectuée sans conversion génique en introduisant 216 mutations potentielles donna, pour tous les tests, des résultats non significatifs.

Pour modéliser la situation correspondant à l'hypothèse H_1 on postula 80 conversions géniques de longueurs variant entre 50 et 200 bases (la séquence du locus gnd comporte 768 bases). Les segments correspondant à ces conversions, localisés de façon aléatoire, furent transférés à des gènes choisis au hasard. Cette deuxième simulation montra que des conversions de peu d'ampleur mais sufisamment nombreuses peuvent homogénéiser rapidement l'ensemble des données et réduire le nombre de sites polymorphes silencieux de façon considérable.

Par expérimentation il fut possible de montrer que, dans de telles conditions, pour obtenir un nombre suffisant de sites polymorphes silencieux, entre 60 et 99, il fallait environ 414 mutations potentielles. Une simulation incluant 414 mutations potentielles et 80 conversions géniques fut effectuée dans les conditions décrites précédemment. Les données obtenues de cette façon furent analysées avec le test de Sawyer. Les valeurs P mesurées pour les variables SSCF et SSUF furent égales à 0.

Pour les deux simulations déterministes le test de Sawyer fournit les résultats attendus. Même si le test semble fonctionner correctement l'évaluation d'une erreur de type II nécessiterait une approche plus systématique, plusieurs milliers de simulations seraient nécessaires. Ce type de modélisation nous renseigne sur certaines limites du test, il suggère que des conversions géniques de peu d'ampleur mais suffisamment nombreuses peuvent rapidement homogénéiser les données et rendre le test de Sawyer inopérant. Même si la simulation de processus évolutif chez les eucaryotes nécessiterait probablement un système de règles plus complexes, la modélisation proposée par Sawyer suggère une façon de résoudre le problème de la puissance des tests.

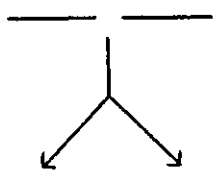
Pour évaluer la puissance des tests de façon adéquate il faudrait mettre au point des modèles déterministes ayant un niveau de complexité au moins égale à celui des modèles de référence (voir p. 115). En répétant ce type de simulation un grand nombre de fois en faisant varier le nombre de séquences, leurs tailles, la fréquence des conversions, la phylogénie, etc... il serait possible d'évaluer l'erreur de type II.

On ne peut pas sous-estimer le problème de l'évaluation de la puissance des tests mais étant donné la complexité des modèles mis en jeu une simulation déterministe, effectuée de façon systématique, dépasserait largement le cadre limité de cette étude et nécessiterait, en fait, la réalisation d'un projet distinct.

3) MÉTHODE DES CODOUBLES

a) Méthode de Balding et al.

Cette méthode peut être utilisée lorsque les gènes orthologues de deux espèces voisines ont été identifiés. Comme pour la méthode de Sawyer, on veut tester l'hypothèse d'une disposition non aléatoire des différences observées parmi les sites silencieux. Contrairement au test de Sawyer, ie test des codoubles n'est pas fondé sur un contraste entre zones de densités différentes mais sur la détection de phénomènes de convergence entre des gènes censés évoluer séparément. Pour cette raison et comme une comparaison est faite entre les familles multigéniques de deux espèces différentes, ce test peut détecter des zones de conversion génique de grande ampleur même si elles recouvrent la totalité des exons. Considérons le cas où l'on compare deux paires de gènes orthologues provenant de deux espèces, A et B. Dans ce cas, on a deux classes d'orthologues, A1B1 et A2B2. L'information connue peut être résumée par le diagramme ci-dessous.

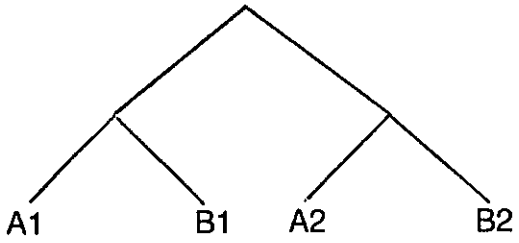


duplication

spéciation

A1_____A2 B1_____B2 gènes orthologues

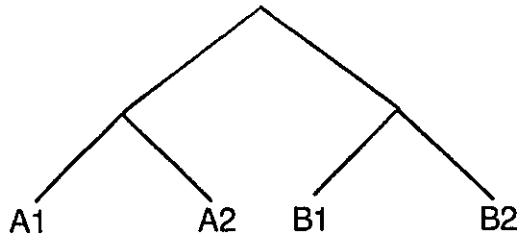
Comme la duplication s'est produite avant la spéciation, cette situation correspond à la phylogénie représentée ci-dessous (fig.a).



duplication

spéciation

Si la bonne phylogénie est celle de la Figure a et s'il y a eu conversion génique, les méthodes de reconstruction phylogénétique risquent de donner une phylogénie éronée, celle de la figure ci-dessous (fig.b).



Dans la mesure où les familles appartenant à A et à B sont isolées par la spéciation, on s'attend à ce que leurs différences synonymes soient réparties de facon aléatoire.

DÉTERMINATION DES SITES INFORMATIFS

Les gènes orthologues des deux espèces sont alignés dans l'ordre suivant : [A1 B1 A2 B2]. On ne considère que les sites correspondant à des codons synonymes pour les quatre séquences. Seules les classes de synonymie 2, 3 et 4 sont envisagées. Pour ces catégories, les seuls sites variables sont les troisièmes sites des codons. Seules les colonnes correspondant à ces sites seront considérées comme informatives. Soit n, le nombre de ces colonnes pour une classe de synonymie donnée. Si d est le nombre de sites correspondant à une différence et si I est le nombre de sites sélectionnés correspondant à deux nucléotides identiques, il y aura d+I pris I á la fois permutations possibles. Étant donné la façon dont les sites ont été sélectionnés d + I sera forcément un multiple de 2 et on aura d + I = 2n. Certaines des n colonnes informatives seront monomorphes et parmi les colonnes polymorphes, on aura quatre arrangements possibles :

Simple (type a):[N1N1, N1N2], [N1N1, N2N1], [N1N1, N2N3]

Simple (type b): [N1N2, N1N1], [N1N2, N2N2], [N1N2, N3N3]

Double: [N1N2, N1N2], [N1N2, N2N1], [N1N2, N3N1], [N1N2, N1N3],

[N1N2, N2N3], [N1N2, N3N2], [N1N2, N3N4]

Où N1, N2, N3 ... sont des nucléotides différents. On définit un codouble comme l'arrangement particulier suivant : [N1N2, N1N2]. On redistribue les différences observées entre les deux familles et l'on veut déterminer quelle proportion d'arrangements obtenus de cette facon comporte au moins x codoubles (x = nombre de codoubles observés dans les données). Comme dans le test de Sawyer les réarrangements sont effectuées à l'intérieur de chaque classe de redondance. Le choix de la variable aléatoire peut s'expliquer soit en relation avec le phénomène de convergence, soit dans le cadre de la méthode de parcimonie.

Considérons deux gènes provenant de la même famille, considérons un site particulier s de ces deux gènes. Avant la spéciation, il n'y a que deux situations possibles :

	gène 1	gène 2
Situation 1	---N1---	---N1---
Situation 2	---N1---	---N2---

Les nucléotides correspondant au site s sont soit identiques, soit différents. Après la spéciation, pour obtenir un codouble à partir de la situation 1 il faudra deux substitutions aux sites homologues, chez la même espèce, dont une convergence.

Exemple :

	gène 1	gène 2
Espèce A	---N1---	---N1---
Espèce B	---N1---	---N1---

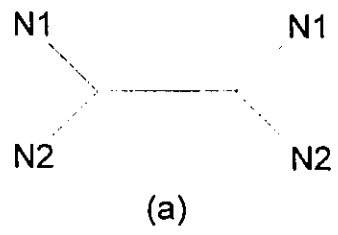
Espèce A	---N1---	---N1---
Espèce B	---N2---	---N2---

Pour obtenir un codouble à partir de la situation 2, il faudra deux convergences (N2-N1 et N1-N2).

	gène 1	gène 2
Espèce A	---N1---	---N2---
Espèce B	---N1---	---N2---
Espèce A	---N1---	---<u>N1</u>---
Espèce B	---<u>N2</u>---	---N2---

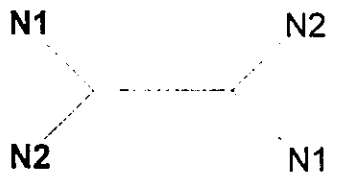
Comme les phénomènes de convergence sont peu fréquents, surtout dans le cas des sites dégénérés quatre fois, s'il n'y a pas eu de conversion génique, on s'attend à ce que les codoubles soient des évènements plutôt rares.

Lorsqu'on effectue une reconstitution phylogénétique avec la méthode de parcimonie, on suppose que la phylogénie la plus probable est celle qui nécessite le plus petit nombre de substitutions. Cette méthode revient, en fait, à dénombrer des colonnes d'un type particulier, appelées colonnes de parcimonie. Pour quatre séquences, il y a trois phylogénies possibles et trois colonnes de parcimonie. La figure ci-dessous représente les arbres correspondant à ces trois colonnes de parcimonie.



(a)

(b)



(c)

Chaque colonne de parcimonie favorise une phylogénie partículière. La phylogénie considérée comme la plus probable (la plus parcimonieuse) sera celle qui correspond aux colonnes de parcimonie les plus nombreuses.

On remarque que, pour obtenir les arbres (a) et (c), il faut au minimum deux substitutions, alors que l'arbre (b) ne nécessite qu'une substitution. Une façon équivalente de procéder consiste donc à chercher laquelle des trois phylogénies possibles va maximiser le nombre de colonnes de parcimonie de type (b).

Un codouble est en fait une colonne de parcimonie. Si les hypothèses de la méthode de Balding and al. sont respectées, le codouble correspondra à la figure (a). S'il n'y a pas eu de conversion génique, on s'attend à ce que les colonnes de type (b) soient plus nombreuses que les colonnes de type (a) ou (c). Une conversion

génique aura pour effet de faire disparaître les colonnes de type (b) ou (c) qui se situent dans la région de conversion. Une conversion génique n'aura, en revanche, aucun effet sur les codoubles. Pour quatre séquences, il y a 14 types de colonnes possibles et donc 11 colonnes qui ne sont pas des colonnes de parcimonie. Certaines de ces colonnes deviendront des codoubles (suivant la polarité de la conversion). Par exemple, si les colonnes sont dans l'ordre : [A1B1 A2B2] et s'il y a conversion génique de A2 vers A1, les colonnes de type [N1N1N2N1] ou [N1N2N3N2] deviendront des codoubles. On voit comment une conversion génique peut perturber le dosage des colonnes de parcimonie et être à l'origine d'une proportion plus grande de codoubles.

Cas de l'isoleucine

La méthode de Balding et al. fut mise au point dans le but d'étudier les gènes codant pour les pigments visuels chez l'Homme et chez le Singe Diana. L'étude portait sur les exons codant pour les pigments vert et rouge. Ces derniers étant assez restreints (au total 264 nucléotides), aucune différence ne fut observée au niveau des codons correspondant à l'isoleucine. Pour cette raison, la méthode de Balding et al. n'envisage pas le cas de l'isoleucine.

Soit n, le nombre de sites sélectionnés, t + t' le nombre de transitions observées entre les séquences des espèces A et B et r + r' le nombre de transversions observées entre les sequences des espèces A et B. Pour l'isoleucine, la probabilité p pour qu'un double formé de deux transitions soit aussi un codouble est égale à 1. La probabilité q pour qu'un double formé de deux transversions soit aussi un codouble est égale à 1/2. Si Pr (i ,j) = probabilité d'avoir i doubles transitions et j doubles transversions, la probabilité d'avoir x codoubles formes de deux transitions et y codoubles formés de deux transversions sera donnée par la formule suivante :

$$P_r\left[\ (X,Y) = (x,y)\ \right] = \sum_{i\ x}^{\frac{d}{2}} \sum_{j\ y}^{\frac{d}{2}\ i} \binom{i}{x} p^{i} (1-p)^{x-i} \binom{j}{y} q^{j} (1-q)^{y-j} P(i,j) \qquad (1)$$

avec

$$P_r\ (i,j) = \sum_{k\ 0}^{\frac{d}{2}} 2^{d-2i-2j} \binom{n}{i,\ j,\ k,\ (t+t'-2i)\ ,\ (r+r'-2j)} \binom{2n}{t+t',r+r'}^{-1}$$

b) Méthode des taux variables

La méthode de Balding et al. peut être appliquée lorsque les taux de mutation des différents gènes de la famille sont comparables. Il peut arriver que ce ne soit pas toujours le cas. Dans ces conditions, on peut utiliser une variante de la méthode des codoubles. En utilisant la même variable aléatoire X que celle qui fut définie pour le test de Balding et al., nous avons mis au point un test qui ne présume rien des taux de mutation (méthode des taux variables). Au lieu de redisposer les sites silencieux entre les séquences A1A2 et les séquences B1B2, on réarrange séparément les différences observées entre A1 et B1 et les différences observées entre A2 et B2. En procédant de cette façon, on génère une population d'arrangements qui est un sous-ensemble des arrangements dénombrés par les équations de Balding et al.

Cas des sites dégénérés deux fois

Pour les sites dégénérés deux fois, la probabilité d'obtenir exactement i doubles différences sera donnée par l'équation suivante :

$$P_r\ (i) \quad = \quad \frac{\begin{pmatrix} n \\ i, t-i, t'-i \end{pmatrix}}{\begin{pmatrix} n \\ t \end{pmatrix}\begin{pmatrix} n \\ t' \end{pmatrix}} \tag{2}$$

La probabilité d'obtenir un codouble, étant donné que l'on a un double, est égale à 1/2 (Balding et al. 1992). La probabilité d'obtenir exactement x codoubles sera donc

$$P_r\ (X = x) \quad \sum_{i=x}^{Min(t,t')} \begin{pmatrix} i \\ x \end{pmatrix} \left(\frac{1}{2}\right)^i \quad P_r(i) \tag{3}$$

La valeur P sera alors calculée de la façon suivante :

$$P_r\ (X \geq x) \quad = \quad \sum_{X=x}^{Min(t,t')} \sum_{i=X}^{Min(t,t')} \begin{pmatrix} i \\ X \end{pmatrix} \left(\frac{1}{2}\right)^i P_r(i) \tag{4}$$

Cas des sites dégénérés quatre fois

Pour les sites dégénérés quatre fois, la probabilité d'obtenir exactement i doubles transitions et j doubles transversions sera donnée par l'équation suivante :

$$P_r(i,j) = \sum_{k=0}^{d-2i-2j} \sum_{l=0}^{k} F(k,l) \binom{n}{t,r}^{-1} \binom{n}{t',r'}^{-1} \quad (5)$$

avec

$$F(k,l) = \binom{n}{i,j,l,(k-l),(t-i-l),[r-j-(k-l)],[t'-i-(k-l)],(r'-j-l)}$$

r = nombre de transversions observées chez la paire A1B1

r' = nombre de transversions observées chez la paire A2B2

d = t + t' + r + r' (nombre total de différences)

La probabilité d'obtenir exactement x codoubles formés de deux transitions et y codoubles formés de deux transversions sera donnée par la loi binômiale :

$$P_r\,(X=x, Y=y) = \sum_{i=x}^{Min(t,t')} \sum_{j=y}^{Min(r,r')} \binom{i}{x} \left(\frac{1}{2}\right)^{i} \binom{j}{y} \left(\frac{1}{4}\right)^{y} \left(\frac{3}{4}\right)^{j-y} P_r(i,j) \quad (6)$$

La valeur P sera considérée comme significative si elle est ≤ 0.05.

4) MÉTHODE DES DENSITÉS

Les méthodes décrites précédemment ne sont pas spécifiques. Elles ne permettent pas de déterminer l'emplacement des zones de conversion génique. Considérons deux séquences préalablement alignées. On a vu que la conversion génique avait pour effet l'apparition de zones à faible densité. On a vu aussi que les zones ayant les densités les plus faibles doivent être recherchées parmi les zones bornées par deux différences (chaque zone appartenant à une catégorie définie par un paramètre k).

Considérons seulement les sites silencieux pouvant être sujets à des variations (comme dans la méthode des codoubles). Pour chaque valeur de k, on peut trouver un intervalle i ayant une densité minimale. Dans ces conditions, on peut considérer que la localisation d'une zone de conversion génique revient à la détermination d'une valeur du paramètre k (voir p.11). Pour chaque valeur de k, on peut définir la variable aléatoire X qui représente la longueur de l'intervalle i.

Appelons d, le nombre total de différences observées sur la paire de séquences considérées. Soit r le nombre total de sites identiques.

Si on tient compte du bord de la paire de séquences, les r sites peuvent être répartis entre d + 1 intervalles. Si k = 0, on veut déterminer la probabilité pour qu'un de ces d + 1 intervalles, pris au hasard, ait une longueur égale à x (mesurée en nucléotides). Ce problème revient à déterminer le nombre de façons de disposer r boules blanches parmi d + 1 cellules, sachant qu'une cellule contient déjà x boules blanches. La probabilité d'obtenir un intervalle de longueur x sera donnée par la formule suivante :

$$P_r(X=x) = \frac{\binom{d+r-x-1}{d-1}}{\binom{d+r}{d}} \qquad (7)$$

La probabilité pour qu'un intervalle pris au hasard ait une longueur ≥ à x sera donnée par l'équation :

$$P_r(X \geq x) = \sum_{i=x}^{r} \left[\frac{\binom{d+r-i-1}{d-1}}{\binom{d+r}{d}} \right] \qquad (8)$$

On a vu que les sites ne sont pas tous équivalents. Ils peuvent appartenir à des classes de synonymie différentes. Si on veut faire une distinction entre les différents types de sites, le problème est un peu plus compliqué. Pour le résoudre, on peut remarquer que l'équation (7) est égale à la suite de termes ci-dessous :

$$\left[\binom{d+r-x-2}{d-2} + \binom{d+r-x-3}{d-2} + \binom{d+r-x-4}{d-2} + \cdots + \binom{d+r-x-2-[d+r-x-2-(d-2)]}{d-2}\right] \binom{d+r}{d}^{-1}$$

On voit que l'expression (7) peut être évaluée de deux façons différentes. Si on fait une distinction entre les différents types de sites, la valeur de P(X = x) ne peut pas être évaluée directement mais la situation est analogue au cas précédent. Il existe une suite de termes qui peuvent être évalués séparément et dont la somme est égale à P(X = x). Une autre façon de procéder, et c'est la méthode adoptée pour cette étude, consiste à obtenir une approximation de la valeur P, en générant un échantillon de données hypothétiques (voir méthode de Sawyer). Comme dans les méthodes décrites précédemment, on considère qu'un résultat est significatif, lorsque la valeur P est ≤ 0.05.

C - MÉTHODE DE COMPATIBILITÉ

1) ANALYSE SITE PAR SITE

Cette méthode peut être employée lorsqu'on dispose de séquences provenant de plusieurs espèces et lorsqu'on connait la phylogénie de ces espèces. On utilise cette information supplémentaire pour détecter d'éventuelles conversions géniques.

L'arbre d'espèces qui sert de point de départ à l'analyse a été obtenu, au préalable, à partir d'autres sources d'information, à partir, par exemple, de données paléontologiques ou morphologiques.

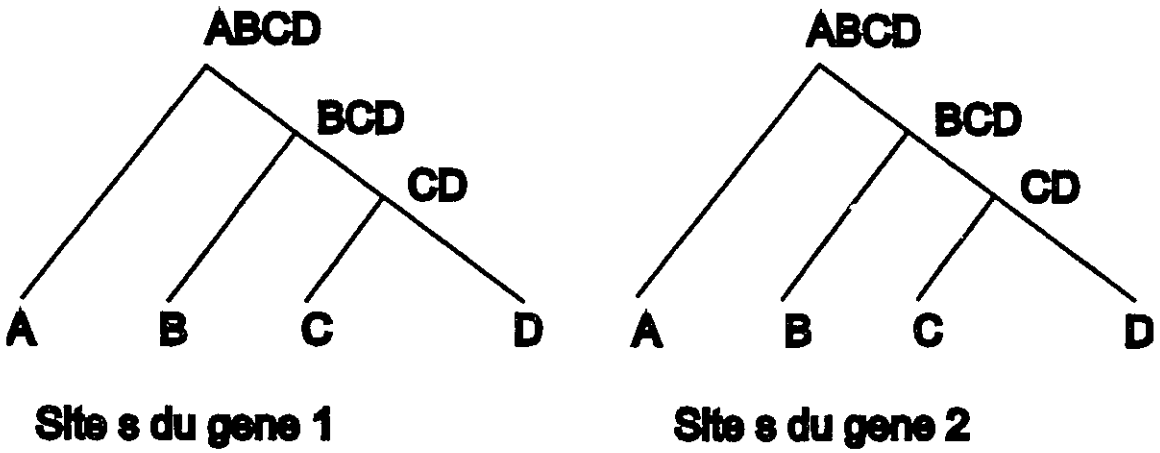
Il s'agit essentiellement :

- de postuler des nucléotides ancestraux en supposant un nombre minimum de substitutions (hypothèse parcimonieuse)
- de coder et de disposer l'information recueillie de façon à faire apparaître d'éventuelles zones de conversion génique.

Si le codage des données fait apparaître une succession quasi ininterrompue de sites compatibles avec l'hypothèse d'une conversion génique, cette disposition ayant peu de chances de se produire par hasard, on supposera qu'elle coïncide avec une zone de conversion génique.

C'est ce type d'approche qui permit de détecter de nombreuses zones de conversion chez la famille des globines (Slightom et al. 1987, Fitch et al.1990). La méthode décrite dans les pages qui suivent est une variante de la méthode de Fitch. Les principes de base sont les mêmes, mais le codage de l'information prend une tournure différente.

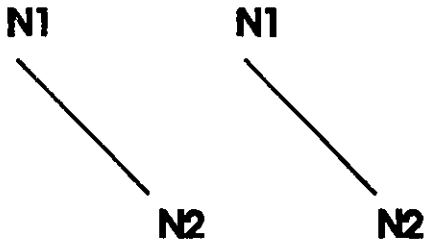
Supposons que l'on cherche à déterminer si des conversions géniques se sont produites entre deux gènes différents. Ces deux gènes étant alignés, ils comportent tous les deux le même nombre de sites. A chaque site, on peut associer deux arbres phylogénétiques correspondant chacun à un des deux gènes étudiés. Les nucléotides correspondant aux noeuds de ces arbres phylogénétiques sont obtenus par la méthode de parcimonie. L'arbre d'espèces peut être considéré comme formé de différents segments. Chaque segment relie soit deux noeuds, soit un noeud et une feuille de l'arbre (figure ci-dessous).



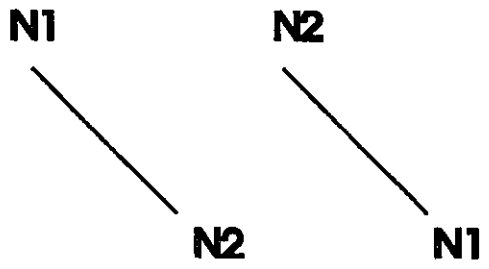
Comme les phylogénies correspondant au site étudié ont la même topologie, on peut les comparer segment par segment. Chaque segment, ou entre-noeud, comporte deux sites. On considère le site le plus récent (par exemple le point D dans le segment CD-D). Si les deux nucléotides correspondant à ce site sont identiques à l'un des deux nucléotides ancestraux, on considère que ce site est compatible avec l'hypothèse d'une conversion génique. Le site correspondant sera identifié par un O.

Comme ces conditions peuvent être remplies de deux façons différentes, on peut associer une polarité au site examiné. La polarité de la conversion sera indiquée par un + ou un -, en précisant la convention de signes.

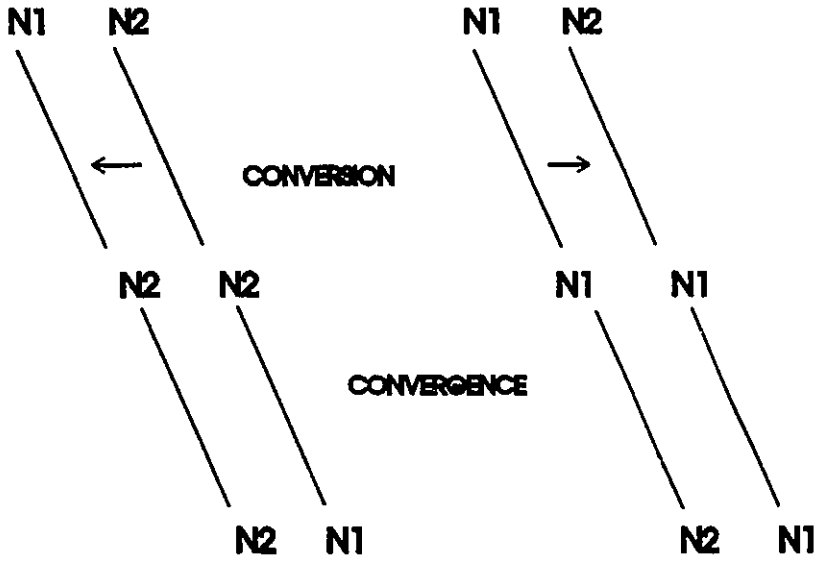
Le symbole "O" est employé isolément lorsque dans l'hypothèse d'une conversion génique, il est impossible de déterminer quel gène aurait pu convertir l'autre, la polarité est indéterminée (chaque ligne represente un segment de chacun des deux arbres considérés):



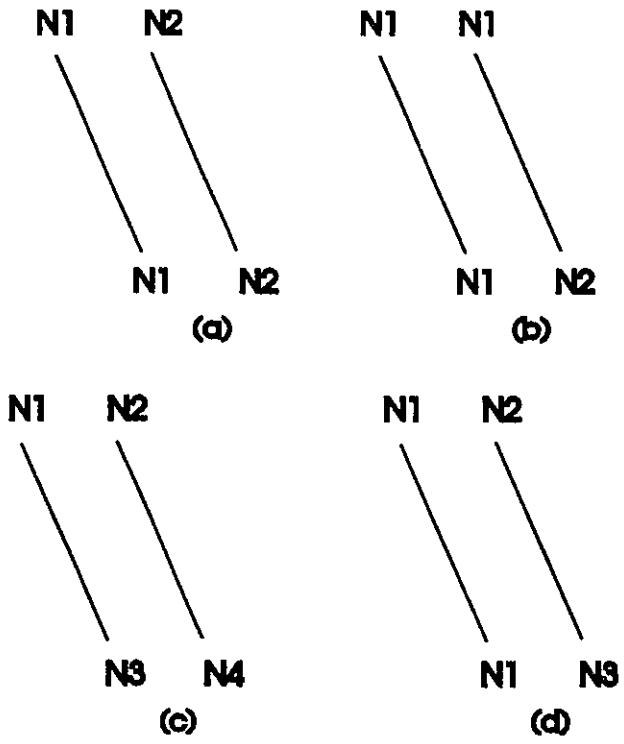
Il existe une troisième possibilité théorique compatible avec l'hypothèse d'une conversion génique à polarité indéterminée mais elle n'a été observée qu'une seule fois dans les données.



Cette possibilité correspondrait à deux convergences ou à une conversion génique (à polarité indéterminée) suivie d'une convergence:

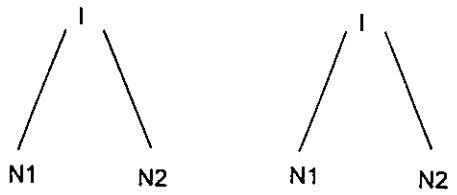


Les autres situations (cas où les deux nucléotides récents sont différents chez les deux séquences) seront considérées comme ne favorisant pas l'hypothèse d'une conversion génique (figure ci-dessous). Les sites correspondant seront indiqués par un X. La situation (a) étant particulièrement improbable dans l'hypothèse d'une conversion génique, elle sera indiquée par un X souligné.

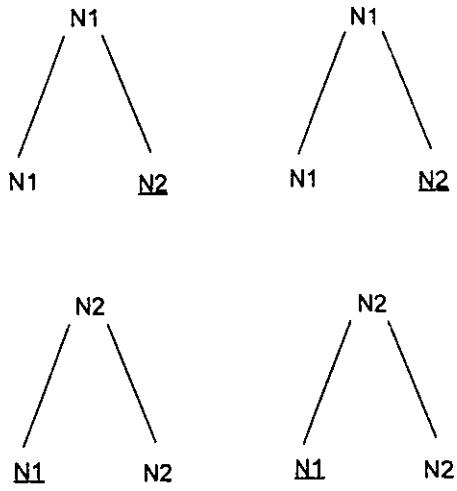


CAS INDÉTERMINÉS

Il peut arriver que certains nucléotides soient indéterminés. Parmi les cas indéterminés compatibles avec l'hypothèse d'une conversion génique, on a la situation suivante (il s'agit en fait d'un codouble) :

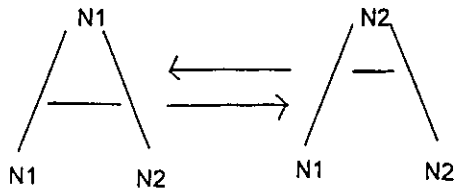


Ce type d'indétermination correspond aux possibilités ci-dessous :



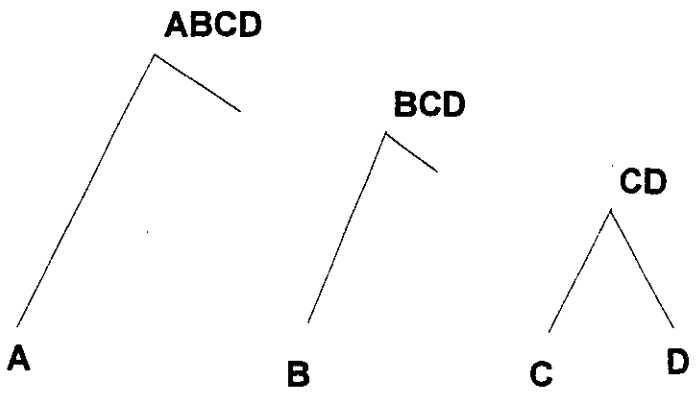
Soit une conversion de N1 vers N2, soit une conversion de N2 vers N1, le site affecté sera donc soit celui qui correspond aux nucléotides N1, soit celui qui correspond aux nucléotides N2 et la polarité est indéterminée.

Une autre possibilité théorique (très improbable) est celle d'une double conversion :



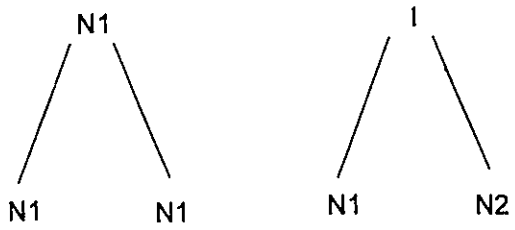
Ce premier type de cas indéterminé ne permet pas de décider si la conversion génique s'est produite entre les noeuds iN1 ou entre les noeuds iN2.

Comme il y a une certaine ambiguïté, on ne sait pas lequel des deux sites a pu être affecté par une conversion génique. Il vaut mieux utiliser un symbole approprié différent du symbole habituel. Par convention, les deux sites seront indiqués par le chiffre 3. Pour interpréter cette notation, on utilise une propriété de la topologie de l'arbre d'espèces. L'arbre peut être considéré comme formé de sous-arbres binaires.



Il est commode de représenter côte à côte les paires de gènes correspondant au même sous-arbre binaire.. La présence d'un 3 sur les deux sites provenant du même arbre binaire signifiera que l'un des deux sites correspondra à l'hypothèse d'une conversion génique.

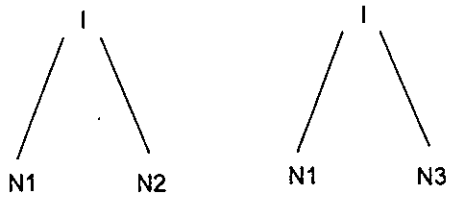
Il existe une autre catégorie d'indétermination (ci-dessous). Elle sera dénotée par le chiffre 2.



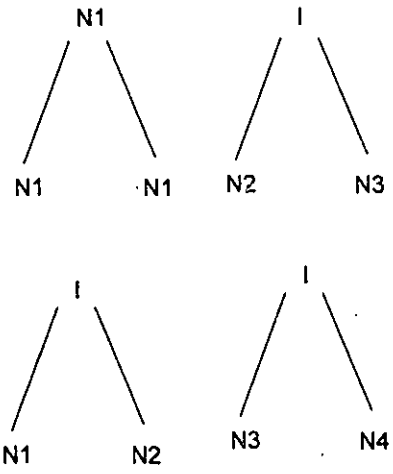
Elle correspond a deux possibilités, soit une conversion génique (cas a), soit une mutation de N1 en N2 (cas b) :

(a) (b)

Il existe une troisième possibilité théorique, compatible avec l'hypothèse d'une conversion génique (représentée ci-dessous) mais rarement observée dans les données. Cette situation sera indiquée par le chiffre 1.



Les indéterminations favorisant l'hypothèse d'une absence de conversion génique sont représentées ci-dessous.



Dans ce cas, les deux sites impliqués seront marqués par un X.

CHAPITRE III
ÉTUDE DES DIFFÉRENTES FAMILLES MULTIGÉNIQUES

A - GROUPE DES GÈNES D'ACTINES VÉGÉTALES

On trouve des actines dans le cytoplasme de tous les eucaryotes. La présence de ces molécules n'a jamais été décelée chez les procaryotes. Même si l'on soupçonne leur présence dans le noyau cellulaire, les actines se cantonnent principalement au cytoplasme.

Les monomères d'actine (actine G ou globulaire) s'assemblent pour former les micro filaments (actine F). Ces micro filaments s'attachent à des protéines liées à l'actine pour former des structures tri-dimensionnelles complexes (cytosquelette d'actine). Il s'agit de structures dynamiques dont les éléments se renouvellent sans cesse. Dans un cycle d'assemblage et de désassemblage, chaque monomère d'actine hydrolyse une molécule d'ATP. Le cytosquelette d'actine est impliqué dans toutes sortes de phénomènes liés à la motilité cellulaire. L'exemple le plus classique du rôle des actines est la contraction des muscles striés. Le rétrécissement des sarcomères est rendu possible par une interaction de l'actine avec la myosine.

Les actines ont surtout été étudiées chez les animaux, les actines végétales demeurent assez mal connues. Comme les filaments d'actine animale, les filaments observés chez les plantes se réorganisent considérablement pendant le cycle

cellulaire. En règle générale, les actines végétales semblent avoir des propriétés assez semblables à celles des actines animales. Chez les plantes supérieures, on a montré que l'actine était impliquée dans divers processus comme la mitose, la division cellulaire, la croissance des tubes polliniques, des radicelles. En utilisant des inhibiteurs comme les cytochalassines et les phalloïdines, on a montré que l'actine jouait un rôle dans la cyclose et les mouvements des chloroplastes. Une étude effectuée chez les cellules de l'endosperme a montré que pendant la cytocinèse, les actines jouent un rôle dans la formation du phragmoplaste (Schmitt and Lambert 1988 cité par Reece et al. 1992).

Sauf chez les champignons et chez quelques protistes, les actines sont généralement organisées en familles multigéniques. Chez les vertébrés, les actines sont classées en deux groupes principaux : actines cytoplasmiques (β et γ) et actines musculaires, ces dernières comprenant les actines des muscles lisses (entériques et vasculaires), cardiaques et striés. Chez les plantes supérieures, les familles d'actines comportent de nombreux gènes. Le nombre exact de ces gènes n'est pas encore connu. Chez le pétunia, on estime que la famille des gènes d'actine comprend entre 100 et 200 membres répartis en au moins six sous-familles (McLean et al. 1990). Chez les familles choisies pour cette étude (tabac, maïs, pomme de terre, tomate et soja), on estime que la famille des actines doit comporter entre 20 et 40 gènes (Moniz de Sá and Drouin 1996).

Alors que chez les animaux, la position et le nombre d'introns varient considérablement, les gènes fonctionnels d'actines d'angiospermes comportent généralement trois introns, toujours situés au même endroit. Certains pseudo gènes d'actines végétales font exception à cette règle. Chez le gène tom 32 de la tomate, le deuxième intron est absent. Chez deux gènes de la pomme de terre, les deux derniers introns sont absents, seul le premier intron subsiste (Moniz de Sá and Drouin 1996).

1) PHYLOGÉNIE DES GÈNES D'ACTINES ÉTUDIÉES

Les familles d'actines des espèces étudiées (Soja, Pomme de terre, Tomate, Maïs, Tabac) comportent de nombreux gènes. Le nombre de ces gènes n'est pas connu mais il est estimé à plusieurs dizaines (Moniz de Sá and Drouin 1996).

Un arbre phylogénétique incluant les gènes d'actines utilisés pour cette étude est représenté à la figure 2. Cet arbre a été obtenu à partir des séquences de nucléotides et en excluant le troisième site des codons. Cet arbre fut obtenu par la méthode Neighbor-Joining, à partir des distance phylogénétiques évaluées par la méthode de Kimura (Kimura 1980). Les familles d'actines végétales demeurent encore assez mal connues. Pour cette raison, l'arbre est incomplet et il est difficile d'identifier tous les gènes orthologues. On remarque cependant un regroupement des gènes de Soja dans un même rameau, qui pourrait laisser présager d'éventuelles conversions géniques. Le fait que la phylogénie des gènes d'actines étudiés soit à ce point incomplète rend l'utilisation de méthodes comme celle de Fitch, peu praticable.

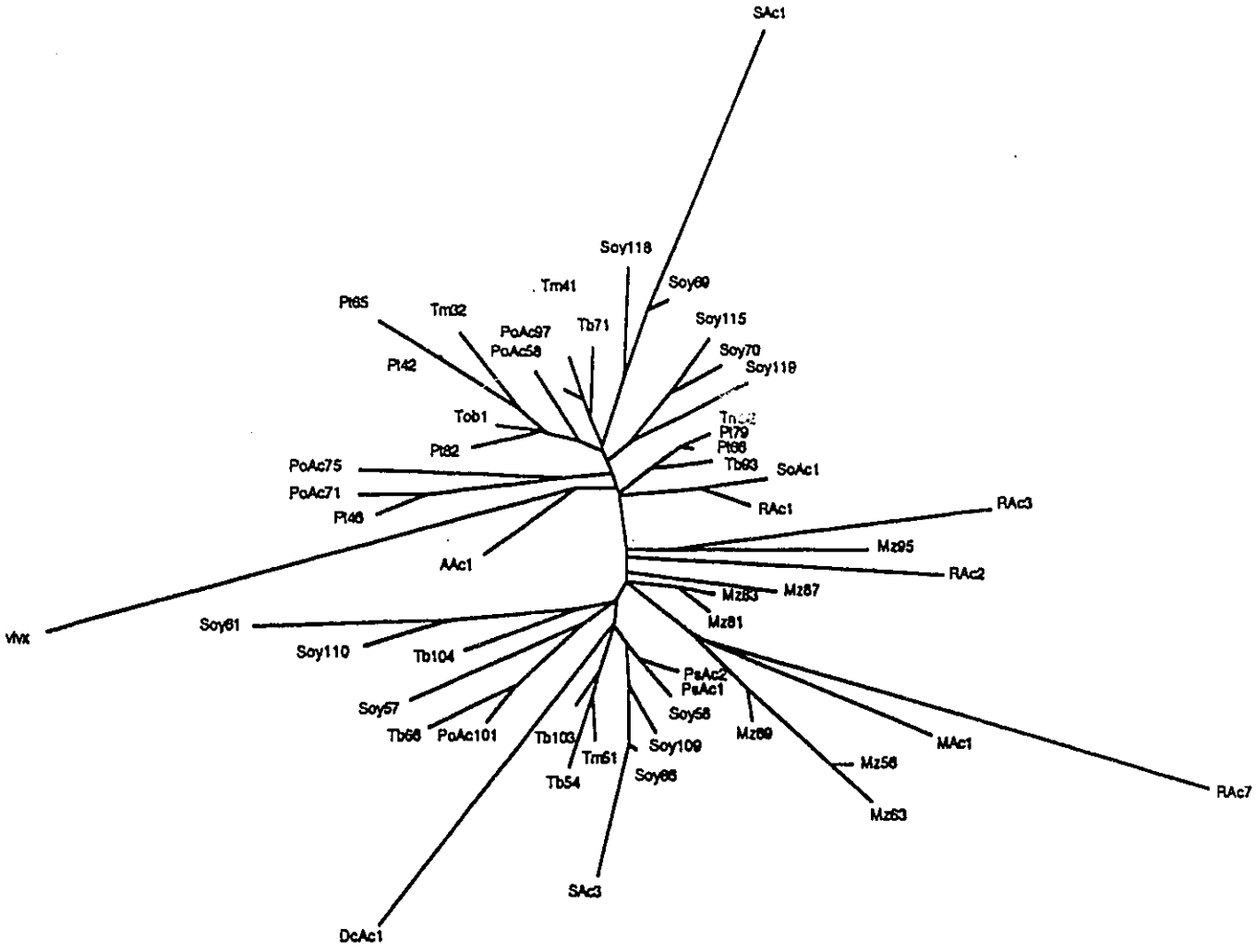
Figure 2. Phylogénie des gènes d'actines végétales Les distances phylogénétiques ont été évaluées par la méthode de Kimura en excluant les troisièmes sites des codons et l'arbre a été obtenu par la méthode Neighbor-joining (reproduit avec la permission de Marió Moniz de Sá)

vlvx : Volvox carteri, Mac, Mz : Zea maize, SAc, Soy : Glycine max,

RAc : Oryza sativa, DcAc : Daucus carota, SoAc : Sorghum vulgare

PsAc : Pisum sativum, AAc : Arabidopsis thaliana, Tm: Lycopersicon esculemtum

PoAc, Pt: Solanum tuberosum, Tb: Nicotiana tabacum


SAc1
Soy118
Tm41
Soy69
Pt65
Tm32
Tb71
Soy115
PoAc97
PoAc58
Soy70
Pt42
Soy119
Tob1
Tm42
Pt79
Pt82
Pt66
PoAc75
Tb93
SoAc1
PoAc71
RAc1
RAc3
Pt46
Mz95
AAc1
RAc2
Mz83
Mz87
Mz81
vivx
Soy61
Soy110
Tb104
PsAc2
PsAc1
Soy57
Soy58
Tb66
PoAc101
Tb103
Soy109
Mz89
MAc1
Tm61
Mz56
Tb54
Soy66
RAc7
Mz63
SAc3
DcAc1
.10 substitutions per site

2) MESURES STATISTIQUES

On a vu que les méthodes de simulation ont l'avantage de pouvoir être utilisées lorsqu'on dispose de peu d'information sur la famille étudiée. Les données ont donc été analysées avec la méthode de Sawyer, la méthode de Balding and al. (quelques orthologues sont identifiables) et la méthode des densités. Les résultats obtenus avec la méthode de Sawyer pour 20 000 données simulées sont indiqués au tableau 3. Les distributions ne sont certainement pas normales. Les moyennes et les écarts-types sont fournis à titre indicatif.

Tableau 3. Résultats du test de Sawyer pour les gènes d'actines

Espèce	Variable	Valeur observée	Moyenne	Écart-type	Valeur P	Valeur Z
Soja	SSCF	43310	40902	2311	0.14	1.04
	SSUF	674234	669787	20083	0.38	0.22
P. de terre	SSCF	72214	72995	7296	0.47	-0.11
	SSUF	724697	725621	25962	0.47	-0.04
Maïs	SSCF	23180	16247	1011	0.0003	6.86
	SSUF	410608	315757	15711	0.0002	6.04
Tomate	SSCF	1271	1199	59.5	0.12	1.21
	SSUF	41302	41061	1043	0.38	0.23
Tabac	SSCF	17314	16734	549	0.14	1.06
	SSUF	256323	251506	4417	0.13	1.09

SSCF = fragments condensés
SSUF = fragments non condensés

Contrairement à ce que l'on attendait, étant donné l'aspect de l'arbre de la Figure 2 les gènes d'actine de Soja donnent un résultat non significatif. Chez le Tabac, la valeur P obtenue pour les fragments non condensés est légèrement plus faible que la valeur P obtenue pour les fragments condensés. Cela vient de ce que beaucoup d'acides aminés sont communs aux deux séquences. On observe le même phénomène chez le maïs. Une similitude au niveau des acides aminés peut toujours s'expliquer en terme de contrainte fonctionnelle (de sélection naturelle). Elle ne peut pas être considérée comme l'indice d'une éventuelle conversion génique, d'où l'utilité de faire le test de deux façons : en utilisant les fragments condensés et en utilisant les fragments non condensés.

Lorsqu'il est appliqué au groupe des gènes de Maïs, le test de Sawyer donne un résultat très significatif (Valeur P = 0.0003).

Lorsqu 'une comparaison est faite entre les différentes paires de séquences et l'ensemble des données simulées on constate que dans plus de 99 % des cas les données simulées ont une valeur indicielle (SSCF) strictement inférieure à l 'indice de la paire maz 56/maz 81. On constate aussi que toute les données simulées ont une valeur indicielle inférieure à l 'indice de la paire maz 56/maz 63.

En appliquant la méthode des densités avec k = 1 à la paire maz 56/maz 81, on obtient un résultat très significatif (Valeur P = 0.0002). La région de densité minimale correspondant à cette valeur de k se situe entre les sites 827 et 954. Les résultats

obtenus suggèrent la présence d'une conversion génique d'environ 130 nucléotides, située à l'extrémité 3' du deuxième exon.

Il peut arriver que des conversions géniques se produisant au niveau de l'ADN soient à cheval entre un intron et un exon. Afin d'examiner cette éventualité, les introns bordant la zone identifiée furent alignés. On constata que les régions 5' des deuxièmes introns des gènes maz 56 et maz 81 étaient quasi identiques. Étant donné qu'en général, les introns ont une très grande variabilité, il paraît difficile d'expliquer ces similitudes autrement que par une conversion génique. L'examen des introns confirme donc l'hypothèse d'une conversion génique incluant la partie des exons située approximativement entre les sites 827 et 954.

En général, un résultat significatif doit être interprété avec prudence. Comme il s'agit de tests statistiques, un résultat significatif ne signifie pas forcément qu'il y a eu effectivement une conversion génique. Il constitue un indice supplémentaire. Le test de Sawyer et le test des densités sont effectués à partir des exons exclusivement. Dans le cas du gène du Maïs, une information supplémentaire, fournie par l'examen de l'intron bordant la zone identifiée, permet de valider les résultats des tests statistiques.

Lorsque le test de Sawyer est appliqué aux gènes de Maïs, on constate que pour la paire maz 56/maz 63, la valeur observée dans le cas des fragments condensés est extrêmement élevée (SSCF = 10364). Cela vient de ce que ces deux gènes ont

énormément de similitudes. Les fragments condensés correspondant sont très étendus. On observe une zone de conversion génique qui s'étend approximativement du début du premier exon à l'extrémité 3' du troisième exon (sites 1 à 825). Comme dans le cas précédent, l'intron 4 ne semble pas avoir été affecté mais l'examen des introns numéros 1, 2 et 3 confirme l'hypothèse d'une conversion génique de très grande ampleur. Les similitudes observées entre les deux gènes indiquent que la conversion génique détectée est probablement assez récente.

L'application de la méthode des codoubles nécessite, au préalable, l'identification de gènes orthologues. Un certain nombre de gènes orthologues ont pu être identifiés en comparant les arbres obtenus par différents types d'analyses phylogénétiques et en repérant les noeuds et les regroupements communs à tous ces arbres (Moniz de Sá and Drouin 1996). L'identification des orthologues à partir de données phylogénétiques fut corroborée par une analyse de type Southern (Moniz de Sá and Drouin 1996). Parmi les orthologues identifiés, un groupe de huit gènes provenant de trois espèces différentes (Pomme de terre, Tomate, Tabac) ont été comparés en utilisant la méthode de Balding and al. Les résultats obtenus sont présentés au tableau 4.

Tableau 4. Résultats du test de Balding et al. pour les gènes d'actines végétales. Comparaison pot 101 pot 97/tob 66 tob 71.

Catégorie	Nombre de sites	d	t + t'	r + r'	DT	DR	ST	SR
2	118	29	29	-	0	-	29	-
3	23	6	4	2	0	0	4	2
4	113	45	21	24	1	2	17	18

Catégorie	Mixtes	Codoubles (T)	Codoubles (R)	P(X = x)	Valeur P	Valeur P totale
2	-	0	-	0.4	1	
3	0	0	0	0.98	1	
4	2	0	0	0.32	1	
						1

Tableau 5. Résultats du test de Balding et al. pour les gènes d'actines végétales. Comparaison tom 52 tom 51/tob 93 tob 54.

Catégorie	Nombre de site	d	t + t'	r + r'	DT	DR	ST	SR
2	113	44	44	-	5	-	34	-
3	24	12	10	2	0	0	10	2
4	115	58	27	31	1	3	21	21

Catégorie	Mixtes	Codoubles (T)	Codoubles (R)	P(X = x)	Valeur P	Valeur P totale
2	-	4	-	0.1	0.14	-
3	-	0	0	0.99	1	-
4	4	0	0	0.26	1	-
						0.93

Comme on peut le constater, aucun de ces résultats n'est significatif. Lorsque les séquences tom 52 et tom 51 sont comparées aux séquences tob 93 et tob 94, on observe quatre codoubles parmi les 113 sites dégénérés deux fois. Ces résultats s'expliquent par le fait que, même si le nombre de sites synonymes est élevé, le nombre de différences est aussi élevé (d = 44). De plus, dans le cas des sites dégénérés deux fois, la probabilité d'obtenir un codouble est assez élevée car ce type de site ne peut varier que de deux façons. Comme l'indique le résultat non significatif obtenu (Valeur P = 0.14), la présence de ces codoubles ne peut pas être considérée comme un indice de conversion génique. Comme la méthode des codoubles est très sensible, il semblerait qu'aucun des gènes examinés n'ait été impliqué dans une conversion génique ou alors, avec des gènes extérieurs au groupe d'orthologues comparés.

Parmi tous les gènes d'actines étudiés, seuls les gènes de Maïs donnent un résultat significatif au test de Sawyer et au test des intervalles. Tous les tests effectués avec la méthode des codoubles et impliquant des séquences de Pomme de terre, de Tabac et de Tomate sont non significatifs. Les résultats obtenus pour la Pomme de terre confirment donc les analyses précédemment effectuées (Drouin and Dover 1990).

Il est possible que le regroupement des gènes de Soja sur l'arbre de la Figure 2 soit dû, en fait, à des duplications récentes. Le regroupement apparent des gènes de Soja pourrait aussi s'expliquer par le fait que certains orthologues non identifiés, qui

devraient normalement apparaître à côté des gènes de Soja ne sont pas pris en compte lors de la reconstruction phylogénétique.

B - GROUPE DES GÈNES DE GLOBINE

La super famille des globines a été étudiée de façon très approfondie et ses caractéristiques sont bien connues. On peut, néanmoins, rappeler brièvement quelques informations de base concernant cette famille.

La super famille des globines est composée de trois familles : la famille de la myoglobine, celle des globines de type α et celle des globines de type β. Chez les mammifères, un gène de globine typique comprend trois exons et deux introns dont les longueurs respectives sont d'environ 4800 pb et 3400 pb. L'hémoglobine est obtenue à partir des produits des familles α et β. Bien que la myoglobine et l'hémoglobine soient toutes les deux impliquées dans le transport de l'oxygène, elles n'ont pas exactement la même fonction. La myoglobine se retrouve dans les tissus musculaires et son affinité pour l'oxygène est beaucoup plus importante. L'hémoglobine a, en revanche, une structure plus élaborée et sa fonction est contrôlée de façon plus complexe. Ainsi, l'hémoglobine des mammifères a la capacité de réguler son affinité pour l'oxygène en fonction du niveau de phosphate organique dans le sang.

L'hémoglobine est constituée de deux chaînes de type α et de deux chaînes de type β. Chez des animaux comme le mouton, le lapin, la souris et, en fait, tous les mammifères étudiés à ce jour, on a constaté que les chaînes de type β ont plusieurs variantes possibles. Ces variantes sont toutes légèrement différentes et elles apparaissent à différents stades de l'existence de l'organisme. Ainsi, les variantes produites pendant l'embryogénèse sont différentes de celles que l'on observe chez l'adulte.

Chez les primates, les chaînes de globines sont disposés comme ci-dessous dans l'ordre de leur expression:

5' - ε - γ1 - γ2 - Ψη - δ - β - 3'

C'est le gène ε qui est exprimé en premier, chez l'embryon. Chez les hématies du foetus, on retrouve les chaînes de type γ1 et γ2. Plus tard, après la naissance et chez l'adulte, ce sont les chaînes δ et β qui sont exprimées (Slightom et al. 1988). Le gène Ψη est un pseudogène inactif (Fitch et al. 1990).

Chez l'homme, le groupe des globines de type β est inclus dans une région d'environ 6 Kb, située sur le chromosome 11. Le groupe des globines de type α comprend quatre gènes fonctionnels : ζ, α1, α2 et θ et les pseudogènes Ψζ, Ψα1 et Ψα2. En se combinant, les chaînes de types α et β forment différents types d'hémoglobines : ζ2ε2 et α2ε2 chez l'embryon, α2γ2 chez le foetus et α2β2 et α2δ2 chez l'adulte. Les conditions d'expression du gène θ ne sont pas encore connues (Li and Graur 1991).

1) PHYLOGÉNIE DES GÈNES DE GLOBINES

La séparation entre, d'une part, la myoglobine et, d'autre part, les familles α et β se serait produite il y a environ 600 à 800 millions d'années (Li and Graur 1991). La subdivision en proto ε et en proto β se serait produite chez les premiers mammifères il y a environ 200 ou 250 millions d'années (Koop et al. 1988, Fitch et al. 1990).

Une étude comparative des lagomorphes, des rongeurs, des artiodactyles et des marsupiaux (Koop et al. 1989) a permis de reconstituer la série ancestrale des gènes de globines de type β. Il semblerait que chez les mammifères euthériens, il y a environ 65 à 85 millions d'années, le groupe des globines β comprenait déjà les gènes ε, γ, η, δ, β. C'est après la séparation des simiens et des autres mammifères euthériens que η serait devenu un pseudogène (Fitch et al. 1990).

On estime que dans la lignée des euthériens, l'apparition des gènes δ et β daterait de 80 à 100 millions d'années. (Fitch et al. 1990). Les gènes δ et β auraient pour ancêtre un gène exprimé chez l'adulte, alors que les gènes ε, γ et η proviendraient d'un gène exprimé à l'origine chez l'embryon . C'est après la séparation des simiens et des prosimiens que l'expression du gène γ serait passée du stade embryonnaire au stade foetal.

Chez le lapin, la souris et les primates prosimiens comme Galago

Crassicaudatus, on ne trouve qu'un seul gène γ exprimé chez l'embryon (Tagle et al. 1988, Hayasaka et al. 1992) alors que chez les primates anthropoïdes, on trouve les gènes γ1 et γ2 (Slightom et al. 1988, Fitch et al. 1990), exprimés chez le foetus (Hayasaka et al. 1992). Il semblerait donc que le décalage de l'expression des gènes γ se soit produit après la séparation des simiens et des prosimiens (Tagle et al. 1988).

Les gènes γ1 et γ2 sont inclus dans une région d'environ 10 Kb. On pense qu'ils sont le résultat de la duplication d'une zone d'environ 5 Kb. Cette duplication se serait produite il y a environ 25 à 35 millions d'années dans l'embranchement des Catarrhini, après la divergence des Platyrrhini (singe du Nouveau-Monde), mais avant la divergence des Cercopithécoïdes (les singes de l'Ancien-Monde) et des Hominoïdes comme l'Homme, le Chimpanzé, le Gorille, l'Orang-outang, le Gibbon (Slightom et al. 1988, Fitch et al. 1990).

En observant les gènes de lapin et ceux de différents primates, on a remarqué certaines séquences très conservées qui pourraient être impliquées dans la régulation transcriptionelle. On a aussi observé certains changements , propres au groupe des Anthropoïdes, qui pourraient être responsables du décalage observé dans l'expression du gène γ (Tagle et al. 1988, Fitch et al. 1990).

En utilisant une méthode de reconstitution, site par site (Fitch et al. 1987, 1990), semblable à celle qui est décrite à la page 51, il fut possible d'identifier chez les

globines γ de nombreuses zones de conversion génique. Les données concernant les primates ne sont pas complètes mais en comparant les quatre espèces suivantes : Singe araignée, Singe rhésus, Gibbon et Humain, il fut possible d'identifier au moins une dizaine de zones de conversion génique (Fitch et al. 1990, Hayasaka et al. 1992).

Chez les mammifères, le gène β code pour la chaîne de type β que l'on trouve chez l'adulte. Le gène δ est généralement peu ou pas du tout exprimé. Chez tous les mammifères étudiés à ce jour, on a observé des conversions du gène δ par le gène β qui se seraient produites, il y a environ 40 millions d'années, chez un ancêtre anthropoïde (Koop et al. 1989).

En comparant les gènes δ de l'Orang-outang, du Rhésus macaque, du Singe araignée, du Chimpanzé, du Gorille, du Macaque et les gènes δ et β de l'Homme, du Tarsier, du Lémure et du Colobus, il fut possible de montrer l'existence de conversions géniques qui se seraient produites dans la lignée des Tarsiers il y a environ 30 millions d'années (Koop et al. 1989).

De nombreuses conversions géniques se sont produites, en plusieurs occasions, chez les globines. Comme les gènes γ1 et γ2 de l'Homme et de l'Orang-outang, les gènes δ et β de l'Homme et du Tarsier sont disponibles, il est possible de les utiliser pour une étude statistique.

2) MESURES STATISTIQUES

Chez les primates, la région codante des globines est très conservée. Elle est, de plus, très restreinte (quelques centaines de nucléotides). On constate, en effet, que chez une espèce donnée, les régions codantes des gènes γ1 et γ2 sont quasi identiques. Une situation analogue est observée lorsque les gènes δ et β sont comparés. D'après l'analyse de compatibilité, cette situation serait due à l'existence de zones de conversion très étendues, englobant la plupart du temps les deux premiers exons et parfois, tous les exons (Fitch et al. 1990).

Dans de telles conditions, le test de Sawyer et la méthode des densités, qui nécessitent un minimum de contraste entre les différentes parties de la chaîne d'exons, ne peuvent être appliqués. Comme les orthologues sont connus, le test des codoubles, qui se fonde sur les différences interspécifiques, peut, en revanche, être utilisé.

Le test des codoubles a été employé pour faire une comparaison entre les gènes γ1 et γ2 de l'Humain et de l'Orang-Outang (Pongo Pygmaeus) et entre les gènes δ et β de l'Humain et du Tarsier (Tarsier Syrichta). Les résultats obtenus sont présentés au tableaux 6 et 7.

Tableau 6. Résultat du test de Balding et al. Pour les gènes de globines γ1 et γ2.

Comparaison Humain/Orang-outang.

Catégorie	Nombre de sites	d	t + t'	r + r'	DT	DR	ST	SR
2	57	6	6	-	2	-	2	-
3	3	0	0	0	0	0	0	0
4	55	1	1	0	0	0	1	0

Catégorie	Mixtes	Codoubles (T)	Codoubles (R)	P(X=x)	Valeur P	Valeur P Totale
2	-	2	-	0.00089	0.0009	
3	0	0	0	-	-	
4	0	0	0	1	1	0.00088

Tableau 7. Résultats du test de Balding et al. pour les gènes de globines δ et β

Comparaison Humain/Tarsier (gènes δ et β)

Catégorie	Nombre de sites	d	t + t'	r + r'	DT	DR	ST	SR
2	49	19	19	-	5	-	9	-
3	0	0	0	0	0	0	0	0
4	50	14	10	4	1	2	8	0

Catégorie	Mixtes	Codoubles (T)	Codoubles (R)	P(X=x)	Valeur P	Valeur P Totale
2	-	4	-	0.0046	0.0049	
3	0	0	0	1	-	
4	0	1	2	0.000004	0.000008	

Dans le cas des gènes γ1 et γ2, les méthodes de compatibilité indiquent que chez l'Orang-outang, les zones de conversion s'étendent bien au-delà des régions codantes et englobent les trois exons. Chez l'Humain, les zones de conversion sont aussi très étendues à l'extérieur des régions codantes, mais elles ne recouvrent que les exons 1 et 2. Les résultats obtenus par la méthode des codoubles sont largement significatifs et ils corroborent donc ces observations.

Dans le cas des gènes δ et β et chez le Tarsier, les zones de conversion génique détectées par les méthodes de compatibilité recouvrent complètement le premier exon et la majeure partie du deuxième exon. La méthode de compatibilité indique l'absence de conversions géniques récentes chez l'Humain. Elle indique cependant l'existence de conversions géniques qui se seraient produites dans la lignée des Anthropoïdes après la séparation de la lignée du Tarsier de celle des Anthropoïdes.

Les valeurs obtenues avec le test des codoubles confirment donc les résultats fournis par la méthode de compatibilité.

La méthode de Fitch et la méthode des codoubles sont deux approches très différentes. Elles n'exploitent pas le même type d'information. Dans le cas des globines, les conclusions obtenues par la méthode de Fitch proviennent

essentiellement d'une étude des introns, alors que l'information utilisée par la méthode des codoubles provient exclusivement d'une comparaison entre exons. Le fait que ces deux méthodes permettent d'arriver aux mêmes conclusions rend très plausible l'hypothèse de conversions géniques chez les espèces considérées.

C - GROUPE DES GÈNES Zfx/Zfy

On sait que pour faire face aux besoins de la cellule, certains gènes doivent être activés, d'autres doivent être inhibés. L'expression des gènes dépend de protéines de régulation. Depuis la découverte de l'opéron lactose, de nombreuses protéines de régulation ont été identifiées chez les procaryotes. C'est seulement depuis une vingtaine d'années que des protéines jouant un rôle analogue ont été identifiées chez les eucaryotes.

Chez les eucaryotes, les arrangements moléculaires permettant une interaction des facteurs de transcription avec l'ADN sont de différents types. Parmi les structures identifiées, on peut mentionner les motifs helix-turn-helix, les homéodomaines, les tirettes à leucine. Un grand nombre de facteurs de transcription possèdent des petites projections appelées "doigts à zinc" qui leur permettent de se fixer à une région spécifique de l'ADN. C'est en 1985 que le premier facteur de transcription possédant des doigts à zinc fut identifié chez Xenopus (Miller et al. 1985) . Cette protéine (appelée TFIIIA) est un des facteurs nécessaires pour l'expression des gènes codant pour les ARN ribosomaux 5S. Depuis cette découverte, on a identifié au moins 200 protéines possédant des doigts à zinc. On appelle doigt à zinc la structure obtenue lorsqu'une chaîne polypeptidique se replie autour d'un atome de zinc, pour former une boucle. Chez TFIIIA, une paire de cystéines et une paire d'histidines se combinent avec un seul ion zinc, pour former la boucle caractéristique. L'ensemble est stabilisé par trois acides aminés hydrophobes. Le doigt ainsi obtenu constitue un minidomaine

capable de se fixer à une région spécifique de l'ADN. Une protéine peut posséder plusieurs modules de type "doigt à zinc". Les possibilités combinatoires de ces arrangements de modules permettent de générer, avec une économie de moyens, des protéines ayant toutes sortes de spécificités. Chaque protéine ainsi formée pourra se combiner à une région particulière de l'ADN.

Des gènes de doigts à zinc ont été identifiés sur les chromosomes sexuels des mammifères. Le gène ZFX, situé sur le chromosome X de l'Homme, a un paralogue sur le chromosome Y (ZFY). Chez les autres espèces euthériennes, les gènes équivalant à ZFX et ZFY sont désignés respectivement par les sigles Zfx et Zfy. Des études d'hybridation ont montré que le gène Zfx a un homologue chez les oiseaux (Page et al. 1987). Chez les reptiles (Valleley et al. 1992) et chez les marsupiaux (Sinclair et al. 1988), les gènes homologues ne sont, cependant, pas liés aux chromosomes sexuels. Comme les gènes ZFX et ZFY ont été identifiés chez tous les mammifères étudiés à ce jour (Page et al. 1987), il est raisonnable de considérer que les gènes Zfx et Zfy sont apparus après la divergence entre les marsupiaux et les mammifères euthériens et avant la radiation des espèces euthériennes (Pamilo et al. 1993). Les gènes de type Zfx/Zfy ont des caractéristiques très semblables. Les protéines qu'ils produisent sont constituées de deux domaines : une région amino-terminale très acide et une région carboxy-terminale incluant 13 doigts à zinc (Mardon et al. 1990, Palmer et al. 1990). Ces doigts à zinc sont de type Kruppel, avec un motif C2 H2 (deux cystéines, deux histidines).

Cette division en deux domaines, l'un acide, l'autre porteur de doigts à zinc, est typique. On la retrouve chez de nombreux activateurs transcriptionnels, comme les récepteurs à glucocorticoïdes et chez les protéines GCN4 de la levure (Godowski et al. 1988, Hollenberg et al. 1988).

1) RÔLES DES GÈNES ZFX ET ZFY

C'est en cherchant à identifier le facteur responsable du phénotype mâle chez les mammifères (appelé TDF : testis differentiating factor) que le gène ZFY fut identifié.

À cause de la délétion d'une petite partie du chromosome Y, certains individus porteurs des chromosomes X et Y sont, malgré tout, des femmes. On a aussi observé le cas d'individus de sexe masculin, porteurs de deux chromosomes X. Cette particularité serait due à une translocation, ayant pour résultat l'adjonction, à l'un des chromosomes X, d'une petite partie du chromosome Y. Chez certains individus de sexe masculin, une partie assez importante du chromosome Y peut être absente.

En étudiant les chromosomes de ces différents groupes d'individus et en les comparant à des chromosomes complets, il fut possible de montrer que la présence d'une région d'environ 140 Kb, appelée intervalle 1A2 et située sur le brin le plus court du chromosome Y, était nécessaire et suffisante pour déterminer la différenciation sexuelle (Page et al. 1987).

Afin d'identifier le facteur de différenciation testiculaire (TDF) et, partant du principe que les zones fonctionnelles sont généralement plus conservées au cours de l'évolution, une comparaison entre différentes espèces de mammifères fut effectuée. En procédant à des expériences d'hybridation, il fut possible de déceler, à l'intérieur de l'intervalle 1A2, une sous-région très conservée d'environ 1.3 Kb. En effectuant le séquençage de cette région, on s'aperçut qu'elle correspondait à un gène, codant pour des doigts à zinc (le gène ZFY). C'est à partir de ce type d'observation que certains auteurs suggérèrent que Zfy pourrait être le TDF (Mardon et al. 1989).

Les études d'hybridation indiquèrent la présence d'une région homologue au gène ZFY, située sur le chromosome X (le gène ZFX). Étant donné la présence de ce paralogue, le rôle de ZFY comme facteur de différenciation sexuelle paraît moins plausible. Il semblerait, d'autre part, qu'en dépit de leur similitude, Zfx et Zfy aient des fonctions assez distinctes. Des expériences effectuées chez la Souris ont montré que Zfx et Zfy ne sont pas exprimés dans les mêmes tissus. Alors que Zfx semble être exprimé dans une grande variété de tissus, aussi bien chez le mâle que chez la femelle, aussi bien chez l'adulte que chez le nouveau-né ou l'embryon, l'expression de Zfy n'a été détectée que dans la région des testicules de l'adulte (Mardon et al. 1990). Si on compare les longueurs de leurs régions non codantes, Zfx et Zfy diffèrent considérablement. Dans la région 3', la partie non traduite du gène Zfx est inférieure à 200 bases, alors que chez le gène Zfy, elle recouvre de 3 à 4 Kb.

Même si Zfx et Zfy semblent avoir des fonctions différentes, l'hypothèse de l'identification de ZFY au TDF paraît, à l'heure actuelle, peu vraisemblable. On pense que les mécanismes mis en jeu sont probablement beaucoup plus complexes. Certains auteurs admettent avec réticence l'hypothèse suivant laquelle le mécanisme de différenciation sexuelle des marsupiaux serait très différent de celui des mammifères. D'après eux, le fait que chez les marsupiaux, les gènes homologues à Zfy ne se trouvent que sur des autosomes, suggère que Zfy n'est probablement pas le TDF (Sinclair et al. 1988).Même si Zfy joue probablement un rôle dans la différenciation sexuelle des mammifères, il semblerait qu'il n'en soit pas le déclencheur (Valleley et al. 1992).

2) CARACTÉRISTIQUES DES GÈNES DE TYPE ZFX/ZFY

C'est surtout chez l'Homme et chez la Souris que les gènes de type Zfx/Zfy ont été étudiés.

Chez la Souris, la situation est un peu particulière. Son génome contient trois gènes paralogues répartis sur les chromosomes sexuels et appelés Zfy-1, Zfy-2 et Zfx et un gène paralogue de type autosomal, dénoté Zfa. Le locus Zfx n'est pas situé au même endroit chez la Souris et chez l'Homme.

La présence de deux copies du gène Zfy serait due à une duplication intrachromosomale qui aurait eu lieu pendant l'évolution des rongeurs (Mardon et

al.1989). Ce phénomène d'amplification du gène Zfy a été aussi observé, à un niveau encore plus élevé, chez le Lemming et chez les rongeurs sud-américains Sigmondotine (Bianchi et al. 1989). Zfy-1 et Zfy-2 sont identiques à 95% et ont probablement la même fonction (Mardon et al. 1990, Mitchell et al. 1989).

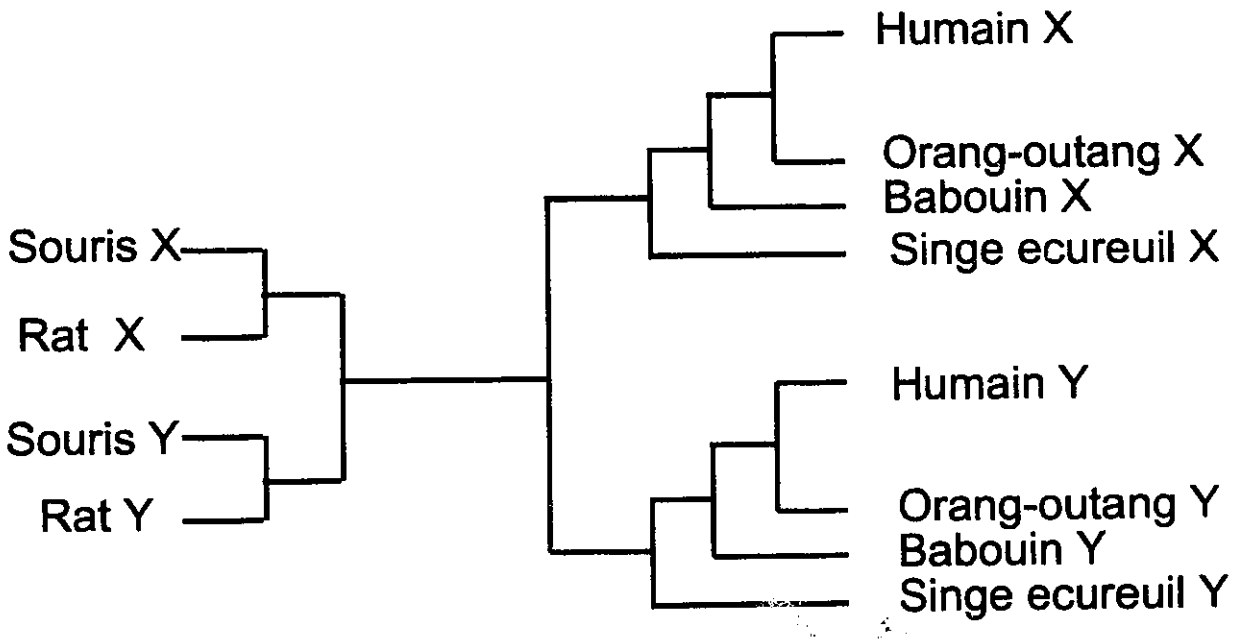
Alors que les gènes Zfy-1, Zfy-2 et Zfx ont des tailles comparables, Zfa est beaucoup plus petit que ses paralogues. Il semblerait que Zfa soit dérivé d'un transcrit de Zfx (Mardon et al. 1991). Le fait que ce gène n'ait pu être détecté chez d'autres mammifères placentaires laisse supposer qu'il serait apparu pendant l'évolution des rongeurs.

3) PHYLOGÉNIE

La découverte des gènes de type Zfx/Zfy est assez récente. Actuellement, seul un petit nombre de séquences de ces gènes sont disponibles et il est encore trop tôt pour tenter de reconstituer leur phylogénie complète.

Chez les gènes Zfx et Zfy, la partie de l'exon codant pour les doigts à zinc a récemment été séquencée chez un groupe de primates (Shimmin et al. 1994). Ces données permettent de tenter de reconstituer un arbre phylogénétique incluant des primates et des rongeurs. La phylogénie de la figure 3 a été obtenue par la méthode de parcimonie.

Figure3.Phylogénie des gènes Zfx/Zfy pour 6 espèces de mammifères. L'arbre phylogénétique a été obtenu par la méthode de parcimonie avec le programme DNApars de la boîte de logiciel PHYLIP (Felsenstein 1991).



Lorsqu'on utilise la méthode de Li (Li 1993), pour calculer les distances phylogénétiques, et la méthode Neighbor-joining (Saitou and Nei 1987) on obtient le même genre de topologie (Shimmin et al. 1994). On constate que toutes les espèces de primates sont regroupées pour former un seul rameau et sont séparées du Rat et de la Souris. Les phylogénies obtenues suggèrent une spéciation, suivie de deux duplications. Cette reconstruction ne cadre pas avec l'information disponible, car les gènes Zfx et Zfy ont été observés chez tous les mammifères placentaires étudiés à ce jour (Mardon et al. 1990, Lanfear and Holland 1991).

Une mesure des valeurs de Ks permet de déceler, chez les séquences de primates, deux zones à faible densité (voir pp.10 et 11). La présence de ces zones pourrait s'expliquer par l'existence de conversions géniques qui se seraient produites aux deux extrémités des séquences considérées (Shimmin et al. 1994).

4) MESURES STATISTIQUES

Afin de déterminer si les anomalies observées à la figure 3 et si la présence de zones à faible densité sont dues à des conversions géniques, on peut recourir aux méthodes statistiques décrites précédemment. Les gènes Zfx et Zfy devraient théoriquement pouvoir être analysés en utilisant le modèle de Balding et al. Une difficulté subsiste cependant. Un certain nombre d'auteurs pensent que les

chromosomes X et Y n'ont pas le même taux d'évolution (Miyata et al. 1987 cité par Shimmin et al.1994).

Chez les mâles, le nombre de divisions cellulaires observées dans la lignée germinale est beaucoup plus élevé que chez les femelles. Pour cette raison, Haldane suggéra que les taux de mutation observés chez les chromosomes X et Y devraient être différents. Un certain nombre de résultats obtenus en étudiant les gènes d'hémophilie et les introns des gènes de doigts à zinc vont dans le sens de la théorie de Haldane. Des différences notables ont été observées entre les taux d'évolution des gènes X et Y. Ces différences se manifestent au niveau des sites non synonymes et des sites synonymes. Les différences observées au niveau des sites non synonymes peuvent s'expliquer par le fait que le chromosome Y est soumis à des contraintes fonctionnelles moins importantes que le chromosome X (Shimmin et al. 1993 , Lanfear and Holland 1991). Les différences observées au niveau des sites synonymes peuvent difficilement s'expliquer de la même façon et semblent corroborer la théorie de Haldane (Shimmin et al. 1993 , Chang et al. 1994).

Même si la théorie de Haldane n'est pas encore parfaitement établie, il convient d'être prudent et il vaut mieux, autant que possible, utiliser un modèle qui ne présume rien des taux de mutation des chromosomes X et Y. On a vu que, pour que la méthode de Balding et al. puisse être appliquée, il faut que les taux de mutation des gènes étudiés soient comparables. Comme cette condition n'est pas remplie, le modèle de Balding et al. est inapplicable.

Pour étudier le groupe de gènes de doigts à zinc, on a donc choisi la méthode des taux variables. C'est le nombre de codoubles qui sert de variable aléatoire. La méthode peut être appliquée même si les taux de mutation sont différents chez les gènes comparés.

Les comparaisons ont été effectuées, d'une part, à l'intérieur du groupe des primates (entre le Babouin et l'Orang-outang) et, d'autre part, entre le groupe des primates et celui des rongeurs (entre le Singe écureuil et le Rat).

Il convient de remarquer que ces tests statistiques sont peu spécifiques. Ils permettent de déterminer si une conversion génique s'est produite depuis la séparation des deux espèces comparées mais ils ne permettent pas de déterminer chez laquelle des deux espèces, la conversion génique s'est produite.

Les résultats du test sont indiqués au tableau 8. Comme le test porte sur la répartition des différences chez les deux espèces d'animaux, il n'a pu être appliqué aux sites dégénérés trois fois. Il convient de faire remarquer que, le fait qu'une seule différence ait été détectée chez les sites dégénérés trois fois, n'a aucune incidence sur l'analyse.

Tableau 8. Résultats du test des codoubles (méthode des taux variables) pour les gènes Zfx/Zfy. Comparaison Singe écureuil/Rat.

Catégorie	Nombre de sites	d	t	t'	r	r'	DT	DR
2	200	80	36	44	-	-	12	-
3	8	1	-	-	-	-	0	0
4	74	47	19	10	10	8	4	3

Catégorie	ST	SR	Mixtes	Codoubles (T)	Codoubles (R)	Valeur P
2	56	-	-	11	-	0.000693
3	1	0	-	0	0	-
4	18	9	3	4	2	0.00203

Tableau 9. Résultats du test des codoubles (méthode des taux variables) pour les gènes Zfx/Zfy. Comparaison Orang-outang/Babouin

Catégorie	Nombre de sites	d	t + t'	r + r'	DT	DR
2	222	14	14	-	1	-
3	15	1	0	1	0	0
4	91	14	10	4	0	0

Catégorie	ST	SR	Mixtes	Codoubles (T)	Codoubles (R)	Valeur P
2	12	-	-	0	-	1
3	0	1	0	0	0	-
4	9	3	1	0	0	1

Dans le cas de la comparaison entre le Singe écureuil et le Rat, les valeurs P calculées pour les sites dégénérés deux fois (0.000693) et pour les sites dégénérés quatre fois (0.00203) sont très significatives. Ces résultats peuvent être confrontés à ceux obtenus pour l'Orang-outang et le Babouin. La comparaison de ces deux espèces ne révèlent la présence d'aucun codouble. L'hypothèse d'une conversion génique qui se serait produite depuis la divergence de l'Orang-outant et du Babouin paraît donc peu probable.

Lorsqu'elle est appliquée à des gènes ayant des taux de mutation différents on s'attend à ce que la méthode des taux variables donne une erreur de type II plus faible

que la méthode de Balding et al. Pour le groupe des gènes Zfx/Zfy les deux méthodes donnent des résultats comparables: les valeurs P obtenues avec la méthode de Balding et al. pour les sites deux et quatre fois dégénérés sont respectivement 0.0007 et 0.003. Pour les deux tests les résultats sont significatifs mais on ne peut rien conclure à partir de l'analyse d'un seul groupe de données. Pour comparer les deux méthodes il faudrait disposer de milliers de données simulées. Pour évaluer les erreurs de type II, il faudrait modéliser des conversions géniques se produisant chez des groupes de gènes ayant des taux de mutations différents.

Les résultats des tests suggèrent la présence d'une conversion génique qui se serait produite après la divergence entre les rongeurs et les primates. Ces tests ne permettent pas de déterminer si la conversion génique s'est produite chez le groupe des rongeurs, chez celui des primates ou chez les deux groupes. On peut se demander aussi si la conversion génique s'est produite au niveau des espèces ancestrales ou, plus récemment, au niveau des espèces dérivées. Enfin, il serait intéressant de pouvoir déterminer la polarité de la conversion génique.

5) ANALYSE SITE PAR SITE

Pour tenter de répondre à toutes ces interrogations, nous avons effectué, à partir de l'hypothèse de parcimonie, un patient travail de reconstruction phylogénétique. Cette reconstitution, site par site, fut réalisée en adoptant la méthode décrite à la page 51. La phylogénie utilisée pour faire cette reconstitution est présentée à la figure 4. Les résultats de cette analyse sont présentés au tableau 10.

Figure4. Arbre d'éspèces utilisé pour l'analyse site par site des gènes Zfx/Zfy.

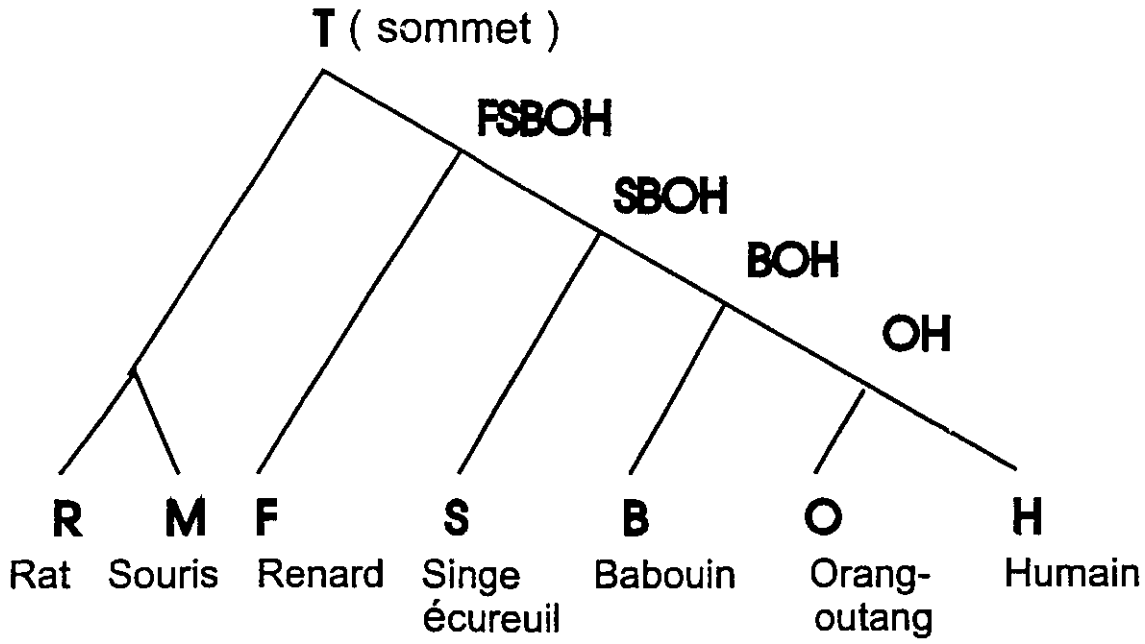


Tableau 10. Analyse site par site des gènes Zfx/Zfy

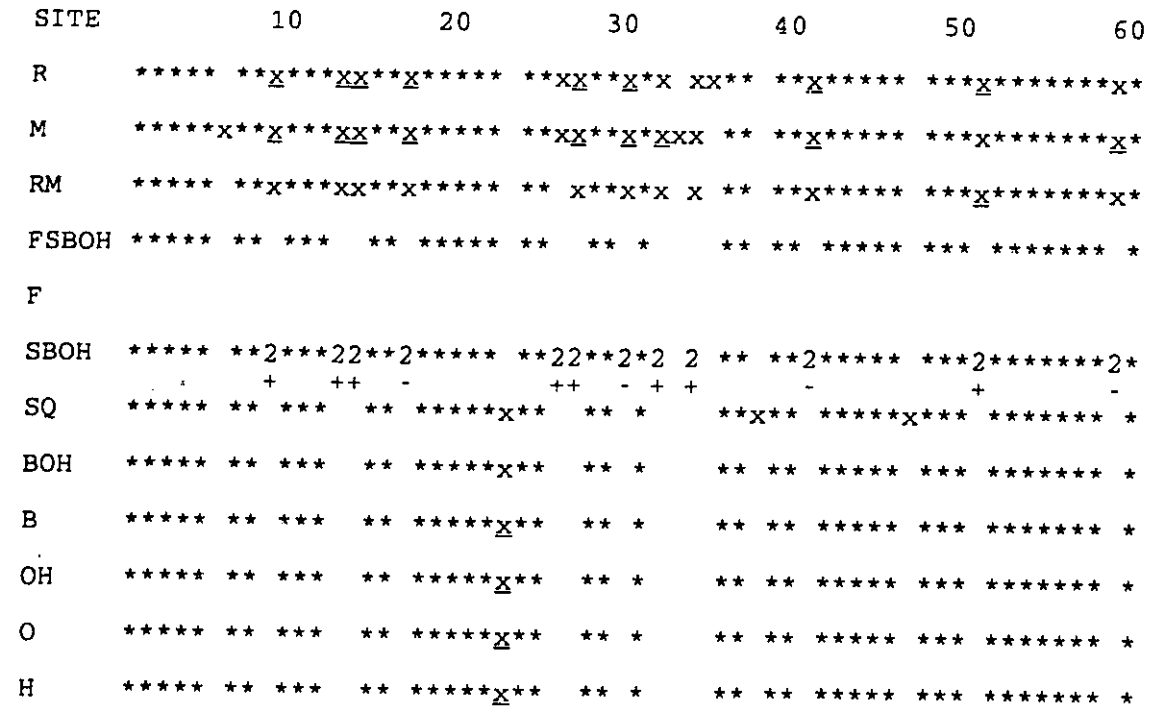
Explication des symboles (voir aussi la section "Méthode"):

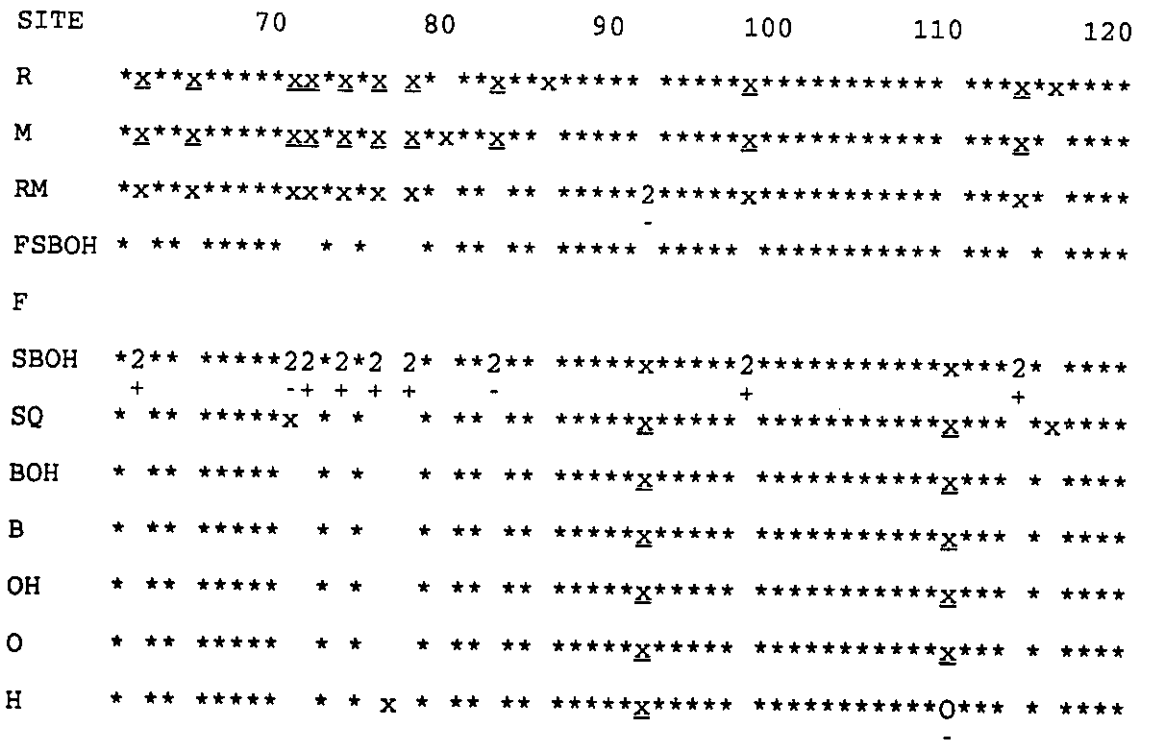
1, 2 2, 3 + -	: désignent les sites comportant une indétermination et correspondant à l'hypothèse d'une conversion génique
O O O + -	: sites entièrement déterminés (par la méthode de parcimonie) et correspondant à l'hypothèse d'une conversion génique
x x	: sites correspondant à l'hypothèse d'une absence de conversion génique
*	: les colonnes d'étoiles indiquent les nucléotides identiques pour toutes les espèces
R	: Rat (Rattus norvegicus)
M	: Souris (Mus musculus)
F	: Renard (Dusicyon thous)
SQ	: Singe écureuil (Siamiri boliviensis)
B	: Babouin (Papio cyanocephalus)
O	: Orang-outang (Pongo pygmaeus)
H	: Humain (Homo sapiens)

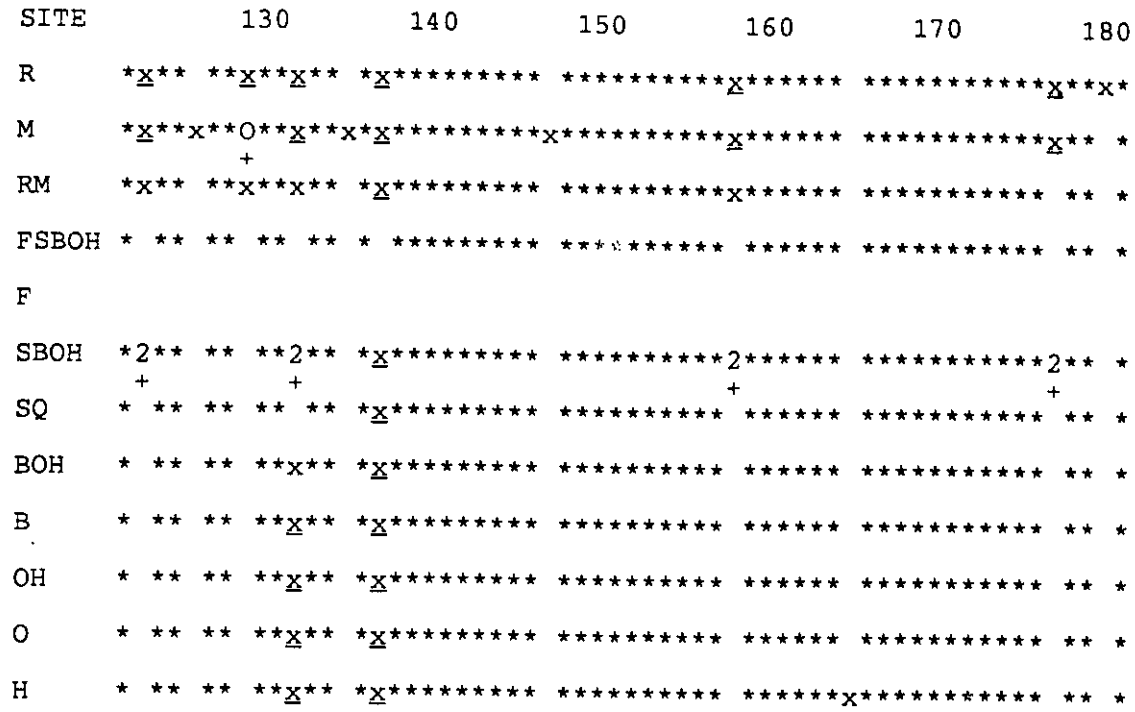
Voir aussi la figure 4.

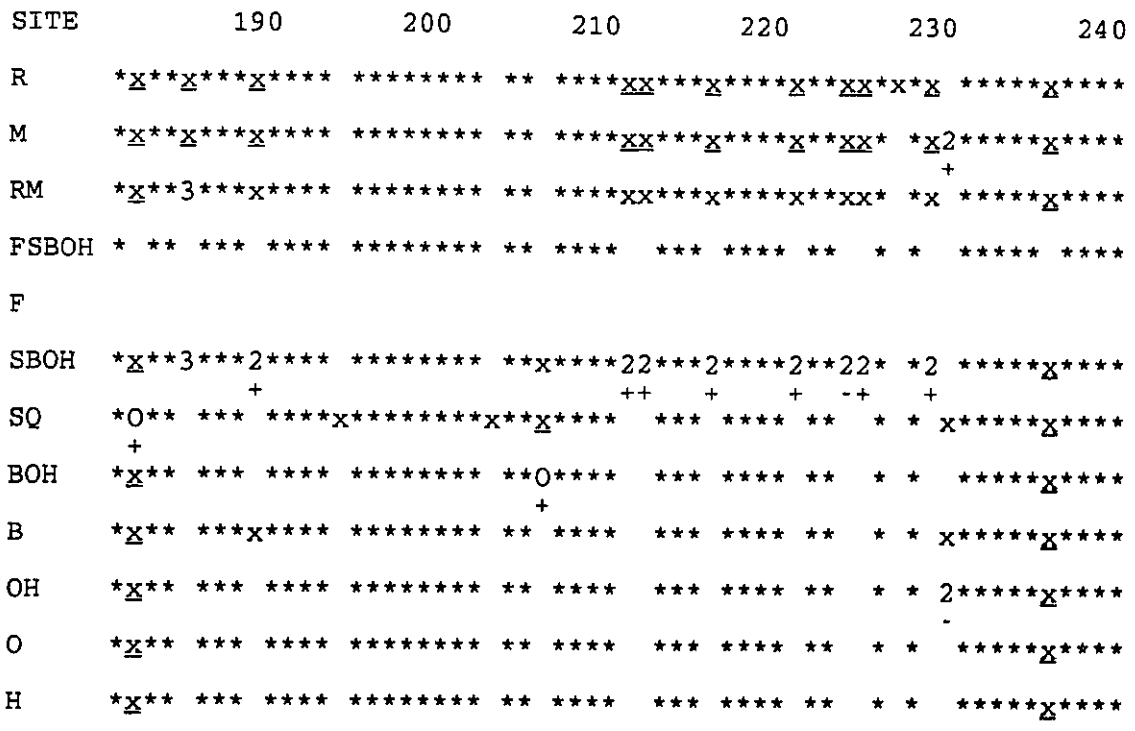
Les signes + et - indiquent, respectivement , une conversion de X vers Y et une conversion de Y vers X.

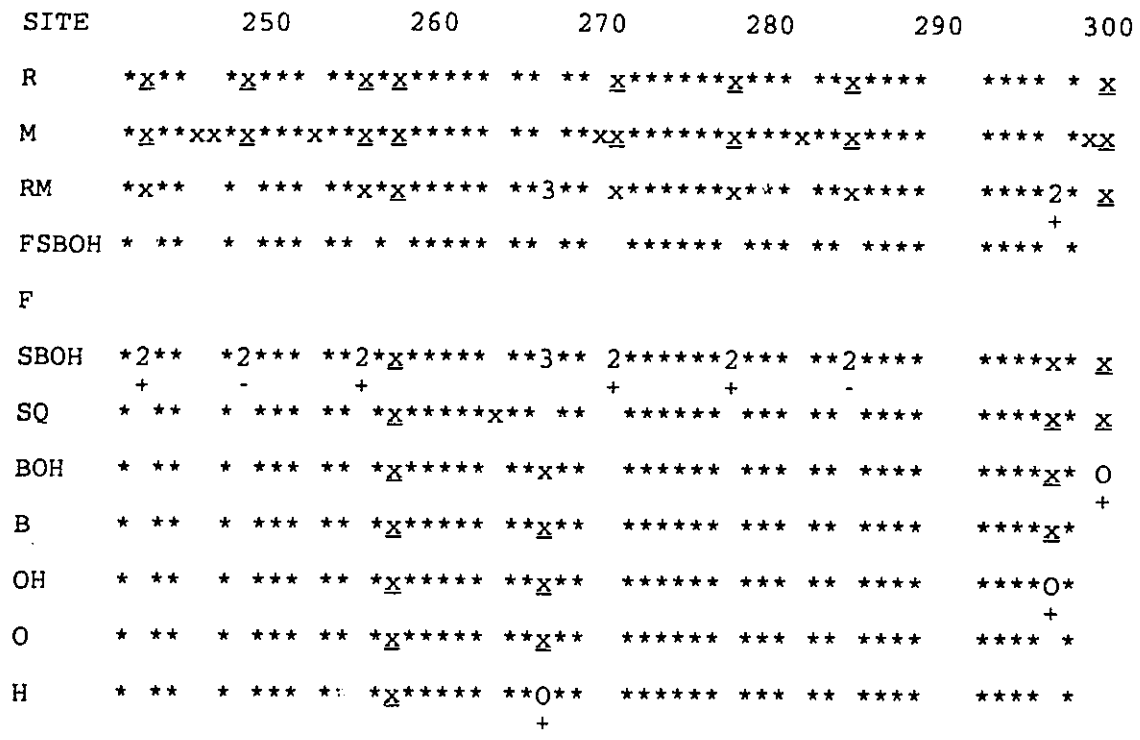
L' alignement et le système de numérotation des sites sont ceux de Shimmin et al.(1994)

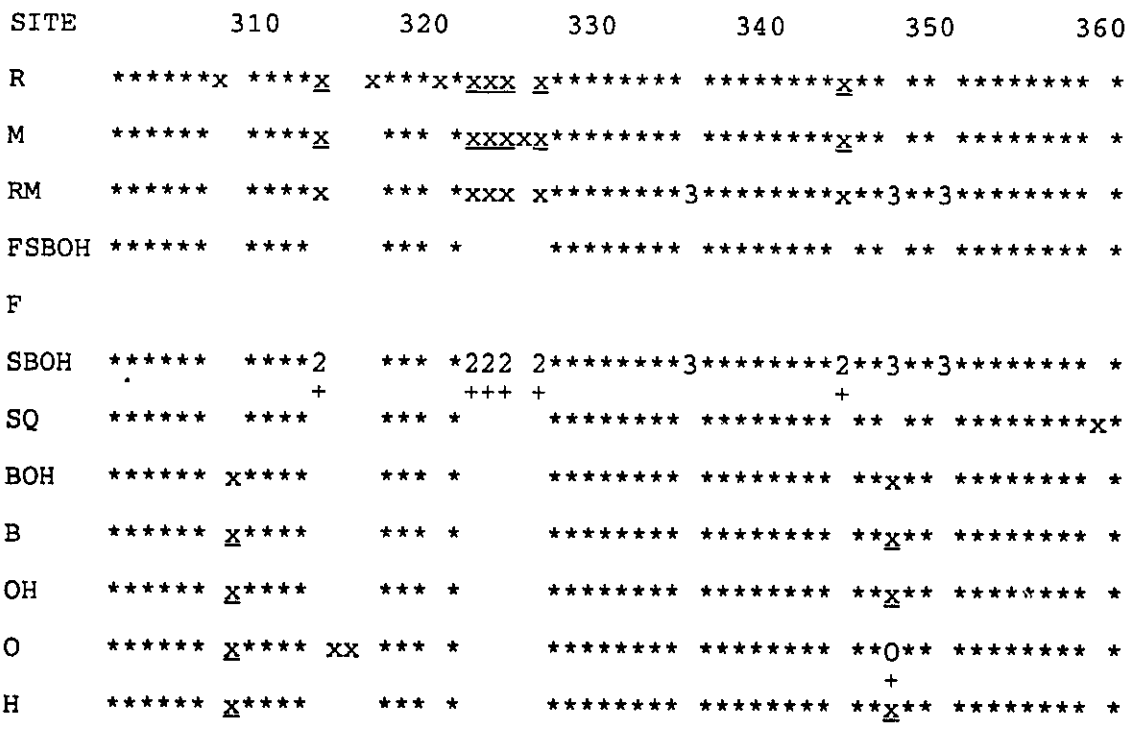

SITE 10 20 30 40 50 60
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

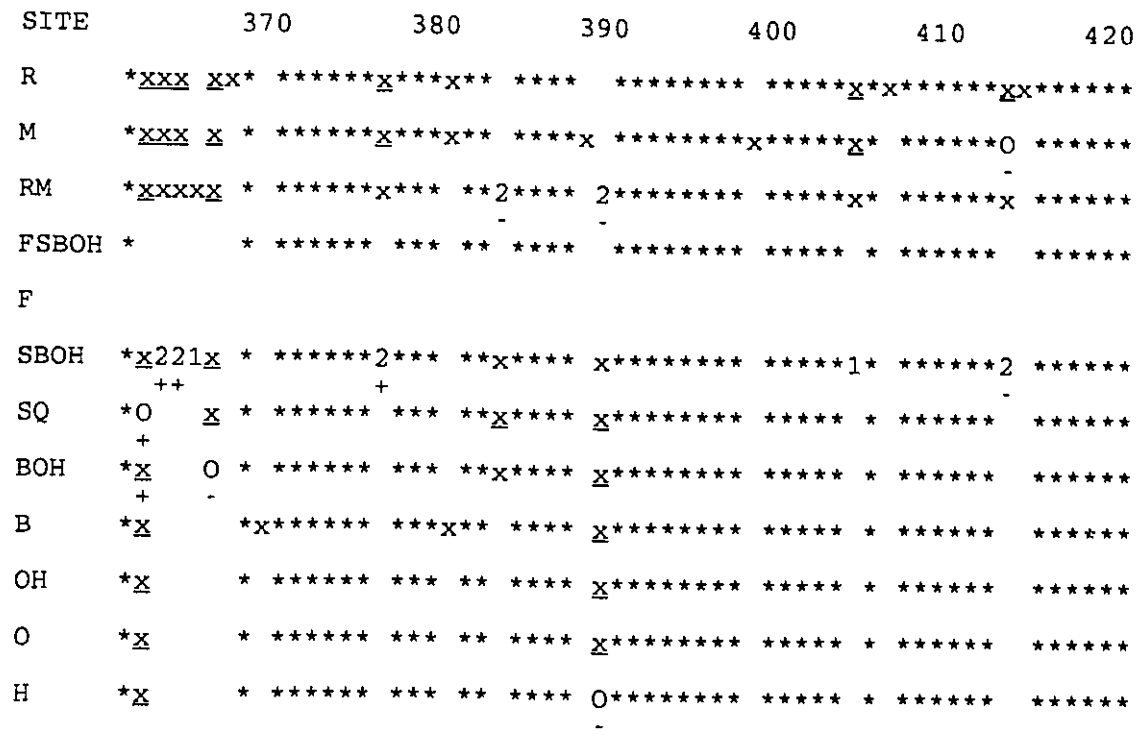

SITE 70 80 90 100 110 120
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

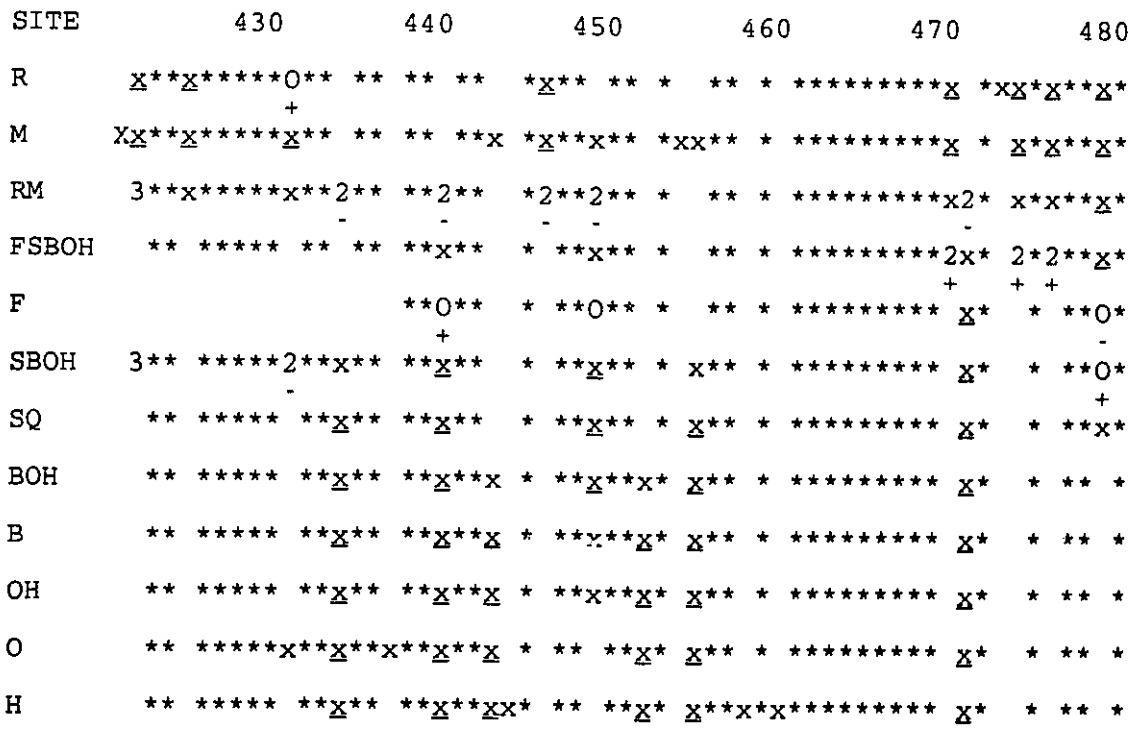

SITE 130 140 150 160 170 180
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

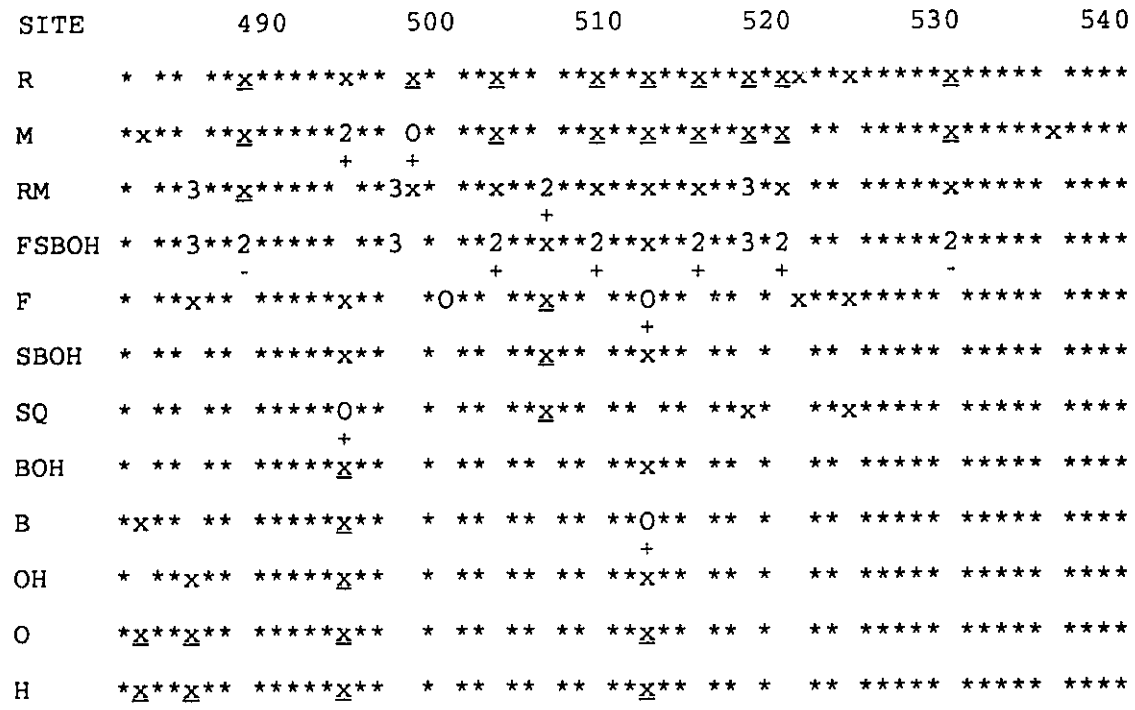

SITE 190 200 210 220 230 240
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

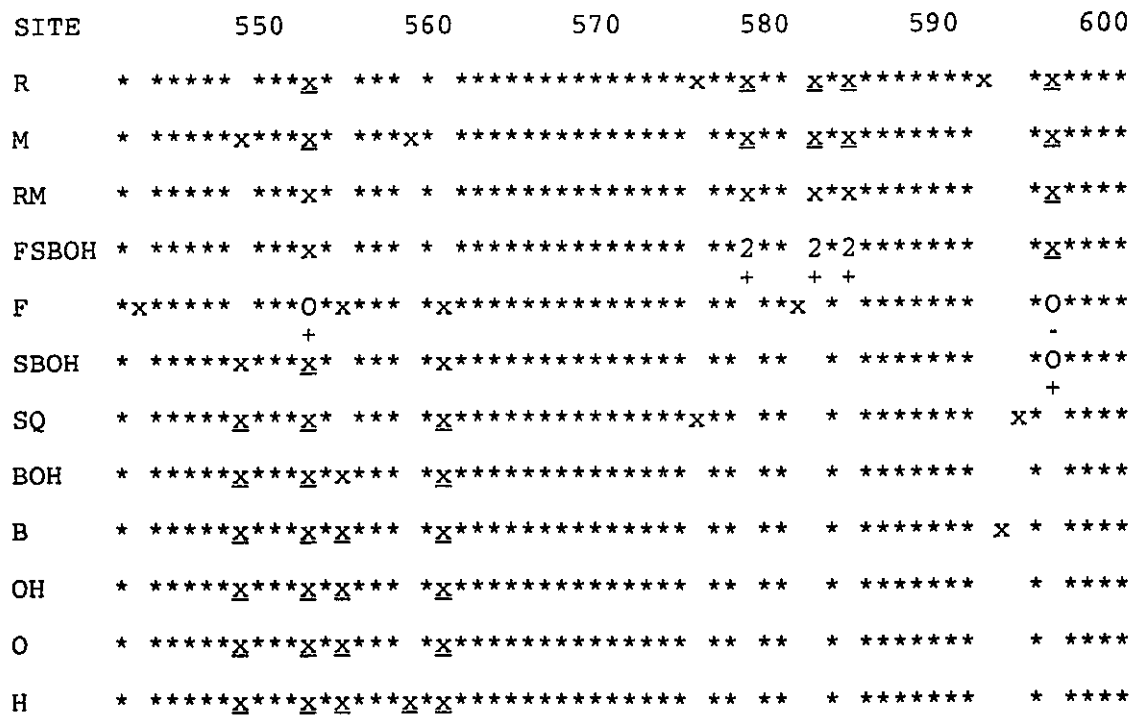

SITE 250 260 270 280 290 300
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

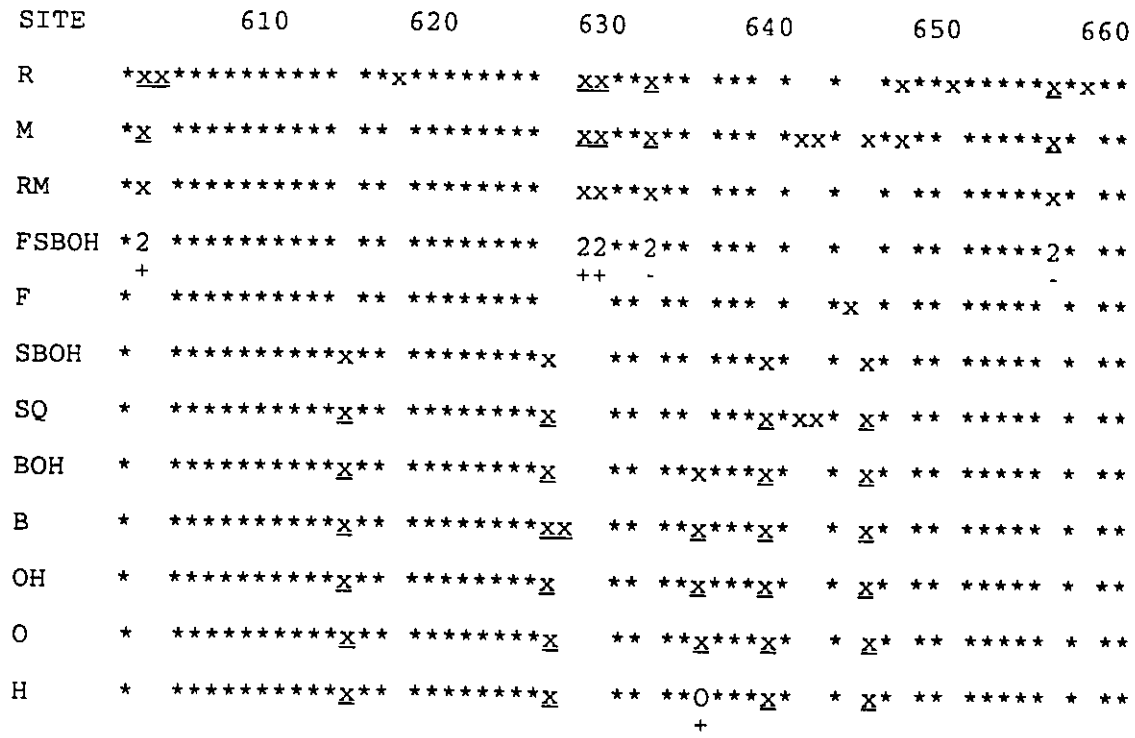

SITE 310 320 330 340 350 360
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

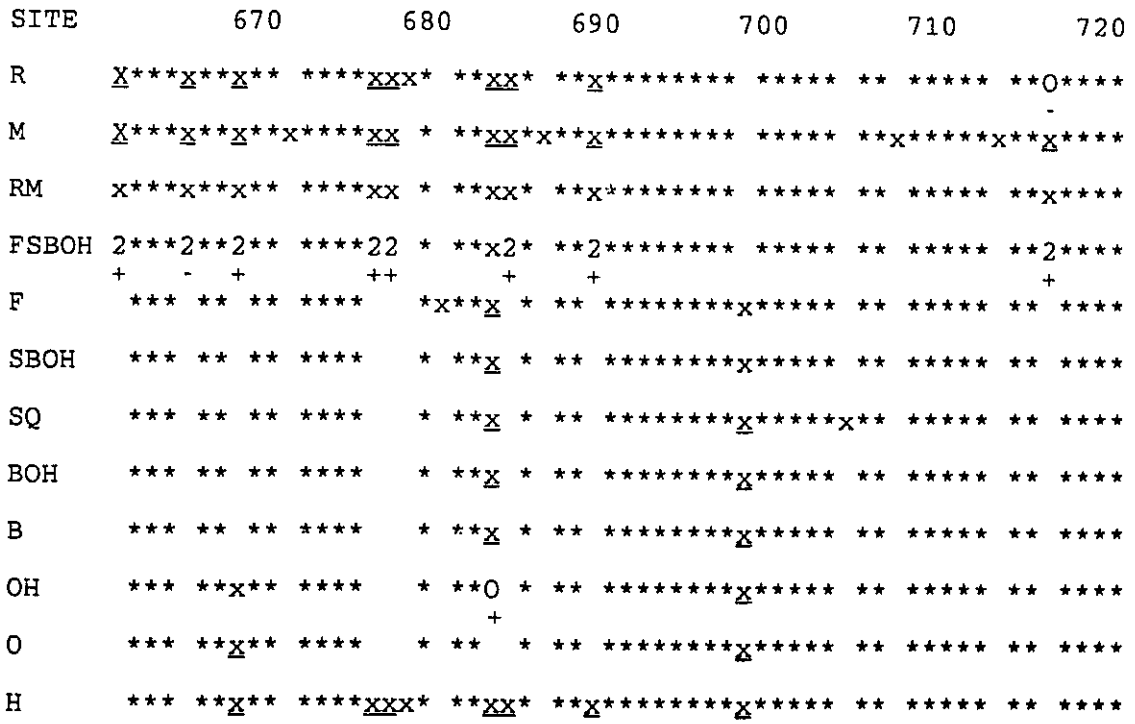

SITE 370 380 390 400 410 420
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

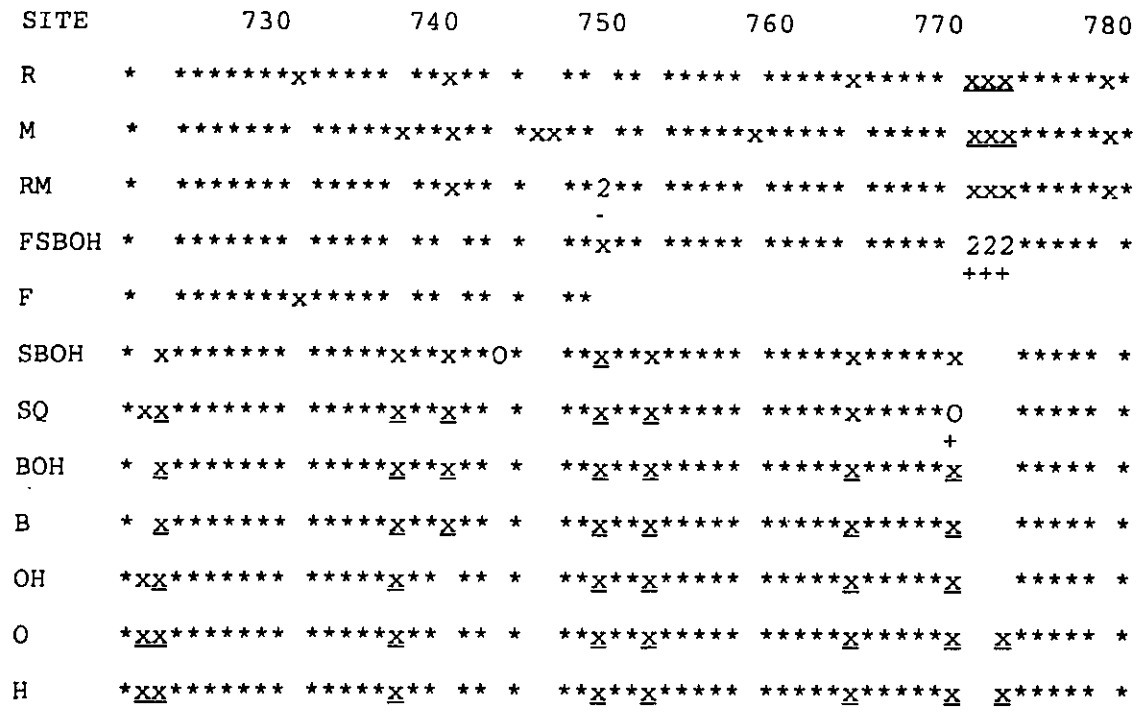

SITE 430 440 450 460 470 480
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

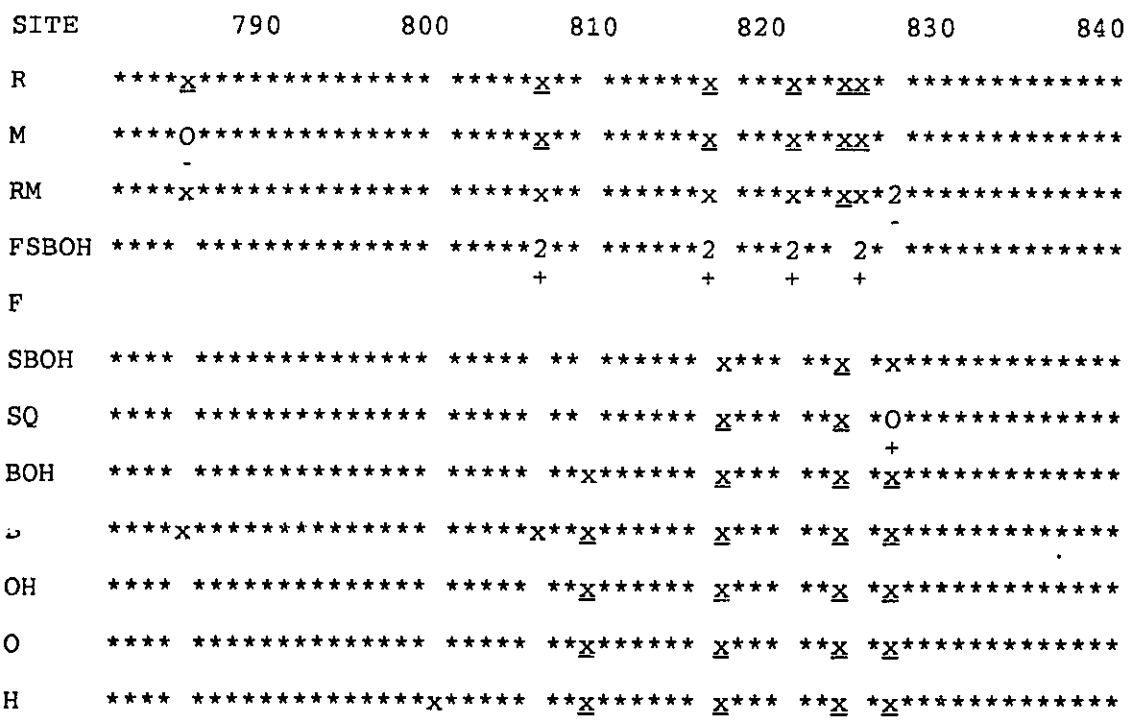

SITE 490 500 510 520 530 540
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H


SITE 550 560 570 580 590 600
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H


SITE 610 620 630 640 650 660
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

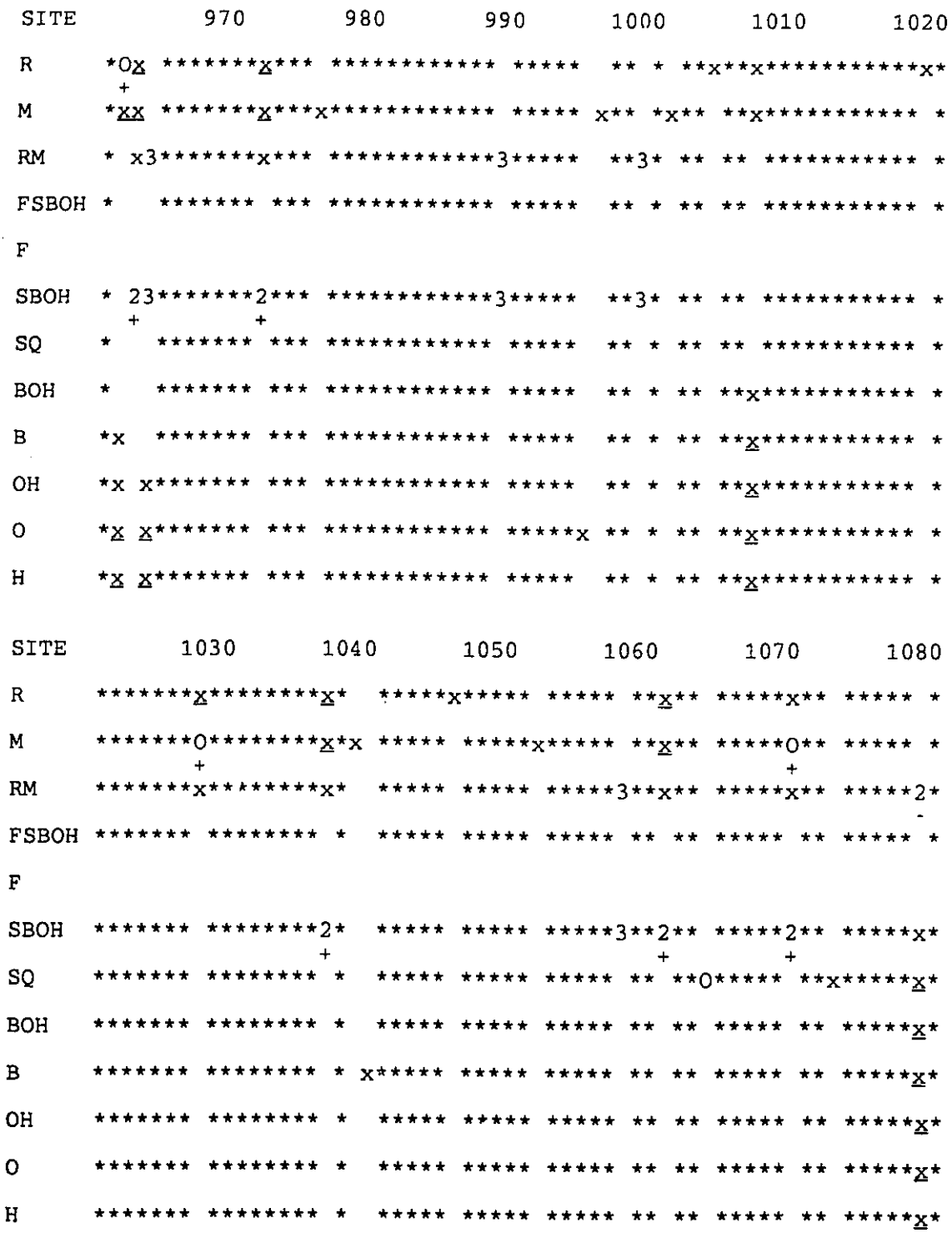

SITE 670 680 690 700 710 720
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

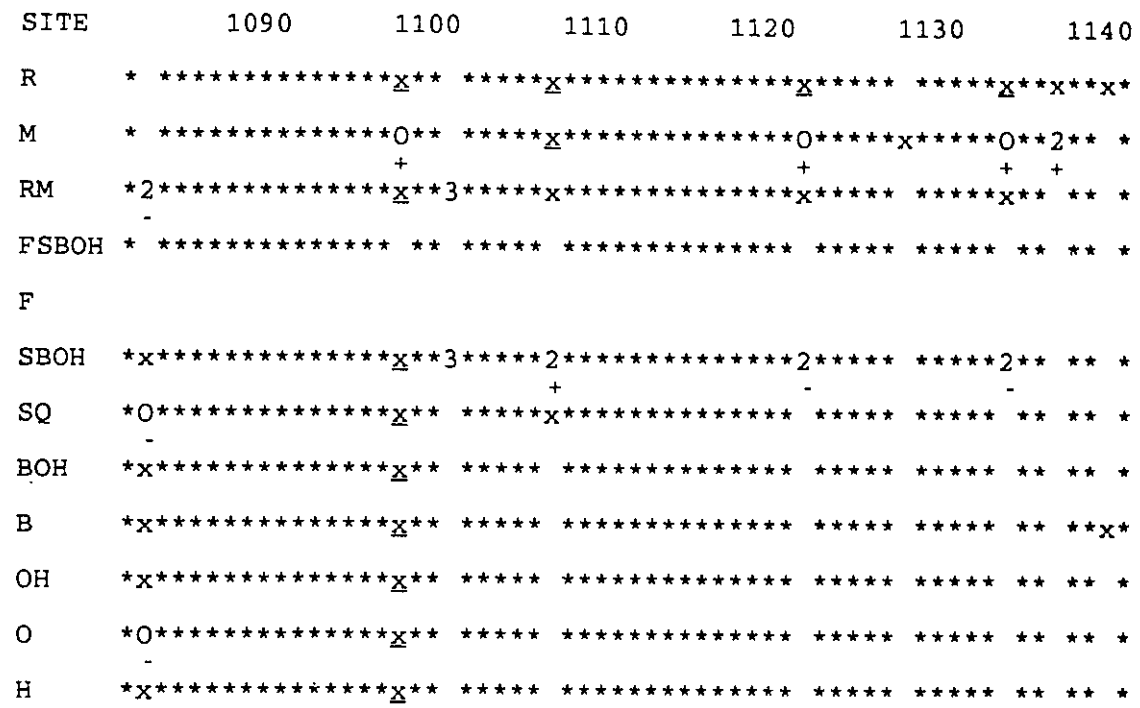

SITE 730 740 750 760 770 780
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

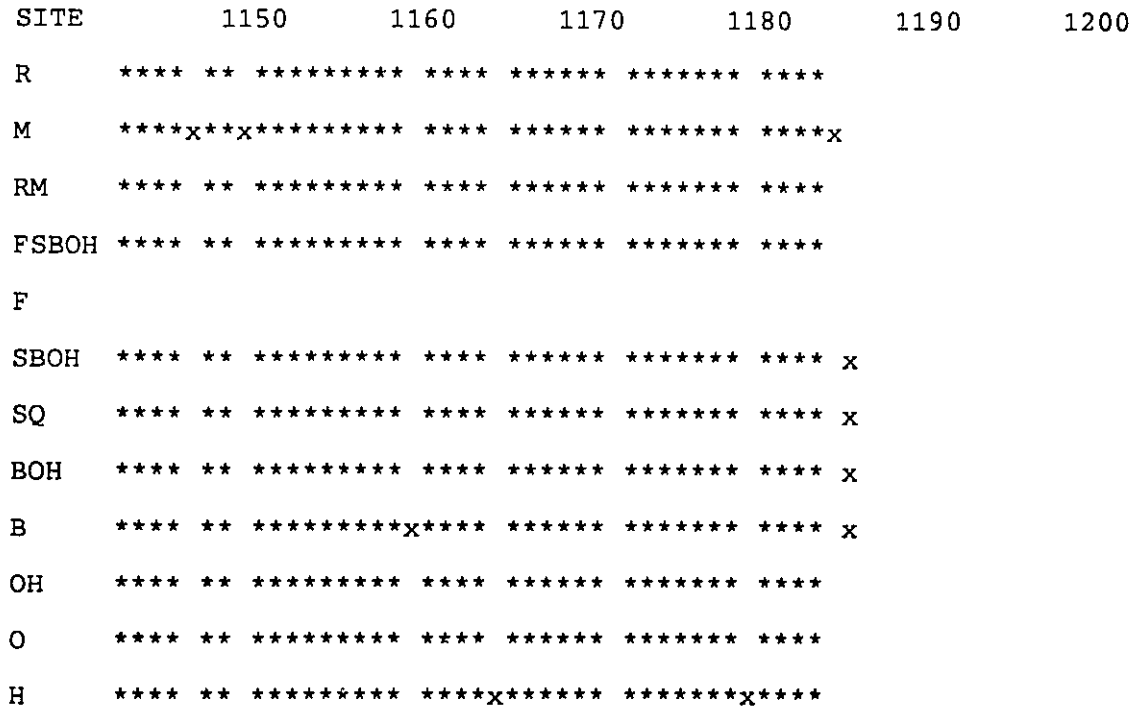

SITE 790 800 810 820 830 840
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

```
SITE          850        860        870        880        890        900
R      *x********x********** *x*x**** **x***x*x** **** **********
M      *x********x**********x*x*x**** **x***x*x**x****x**********
RM     *x********x********** *x*x****2**x***x*x** **** **********
FSBOH  * ******** ********** * * **** ** *** * ** **** **********
F
SBOH   *x********2********** *2*2****x**x***2*1** **** **********
SQ     *x******** ********** * * ****x**x*** * ** **** **********
BOH    *x******** ********** * * ****x**x*** * ** **** **********
B      *x******** ********** * * ****x**x*** * ** **** **********
OH     *x******** ********** * * ****x** *** * ** **** **********
O      *x******** ********** * * ****x** *** * ** **** **********
H      *x******** ********** * * ****x** *** * ** **** **********

SITE          910        920        930        940        950        960
R      *x* ***x******x*xx**** ** *x*** **********x*** * ***   x*****
M      *x*x***x******x*xx**** **x*x*** **********x***x*x***  xx*****
RM     *x* ***x******x*xx**** ** *x*** **********x*** * *** 3 x*****
FSBOH  * * *** ****** *  **** ** * *** ********** *** * ***    *****
F
SBOH   *2* ***2******2*22**** ** *2*** **********2*** * *** 3 2*****
SQ     * * *** ****** *  ****x** * *** ********** *** * ***x   *****
BOH    * * *** ****** *  **** ** * ***x********** *** * ***    *****
B      *x* *** ****** *  ****x** * ***x********** *** * ***    *****
OH     *x* *** ****** *  **** ** * ***x********** *** * ***    *****
O      *x* *** ****** *  **** ** * ***x********** *** * ***    *****
H      *x* *** ****** *  **** ** * ***x********** *** * ***    *****
```


SITE 970 980 990 1000 1010 1020
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H
SITE 1030 1040 1050 1060 1070 1080
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H


SITE 1090 1100 1110 1120 1130 1140
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

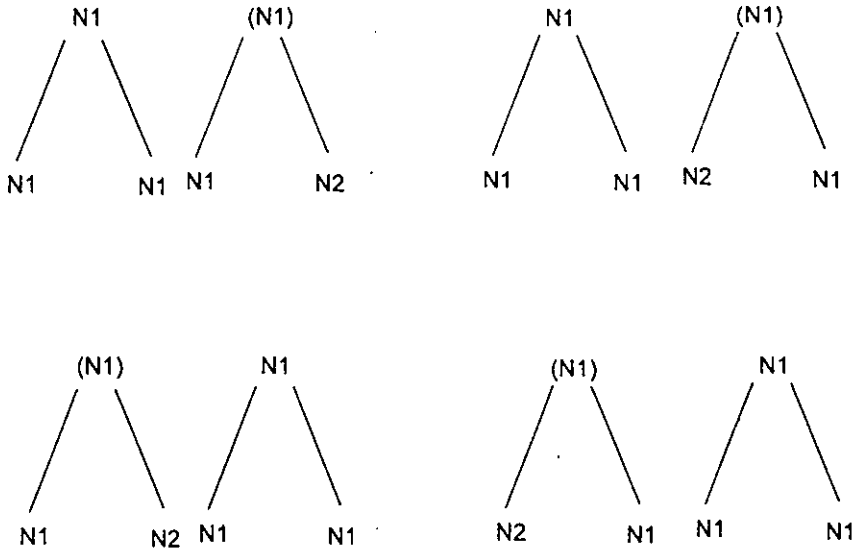

SITE 1150 1160 1170 1180 1190 1200
R
M
RM
FSBOH
F
SBOH
SQ
BOH
B
OH
O
H

Un premier résultat de l'analyse, site par site, est le fait qu'aucune conversion génique ne soit détectable entre les séquences Zfx et Zfy du Rat ou des espèces récentes de primates.

Dans l'arbre d'espèces, les sites SBOH et FSBOH sont au niveau le plus élevé de la hiérarchie (fig.4). Pour cette raison, ils correspondent le plus souvent à des cas indéterminés. Même si à ce niveau de la hiérarchie, l'analyse site par site donne de moins bons résultats, elle permet parfois de fournir des informations utiles. On a vu que l'hypothèse d'une conversion génique est retenue lorsque deux nucléotides correspondant au même site sont identiques à l'un des deux nucléotides ancestraux. Dans ces conditions, à ce niveau de la hiérarchie, les sites favorisant l'hypothèse d'une conversion génique correspondront forcément à une indétermination. L'examen des données indique que la majorité des sites sont de catégorie 2 et que les sites qui ne correspondent pas à une indétermination sont peu nombreux. Ces quelques sites correspondent, bien entendu, à l'hypothèse d'une absence de conversion génique.

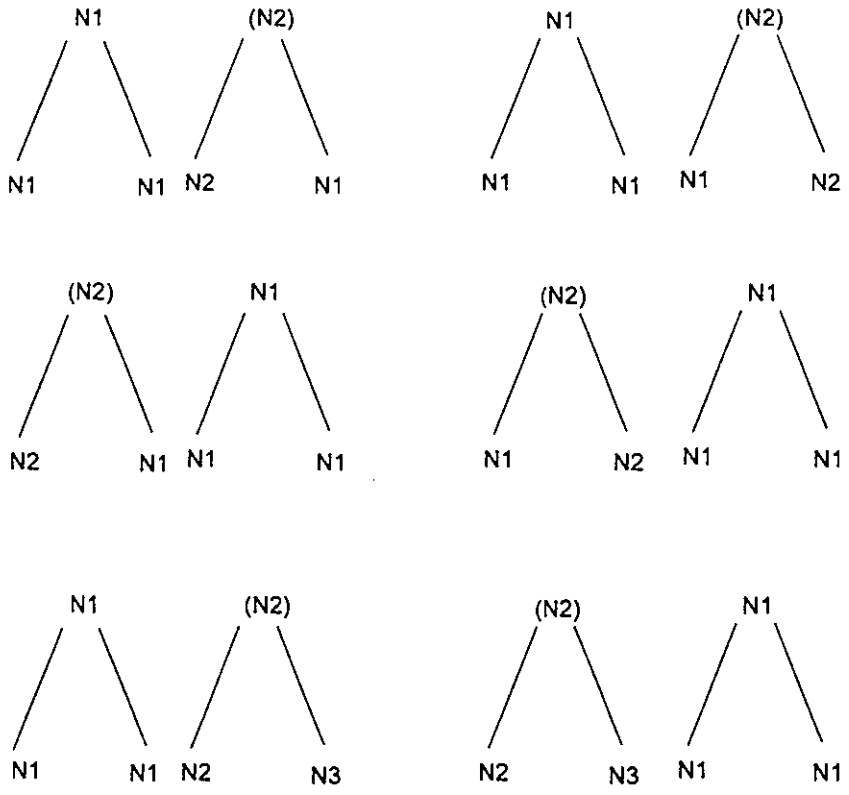
L'examen détaillé des arrangements observés au niveau des sites SBOH et FSBOH révèle certaines disparités. Considérons le niveau le plus élevé de la hiérarchie dans l'arbre des espèces (le point T). Au point T, et pour un site donné, on a deux nucléotides correspondant chacun à l'un des deux gènes étudiés (X ou Y). Ces deux nucléotides sont soit identiques, soit différents.

Considérons les arrangements partant du point T et qui, d'après l'hypothèse de parcimonie, ne comportent qu'une seule substitution. Dans le cas où le nucléotide provenant du gène X et le nucléotide provenant du gène Y sont identiques, on a quatre arrangements possibles :



N1 et N2 représentent deux nucléotides différents.

Dans le cas où les deux r cléotides correspondant au point T sont différents, on a six arrangements possibles :



Si on ne considère que les nucléotides récents (identifiables par reconstruction), ces dix arrangements appartiennent donc à six catégories :

(N1N1) (N2N1), (N1N1) (N1N2), (N2N1) (N1N1), (N1N2) (N1N1), (N1N1) (N2N3), (N2N3) (N1N1). Tous ces arrangements devraient être observés dans les données. On constate

cependant certaines disparités.

Parmi les quatre arrangements correspondant aux cas indéterminés de la catégorie 2 : (N1N1N1N2, N1N1N2N1, N1N2N1N1, N2N1N1N1), on observe une très nette prépondérance des arrangements de types N1N1N2N1 et N2N1N1N1. Sur les 117 arrangements de type 2 recensés, 102 appartiennent à ces catégories. Cette prépondérance pourrait s'expliquer soit par des phénomènes de conversion génique, soit en supposant des taux d'évolution différents dans le groupe des primates et dans celui des rongeurs, soit par un effet cumulé de ces deux types de phénomènes.

À l'intérieur de cette sous-classe d'arrangements, on pourrait aussi expliquer la prépondérance des arrangements de type N1N1N2N1 (ils représentent 82% de cette sous-classe), en supposant que les gènes portés par le chromosome Y ont un taux d'évolution plus élevé que les gènes situés sur le chromosome X.

L'inconvénient de cette hypothèse est qu'elle ne permet pas d'expliquer la prépondérance des arrangements de types N1N2N1N1, par rapport aux arrangements de types N1N1N1N2. Sur les 15 sites appartenant à ces deux catégories, 12 sont de type N1N2N1N1.

On peut aussi expliquer la prépondérance des arrangements de type N1N1N2N1 par l'effet d'une conversion génique à polarité X-Y. Cette conversion n'affectant que les

noeuds SBOH et FSBOH, elle est parfaitement compatible avec une prépondérance des arrangements de type N1N2N1N1.

Une autre anomalie observée est l'absence d'arrangements de type N1N1N2N3 ou N2N3N1N1. Cette situation correspond au cas où les nucléotides correspondants au sommet T sont différents. Le fait que ces arrangements n'apparaissent même pas au niveau des sites synonymes dégénérés quatre fois est difficile à interpréter. On pourrait expliquer cette anomalie en supposant que tous les sites de niveau T correspondant aux arrangements examinés étaient, au départ, identiques chez les deux gènes. À moins de supposer que la duplication des gènes Zfx et Zfy s'est produite juste avant la divergence des primates et des rongeurs, cette explication paraît improbable. Une autre explication serait la possibilité d'une conversion génique qui se serait produite entre le point T et le point SBOH ou FSBOH.

Les résultats des tests statistiques suggèrent qu'une conversion génique s'est produite après la divergence des rongeurs et des primates, soit dans la lignée du Rat, soit dans celle du Singe écureuil. L'analyse site par site suggère une absence de conversion génique chez les espèces récentes de primates. La distribution des arrangements correspondant à des cas indéterminés est parfaitement compatible avec l'hypothèse d'une conversion génique qui se serait produite au niveau du noeud SBOH ou FSBOH.Toutes ces observations font pencher en faveur de l'hypothèse de conversions géniques anciennes qui se seraient produites dans le groupe des primates, peu après la divergence

des primates et des rongeurs. L'analyse, site par site, indique, de plus, un sens de conversion. La polarité la plus fréquente suggère une conversion du gène Y par le gène X.

On a vu que le nombre de codoubles est un indice de la présence de conversions géniques. La façon dont ces codoubles sont répartis peut fournir des indications concernant l'emplacement des zones de conversion. Lorsqu'il n'y a pas de conversion génique, un codouble est un événement plutôt rare et son apparition nécessite au moins une convergence. Lorsqu'il y a eu conversion génique, on s'attend à une recrudescence de codoubles dans la région impliquée dans la conversion.

Lorsqu'on examine la répartition des 256 sites sélectionnés pour faire le test de comparaison entre le singe écureuil et le rat, on constate la présence de codoubles dans presque toutes les régions des séquences considérées. Seule une région centrale de 217 nucléotides ne contient aucun codouble. Ces observations suggèrent la présence de conversions géniques qui se seraient produites aux deux extrémités des séquences étudiées.

Cette subdivision en trois régions est en accord avec les résultats fournis par l'examen des zones à faible densité. En évaluant les variations des valeurs de Ka et de Ks le long des séquences de primates, Shimmin et al. ont délimité deux zones à faible densité, correspondant aux intervalles 1 à 437 et 750 à 1184 (Shimmin et al. 1994).

Même si l'examen de la répartition des codoubles délimite une région intermédiaire un peu plus restreinte (sites 498 à 715), étant donné le caractère approximatif de ce type de comparaison, on peut considérer que les résultats obtenus sont en accord avec les observations faites par Shimmin et al.

On peut constater la présence de traces de conversion génique dans la partie terminale des séquences de la Souris. On observe, en effet, chez la Souris, une agrégation de sites correspondants à l'hypothèse d'une conversion du géne Zfy par le géne Zfx (sites 1028, 1070, 1097, 1121, 1133, voir Tableau 10). Les données concernant le Renard sont incomplètes, seule une partie du domaine codant pour les doigts à zinc a été séquencée (site 438 à 749). L'analyse site par site permet, malgré tout, de déceler les indices d'une conversion génique qui se serait produite tardivement dans la lignée du Renard. Cette hypothèse a déjà été envisagée (Shimmin et al. 1994, Pamilo et Bianchi 1993) mais l'analyse, en terme de phylogénie ou à partir des valeurs de Ks, ne permettait pas de déterminer l'emplacement, même approximatif, des zones de conversion génique (Shimmin et al. 1994). L'analyse site par site permet de déceler une agrégation de sites correspondants à l'hypothèse d' une conversion de Zfx par Zfy (sites 500, 512, 552). L'analyse site par site corrobore donc l'hypothèse d'une conversion génique chez le Renard. Elle suggère, de plus, que la conversion se serait produite dans la première moitié de la séquence examinée (voir Tableau 10).

LIMITES ET FIABILITÉ DES MÉTHODES EMPLOYÉES

Plus on dispose d'information sur une famille multigénique, plus on a de chances d'identifier correctement les zones de conversion génique.

Les méthodes de compatibilité comme la méthode de Fitch sont assez fiables, car elles s'appuient sur une grande quantité d'information préalable. Il est toujours bon, cependant, de recourir, dans la mesure du possible, à des méthodes statistiques (Smith 1992, Hughes 1991). Les méthodes de simulation ont aussi l'avantage d'être économiques. Lorsqu'on dispose de peu d'information, elles constituent une alternative intéressante.

1) LIMITES DES MÉTHODES EMPLOYÉES

Les tests statistiques sont élaborés en fonction d'une certaine représentation de la façon dont les gènes évoluent. L'adéquation de cette représentation relève de la théorie de l'évolution moléculaire. Il est possible que ce modèle ne soit pas définitif et il peut encore être perfectionné. La validité des méthodes statistiques ne peut être déterminée que par rapport au modèle général dans son état actuel. Même si cette éventualité est très improbable, on ne doit pas exclure la possibilité que des incohérences dans les résultats puissent être liées à certaines inadéquations du modèle général.

Les simulations se font à partir d'un certain nombre d'hypothèses. Ces hypothèses sont parfois des simplifications du modèle théorique. La nécessité de simplifier le modèle théorique ne s'impose pas lorsqu'on utilise des méthodes de compatibilité. Les hypothèses nécessaires à l'utilisation des méthodes de compatibilité sont minimales. Comme on l'a vu, les méthodes de simulation peuvent être utilisées même si l'on dispose de peu d'information sur la famille étudiée. Cet avantage a cependant une contrepartie : la nécessité de faire des hypothèses supplémentaires. Les méthodes de simulation sont extrêmement tributaires des hypothèses qui les sous-tendent.

Le choix de ces hypothèses se fait en fonction du modèle général. Il se fait aussi en tenant compte des nécessités imposées par la nature même des techniques statistiques. Il est nécessaire, par exemple, que les échantillons considérés aient suffisamment d'ampleur.

Ainsi, dans le modèle de Sawyer, on fait l'hypothèse d'un taux de mutation constant tout le long du gène (modèle à un paramètre). On sait que les modèles théoriques font une distinction entre les transitions et les transversions. Certains modèles théoriques envisagent même 12 paramètres. L'utilisation d'un modèle simplifié permet d'obtenir de meilleurs résultats statistiques.

On a vu que les permutations sont faites en tenant compte des classes de synonymie des différents codons. Comme les séquences ont une longueur finie, pour que les statistiques aient un sens, il ne faut pas que le nombre de classes correspondant à des réarrangements soit trop élevé.

Les relations qui existent entre le modèle théorique, ses simplifications et les statistiques sont représentées ci-dessous.

Théorie de
l'évolution

Génétique des
populations

Réalité → Modèles théoriques → Modèle simplifié → Statistiques

On peut voir, d'après ce schéma, que la probabilité d'une erreur de type I ne sera pas calculée en fonction du modèle théorique, mais en fonction du modèle simplifié. La probabilité de détecter une conversion génique alors qu'aucune conversion ne s'est produite dépendra donc de l'adéquation entre le modèle choisi et la réalité.

2) FIABILITÉ DES MÉTHODES EMPLOYÉES

Pour évaluer la fiabilité des méthodes de simulation, deux approches sont

envisageables : comparer les résultats obtenus (par ces méthodes) avec les résultats fournis par les méthodes de compatibilité ou effectuer une simulation déterministe.

1 - Comparaison avec les méthodes de compatibilité

Une façon de tester la fiabilité des méthodes de permutation est le critère de cohérence. Si un résultat significatif est corroboré par les analyses de compatibilité et, d'une façon générale, si toute l'information dont on dispose confirme le résultat obtenu, on peut considérer que le test a fonctionné correctement.

Cette façon de procéder a cependant ses limites car, pour obtenir une bonne évaluation de la probabilité d'une erreur de type I (par rapport à la réalité, ce qui s'est réellement passé), il faudrait un grand nombre de cas, suffisamment documentés pour que le critère de cohérence puisse être appliqué.

Même si beaucoup de cas de conversion génique ont été répertoriés chez un nombre assez important de familles, l'information dont on dispose n'est pas encore assez complète pour qu'une évaluation de la fiabilité des méthodes de simulation puisse être faite de façon satisfaisante. C'est seulement à la longue que la fiabilité des méthodes de simulation pourra être évaluée.

2 - Simulation déterministe

On a vu que les méthodes utilisées sont appliquées à une matrice de départ qui représente les données observées. À partir d'un grand nombre de matrices de données, il serait possible d'évaluer le pourcentage d'erreurs des méthodes utilisées (on s'attendrait à ce qu'il se situe autour de 5%). Ces matrices de données peuvent être obtenues en simulant l'évolution d'un gène (le gène ancestral), en recréant des duplications, des conversions géniques, etc.

Ce type de simulation que l'on pourrait appeler "simulation déterministe" a l'avantage de ne pas être soumise à la nécessité d'obéir à des critères statistiques. Une simulation déterministe peut donc être effectuée en collant étroitement à la théorie, sans avoir recours à des hypothèses simplificatrices.

Une simulation déterministe ne permettrait pas d'évaluer la fiabilité des méthodes employées par rapport à la réalité (le passé que l'on veut reconstituer) mais elle permettrait d'évaluer le bien-fondé des simplifications apportées au modèle théorique.

CONCLUSION

Chez les gènes d'actines végétales, des phénomènes de conversion génique n'ont été identifiés que chez le Maïs. Les résultats significatifs obtenus ont été corroborés par l'examen des introns bordant les zones de conversion. En ce qui concerne les globines, les tests statistiques confirment les résultats de l'analyse de compatibilité. Dans le cas des gènes Zfx/Zfy, des résultats significatifs sont obtenus lorsqu'une comparaisom est faite entre le rat et le singe écureuil. L'analyse de compatibilité permet d'interpréter ces résultats . Elle montre que la conversion génique décelée s'est probablement produite dans la lignée des primates, peu après la divergence des primates et des rongeurs. L'analyse de compatibilité permet aussi d'identifier des conversions géniques récentes chez le Renard et chez la Souris.

L'étude de ces trois groupes de gènes montre qu'il faut utiliser les différentes méthodes de détection disponibles avec beaucoup de discernement. On ne peut pas appliquer les mêmes méthodes à toutes les familles multigéniques. La détection des phénomènes de conversion génique doit toujours être précédée par une étude des caractéristiques de la famille étudiée. Chaque famille étant un cas particulier, on doit commencer par recueillir le maximum d'information sur la phylogénie et le rôle des gènes considérés.

Cette étude a accordé une large part aux méthodes statistiques. Même si ces

méthodes sont très utiles, elles ont leurs limites. Il convient de rappeler qu'un résultat significatif n'implique pas forcément qu'il y a eu conversion génique. Dans le cadre du modèle de référence, il signifie simplement que, probablement, les arrangements observés sont non aléatoires.

Du fait de leur nature, les tests statistiques n'apportent évidemment aucune réponse définitive. Une réponse en termes de vrai et de faux serait, bien sûr, préférable à une réponse probabiliste. Dans le cas des actines végétales, l'existence d'introns très conservés indique que les résultats significatifs obtenus correspondent à des conversions géniques. Une confirmation aussi immédiate est plutôt une exception. La plupart du temps, et ce fut le cas pour deux autres groupes de gènes étudiés, l'analyse devra être poussée beaucoup plus loin, il faudra comparer les résultats obtenus à l'information déjà disponible.

L'interprétation d'un résultat significatif doit se faire en respectant le critère de cohérence. De ce point de vue, un résultat significatif doit plutôt être interprété comme un indice supplémentaire. Il doit être considéré comme un morceau d'un puzzle à reconstituer. Les méthodes de compatibilité et les méthodes de simulation ne s'opposent pas. Elles n'exploitent pas le même type d'information et en donnant deux éclairages différents du même phénomène, elles sont complémentaires. C'est ce qu'on peut constater lors de l'étude du groupe des gènes Zfx/Zfy.

On a vu que la présence de zones converties pouvait gêner considérablement la reconstruction phylogénétique. L'identification des zones de conversion est donc souhaitable lors de la détermination d'une phylogénie. Réciproquement la connaissance du maximum de données phylogénétiques peut permettre de corroborer les résultats de certaines méthodes de détection de la conversion génique.

De ce point de vue, les méthodes de reconstruction phylogénétique peuvent être considérées comme des méthodes de compatibilité. On s'attend à ce que les résultats fournis par les méthodes de reconstruction et les méthodes de détection de la conversion génique soient cohérents. Cela vient de ce que ces deux approches sont, en fait, subordonnées à un objectif plus général : la reconstitution d'un passé lointain, d'une histoire, celle d'une famille de gènes. Les différents aspects de cette histoire :

- l'arbre des espèces
- l'arbre des gènes
- les conversions géniques

doivent s'intégrer pour former un ensemble cohérent. On cherche à reconstituer le scénario que constitue une phylogénie et ses incidents de parcours : d'éventuelles conversions géniques. Plus exactement, comme on n'a aucune certitude, on cherche à reconstituer le scénario le plus probable.

RÉFÉRENCES

Aguilera A. 1988, Mitotic gene conversion of large DNA heterologies in Saccharomyces cerevisiae, Mol. Gen. Genet Vol.211 pp 455-458

Andersson L., Gustafsson K., Jonsson A-K., Rask L. 1991, Concerted evolution in a segment of the first domain exon of polymorphic MHC class IIβ loci, Immunogenetics Vol.33 pp 235-242

Atchison M. and Adesnik M. 1986, Gene conversion in a cytochrome P-450 gene family, Proc. Natl. Acad. Sci, USA Vol.83 pp 2300-2304

Balding D. J., Nichols R. A. and Hunt D. M. 1992, Detecting gene conversion : primate visual pigment genes, Proc. R. Soc. Lond. B Vol.249 pp 275-280

Becker R. S. and Knight K. L. 1990, Somatic diversification of immunoglobulin heavy chain VDJ genes : evidence for somatic gene conversion in rabbits, Cell. Vol.63 pp 987-997

Bianchi N. O., de la Chapelle A., Vidal-Rioja L. and Merani S. 1989, The sex determining zinc finger sequences in XY females of Akodon azarea (Rodentia Cricetidae), Cytogenet Cell. Genet. Vol.52 pp 162-166

Borts R. M. and Haber J. E. 1989, Length and distribution of meiotic gene conversion tracts and crossovers in Saccharomyces cerevisiae, Genetics Vol.123 pp 69-80

Bowman C. M., Barker R. F. and Dyer T. A. 1988, In wheat ctDNA, segments of ribosomal protein genes are dispersed repeats, probably conserved by non reciprocal recombination, Current Genetics Vol.14 pp127-136

Chang B. H-J., Shimmin L. C., Shyue S-K., Hewett-Emmett D. and Il W-H. 1994, Weak male-driven molecular evolution in rodents, Proc. Natl. Acad. Sci. USA Vol.91 pp 827-831

Chang B. H-J. and Li W-H. 1995, Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked ube 1 and pseudogenes, J. Mol. Evol. Vol.40 pp 70-77

Derr L. K. and Strathern J. N. 1993, A role for reverse transcripts in gene conversion, Nature Vol. 361 pp 170-173

Dressler D., Potter H. 1985, Molecular mechanisms of genetic recombination, in Wilson J. H., Genetic recombination, Benjamin/Cummings Publishing Company, Menlo Park

Drouin G. and Dover, G. A. 1990, Independent gene evolution in the potato actin gene family demonstrated by phylogenetic procedures for resolving gene conversions and the phylogeny of angiosperm actin genes, Journal of Molecular Evolution Vol.31 pp 132-150

Enea V. and Corredor V. 1991, The evolution of plasmodial stage-specific rRNA genes is dominated by gene conversion, J. Mol. Evol. Vol.32 pp 183-186

Engebrecht J., Hirsch J. and Roeder G. S. 1990, Meiotic gene conversion and crossing over : their relationship to each other to chromosome synapsis and segregation, Cell Vol.62 pp 927-937

Felsenstein J. 1993, Phylogeny Inference Package, Version 3.5 (PHYLIP), University of Washington

Fitch, D. H. A., Goodman, M. 1991, Phylogenetic scanning: a computer assisted algorithm for mapping gene conversions and other recombinational events, CABIOS Vol 7 pp 207-215

Fitch, D. H. A., Mainone C., Goodman, M. and Slightom J. L. 1990, Molecular history of gene conversions in the primate fetal γ-globin genes, The Journal of Biological Chemistry Vol.265 pp 781-793

Gant B. S. and Clegg M. T. 1993, Molecular evolution of the Adh1 locus in the genus Zea, Proc. Natl. Acad. Sci. USA Vol.90 pp 5095-5099

Godowski P. J., Picard D. and Yamamoto K. R. 1988, Signal transduction and transcriptional regulation by glucocorticoid receptor Lex A fusion proteins, Science Vol.241 pp 812-816

Gottesman M. E., Vogel H. J. 1992, Mechanisms of eukaryotic DNA recombination, Academic Press, San Diego

Hawkins J. D. 1991, Gene structure and expression, 2nd edition, Cambridge University Press, Cambridge

Hayasaka K., Fitch D. H. A., Slightom J. L. and Goodman M. 1992, Fetal recruitment of anthropoid γ -globin genes, J. Mol. Biol. Vol.224 pp 875-881

Hein J. 1993, A heuristic method to reconstruct the history of sequences subject to recombination, J. Mol. Evol. Vol.36 pp 396-405

Herbomel P. 1993, L'expression du génome, Éditions scientifiques, techniques et médicales (ESTEM), Paris

Hollenberg S. M. and Evans R. M. 1988, Multiple and cooperative trans-activation domains of the human glucocorticoid receptor, Cell Vol.55 pp 899-906

Holliday R. 1974, Molecular aspects of genetic exchange and gene conversion, Genetics Vol. 78 pp 273-285

Hope I. A. and Struhl K. 1986, Functional dissection of a eukaryotic transcriptional activator protein GCN4 of yeast, Cell Vol.46 pp 885-894

Hugues A. L. 1995, Origin and evolution of HLA class I pseudogenes, Mol. Biol. Evol. Vol.12 pp 247-258

Hugues A. L. 1991, Testing for interlocus genetic exchange in the MHC : A reply to Andersson and co-workers, Immunogenetics Vol.33 pp 243-246

Jinks-Robertson S., Petes T. D. 1985, High frequency meiotic gene conversion between repeated genes in nonhomologous chromosomes in yeast, Proc. Natl. Acad. Sci. USA Vol.82 pp 3350-3354, Genetics

Jinks-Robertson S. and Petes T. D. 1993, Experimental determination of rates of concerted evolution, Methods in Enzymology Vol.224 pp 631-646

John B., Miklos G. L. G. 1988, The eukaryotic genome in development and evolution, Allen and Unwin, London

Judd R. S. and Petes T. D. 1988, Physical length of meiotic and mitotic gene conversion tracts in Saccharomyces cerevisiae, Genetics Vol.118 pp 401-410

Kimura M. 1980, A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotides sequences, J. Mol. Evol. Vol.16 pp 111-120

Koop B. F., Siemieniak D., Slightom J. L., Goodman M., Dunbar J., Wright P. C. and Simons E. L. 1989, Tarsius δ- and β-globin genes : conversions, evolution and systematic implications, The Journal of Biological Chemistry Vol.264 pp 68-79

Kourilsky, P. 1986, Molecular mechanisms for gene conversion in higher cells, Trends in Genetics Vol.2 pp 60-3

Lanfear J., Holland P.W.H. 1991, The molecular evolution of Zfy-related genes in birds and mammals, J. Mol. Evol. Vol. 32 pp 310-315

Leung W. Y., Lindgren V., Lau Y. F. and Yang-Feng Y. L. 1990, Regional assignments of the zinc finger Y-linked gene (ZFY) and related sequences on human and mouse chromosomes, Cytogenet. Cell. Genet. Vol.54 pp 151-153

Li W-H., Wu C-I. and Luo C-C. 1985, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes, Mol. Biol. Evol. Vol.2 pp 150-174

Li W-H. and Graur D. 1991, Fundamentals of molecular evolution, Sinauer Associates Inc., Sunderland, Mass., USA

Li W-H. 1993, Unbiased estimation of the rates of synonymous and nonsynonymous substitution, J. Mol. Evol. Vol.36 pp 96-99

Lloyd A. T. and Sharp P. M. 1992, Codons : a microcomputer program for codon usage analysis, J. Hered. Vol.83 pp 239-240

Lloyd S. L., Sargent C. A., Chalmers J., Lim E., Habeebu S. S. and Affara N. A. 1991, An X-linked zinc finger gene mapping to Xq 21.1, q21.3 closely related to ZFX and ZFY : possible origins from a common ancestral gene, Nucleic Acids Res. Vol.19 pp 4835-4841

McLean M., Gerats A. G. M., Bairds W. V. and Meagher R. B. 1990, Six actin gene subfamilies map to five chromosomes of Petunia Hybrida, Journal of Heredity Vol.81 pp 341-346

Malone R. E., Bullard S., Lundquist S., Klim S. and Tarkowski T.1991, A meiotic gene conversion gradient opposite to the direction of transcription, Nature, Vol.359 pp 154-155

Mardon G., Luoh S-W., Simpson E. M., Gill G., Brown L. G. and Page D. C.1990, Mouse Zfx protein is similar to Zfy-2. Each contains an acidic activating domain and 13 zinc fingers, Molecular and Cellular Biology Vol.10 p 6

Mardon G. and Page D.C. 1989, The sex determining region of the mouse Y chromosome encodes a protein with a highly acidic domain and 13 zinc fingers, Cell V.56 pp 765-770

Marino M., Archidiacono N., Franze A., Rosati M., Rocchi M., Ballabio A. and Grimaldi G. 1993, A novel X-linked member of the human zinc finger protein gene family : isolation, mapping and expression, Mamm. Genome Vol.4 pp 252-257

Melamed C., Nevo Y., Kupiec M. 1992, Involvement of cDNA recombination between Ty elements in Saccharomyces cervisiae, Molecular and Cellular Biology, pp 1613-1620

Menotti-Raymond M., Starmer M. W. T. and Sullivan D.T. 1991, Characterization of the structure and evolution of the Adh region of Drosophila hydei, Genetics Vol.127 pp 355-366

Mézard C., Pompon D., Nicolas A. 1992, Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity, Cell Vol.70 pp 659-670

Miller J., McLachlan A. D. and Klug A. 1985, Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes, EMBO Journal Vol.4 pp 1609-1614

Mitchell M., Simon D., Affara N., Ferguson-Smith M., Avner P. and Bishop C. 1989, Localization of murine X and automosal sequences homologous to the human Y-located testis determining region, Genetics Vol.121 pp 803-809

Miyamoto M. M., Goodman M. 1990, DNA systematics and evolution of primates, Annu. Rev. Ecol. Syst. Vol.21 pp 197-220

Moniz de Sá M. 1995, Ph.D. Thesis, University of Ottawa

Moniz de Sá M., Drouin G.,1996, Phylogeny and substitution rates of angiosperm actin genes, Mol. Biol. Evol. Vol.13 pp 1198-1212

Morton B. R., Clegg M. T. 1993, A chloroplast DNA mutational hotspot and conversion in a noncoding region near rbcL in the grass family (Poaceae), Current Genetics Vol.24 pp 357-365

Nicolas A., Rossignol J-L. 1985, Mechanism for homologous recombination, Nature Vol.314 p 62

Ogihara Y., Terachi T. and Tetsuo S. 1988, Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species, Proc. Natl. Acad. Sci. USA, Genetics Vol.85 pp 8573-8577

Ohta T. 1991, Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci, Proc. Natl. Acad. Sci. USA Vol.88 pp 6716-6720

Ohta T. and Basten C. J. 1992, Gene conversion generates hypervariability at the variable regions of Kallikreins and their inhibitors, Molecular Phylogenetics and Evolution Vol.1 pp 87-90

Page D. C., Mosher R., Simpson E. M., Fisher E. M. C., Mardon G., Pollack J., Mc Gillivray B., de la Chapelle A. and Brown L. G. 1987, The sex-determining region of the human Y chromosome encodes a finger protein, Cell Vol.51 pp 1091-1104

Palmer M. S., Berta P., Sinclair A. H., Pym B. and Goodfellow P. N.1990, Comparison of human ZFX and ZFY transcripts, Proc. Natl. Acad. Sci. USA Vol.87 pp 1681-1685

Pamilo P. and Bianchi N. O. 1993, Evolution of the Zfx and Zfy genes : rates and interdependence between the genes, Mol. Biol. Evol. Vol.10 pp 271-281

Porter C. A., Sampaio I., Schneider H., Schneider M. P. C., Czelusniak J., Goodman M. 1995, Evidence on primate phylogeny from ε-globin gene sequences and flanking regions, J. Mol. Evol. Vol.40 pp 30-55

Powers P. A. and Smithies O. 1986, Short gene conversions in the human fetal globin gene region : a by-product of chromosome pairing during meiosis?, Genetics Vol.112 pp 343-358

Reece S. K., Mc Elroy D. and Wu R. 1992, Function and evolution of actins, Evolutionary Biology Vol.26 pp 1-34

Regier J. C., Wiegmann B. M., Leclerc R. F. and Fiedlander T. P. 1994, Loss of phylogenetic information in chorion gene families of Bombyx mori by gene conversion, Mol. Biol. Evol. Vol.1 pp 72-87

Roeder G. S. and Stewart S. E. 1988, Mitotic recombination in yeast, Trends in Genetics Vol 4 p 9

Rudikoff S., Fitch W. M. and Heller M. 1992, Exon specific gene correction (conversion) during short evolutionary periods : homogenization in a two gene family encoding the β gene constant region of the T-lymphocyte antigen receptor, Mol. Biol. Evol. Vol.9 N°1 pp 14-26

Saitou N. and Nei M. 1987, The neighbor-joining method : a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution Vol.6 pp 514-525

Sanderson M. J., Doyle J. J. 1992, Reconstruction of organismal and gene phylogenies from data on multigene families : concerted evolution homoplasy and confidence, Syst. Biol. Vol.41 pp 4-17

Sawyer S. 1989, Statistical tests for detecting gene conversion, Mol. Biol. Evol. Vol.6 pp 526-538

Schneider-Gädicke A., Beer-Romero P., Brown L. G., Mardon G., Luoh S-W. and Page D. 1989, Putative transcription activator with alternative isoforms encoded by human ZFX gene, Nature Vol.342 pp 708-711

Shah D.M., Hightower R.C. and Meagher R.B. 1982, Complete nucleotide sequence of a soybean actin gene. Proc. Natl. Acad. Sci. USA, Vol.79 pp1022-1026.

Shimmin L. C., Chang B. H-J., Li W-H. 1993, Male-driven evolution of DNA sequences, Nature Vol.362 pp 745-747

Shimmin L. C., Chang B. H-J., Li W-H. 1994, Contrasting rates of nucleotide substitution in the X-linked and Y-linked zinc finger genes, J. Mol. Evol. Vol.39 pp 569-578

Sinclair A. H., Foster J. W., Spencer J. A., Page D. C., Palmer M. S., Goodfellow P. N. and Graves J. A. M. 1988, Nature (London) Vol.336 pp 780-783

Slightom J. L., Blecht A. E. and Smithies O. 1980, Human fetal Gγ and Aγ globin genes : complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes, Cell Vol.21 pp 627-638

Slightom J. L., Theisen T. W., Koop B. F. and Goodman M. 1987, Orangutan fetal globin genes, The Journal of Biological Chemistry Vol.262 pp 7472-7483

Slightom J. L., Koop B. F., Xu P. and Goodman M. 1988, Rhesus fetal globin genes, The Journal of Biological Chemistry Vol.25 pp 12427-12438

Smith J. M. 1992, Analyzing the mosaic structure of genes, J. Mol. Evol. Vol.34 pp126-129

Stachelek J. L. and Liskay R. M. 1988, Accuracy of intrachromosomal gene conversion in mouse cells, Nucleic Acids Research Vol.16 p. 4069

Stephens J. C. 1985, Statistical methods of DNA sequence analysis : detection of intragenic recombination of gene conversion, Mol. Biol. Evol. Vol.2 pp 539-556

Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L. and Jones R. T. 1988, Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus), J. Mol. Biol. Vol.203 pp 439-455

Thangavelu M., Belostotsky D., Bevan M. W., Flavell R. B., Rogers H.J. and Lonsdale D.M. 1993, Partial caracterisation of the Nicotiana tabacum actin gene family: evidence from pollen specific expression of one of the gene family members. Molecular and General Genetics Vol. 240 pp 290-295

Thomson C. B. 1992, Creation of immunoglobin diversity by intrachrosomal gene conversion, Trends in Genetics Vol.8 pp 416-422

Valleley E. M., Muller U., Ferguson M. W. and Sharpe T. 1992, Cloning and expression analysis of two ZFY-relating finger genes from Alligator mississipiensis, a species with temperature-dependent sex determination, Genes Vol.119 pp 221-228

Walsh J. B. 1986, Selection and biased gene conversion in a multigene family : consequences of interallelic bias and threshold selection, Genetics Vol.112 pp 699-716

Wheeler C. J., Maloney D., Fogel S., Goodenow R. S. 1993, Microconversion between murine H-2 genes integrated into yeast, Nature Vol.347 p 47