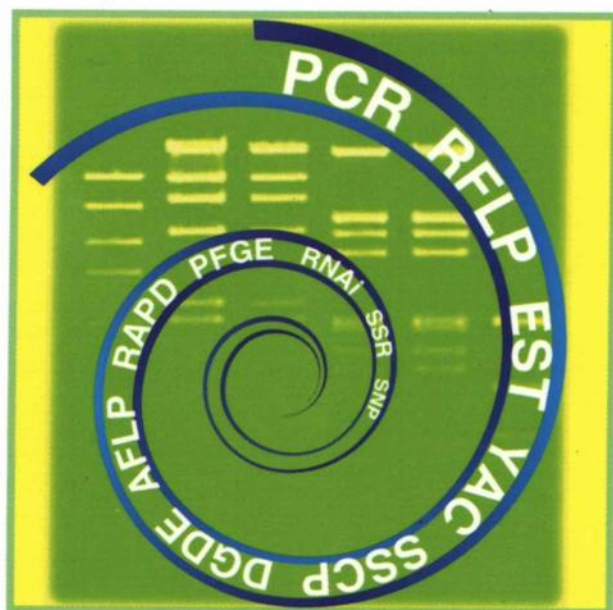


Principes des techniques de biologie moléculaire

2^e édition revue
et augmentée

D. Tagu, C. Moussard, eds



 **INRA**
EDITIONS



Principes des techniques de biologie moléculaire

2^e édition revue
et augmentée

© INRA, Paris, 2003

ISBN : 2-7380-1067-9

ISSN : 1144-7605

Le code de la propriété intellectuelle du 1^{er} juillet 1992 interdit la photocopie à usage collectif sans autorisation des ayants droit. Le non respect de cette disposition met en danger l'édition, notamment scientifique. Toute reproduction, partielle ou totale, du présent ouvrage, est interdite sans autorisation de l'éditeur et du Centre français d'exploitation du droit de copie (CFC), 20 rue des Grands Augustins, 75006 Paris.

Principes des techniques de biologie moléculaire

2^e édition revue
et augmentée

D. Tagu, C. Moussard, eds

MIEUX COMPRENDRE

Ouvrages parus dans la même collection :

Éléments de génétique quantitative

2^e édition revue et augmentée

L. OLLIVIER

2002, 184 p.

Génie génétique

Une histoire, un défi

E. HEBERLE-BORS

trad. M.L. SPIRE, R. JUDOR

2001, 304 p.

L'eau dans l'espace rural

Vie et milieux aquatiques

A. NEVEU, C. RIOU, R. BONHOMME,

P. CHASSIN, F. PAPY (éd.)

2001, 300 p.

Principes de virologie végétale

Génome, pouvoir pathogène, écologie
des virus

S. ASTIER, J. ALBOUY, Y. MAURY, H. LECOQ

2001, 488 p.

Le grain de blé

Composition et utilisation

P. FEILLET

2000, 310 p.

Biology of lactation

J. MARTINET, L.-M. HOUEBINE, H.H. HEAD

1999, 686 p.

Sol : interface fragile

Pierre STENGEL et Sandrine GELIN

1998, 222 p.

**Les marqueurs moléculaires
en génétique et biotechnologies végétales**

Dominique DE VIENNE

1998, 200 p.

Assimilation de l'azote chez les plantes

Aspects physiologique, biochimique
et moléculaire

Jean-François MOROT-GAUDRY (éd.)

1997, 422 p.

L'eau dans l'espace rural

Production végétale et qualité de l'eau

C. RIOU, R. BONHOMME, P. CHASSIN,

A. NEVEU, F. PAPY (éd.)

1997, 414 p.

La pomme de terre

P. ROUSSELLE, Y. ROBERT

et J.C. CROSNIER (éd.)

1996, 640 p.

**Vie microbienne du sol et production
végétale**

Pierre DAVET

1996, 380 p.

Nutrition des ruminants domestiques

R. JARRIGE, Y. RUCKEBUSH,

C. DEMARQUILLY, M.-H. FARCE

et M. JOURNET (éd.)

1995, 921 p.

**Amélioration des espèces végétales
cultivées**

Objectifs et critères de sélection

André GALLAIS et Hubert BANNEROT

1992, 768 p.

**La régression non linéaire :
méthodes et applications en biologie**

Sylvie HUET, Emmanuel JOLIVET

et Antoine MESSÉAN

1992, 250 p.

**L'épidémiologie en pathologie végétale :
mycoses aériennes**

Frantz RAPILLY

1991, 318 p.

**Les oligo-éléments en agriculture
et en élevage**

Yves COÏC et Marcel COPPENET

1989, 114 p.

Cette seconde édition est une version revue et augmentée de la précédente. Nous tenons à remercier les rédacteurs de ces fiches qui ont su donner de leur temps pour mettre à jour cet ouvrage.

SOMMAIRE

Définitions

1. Structure et expression d'un gène eucaryote codant un ARNm et une protéine	6
2. Paramètres de description d'un génome	8
3. Séquençage de génomes entiers	12

Vecteurs et clonage

4. Enzymes de restriction	14
5. Electrophorèse des acides nucléiques	16
6. Description d'un plasmide et d'un phagemide	18
7. Description d'un bactériophage et d'un cosmide	20
8. Description d'un YAC et autres grands vecteurs	22
9. Clonage moléculaire	24
10. Transformation génétique des bactéries et des levures	27

Marquage d'acides nucléiques et hybridations

11. Marquage de l'ADN	29
12. Hybridation moléculaire	32
13. Hybridation <i>in situ</i> des ARNm	36

Banques d'ADN et criblage

14. Construction d'une banque d'ADN génomique	38
15. Construction d'une banque d'ADNc	40
16. Criblage d'une banque	42
17. Criblage différentiel : banques d'ADNc soustraites, AFLP-ADNc	46
18. Criblage différentiel par DD RT-PCR : tri d'ARNm (<i>differential display RT-PCR</i>)	50
19. Criblage différentiel par SSH : hybridation soustractive et suppressive (<i>suppression subtractive hybridization</i>)	54
20. Criblage différentiel par RDA : analyse par différence de représentation (<i>representational difference analysis</i>)	58

21. EST : étiquettes de gènes exprimés (<i>expressed sequence tags</i>)	62
22. Réseaux d'ADN : puces à ADN, filtres d'ADNc	65

Caractérisation d'un gène

23. Séquençage d'ADN	76
24. PCR (<i>polymerase chain reaction</i>)	80
25. RACE : amplification rapide d'extrémités d'ADNc (<i>rapid amplification of cDNAs ends</i>)	82
26. Marche génomique par PCR	86
27. RT-PCR : PCR sur ARNm (<i>reverse transcriptase PCR</i>)	88
28. Transcription <i>in vitro</i>	90
29. Détermination du site d'initiation de la transcription	92
30. Analyse fonctionnelle de promoteurs	94
31. Gel-retard	96
32. Empreinte à la DNase I (<i>footprinting</i>)	98

Transformations génétiques d'eucaryotes

33. Transformation génétique des végétaux par <i>Agrobacterium tumefaciens</i>	100
34. Transfert direct de gènes dans des protoplastes végétaux	105
35. Transfert direct de gènes par biolistique	106
36. Transformation génétique de cellules animales	107
37. Le clonage des animaux	110
38. Expression transitoire	114

Analyse de la fonction d'un gène

39. Protéines recombinantes	116
40. Les baculovirus d'insectes, vecteurs d'expression de gènes étrangers	120
41. Système du double hybride	124
42. Mutagenèse dirigée	128
43. Complémentation de mutation chez la Levure	132
44. Inactivation de gènes (<i>knock-out</i>)	134
45. Étiquetage moléculaire	136
46. RNAi : inactivation de gènes par interférence d'ARN (<i>RNA interference</i>)	140

Polymorphisme d'un génome

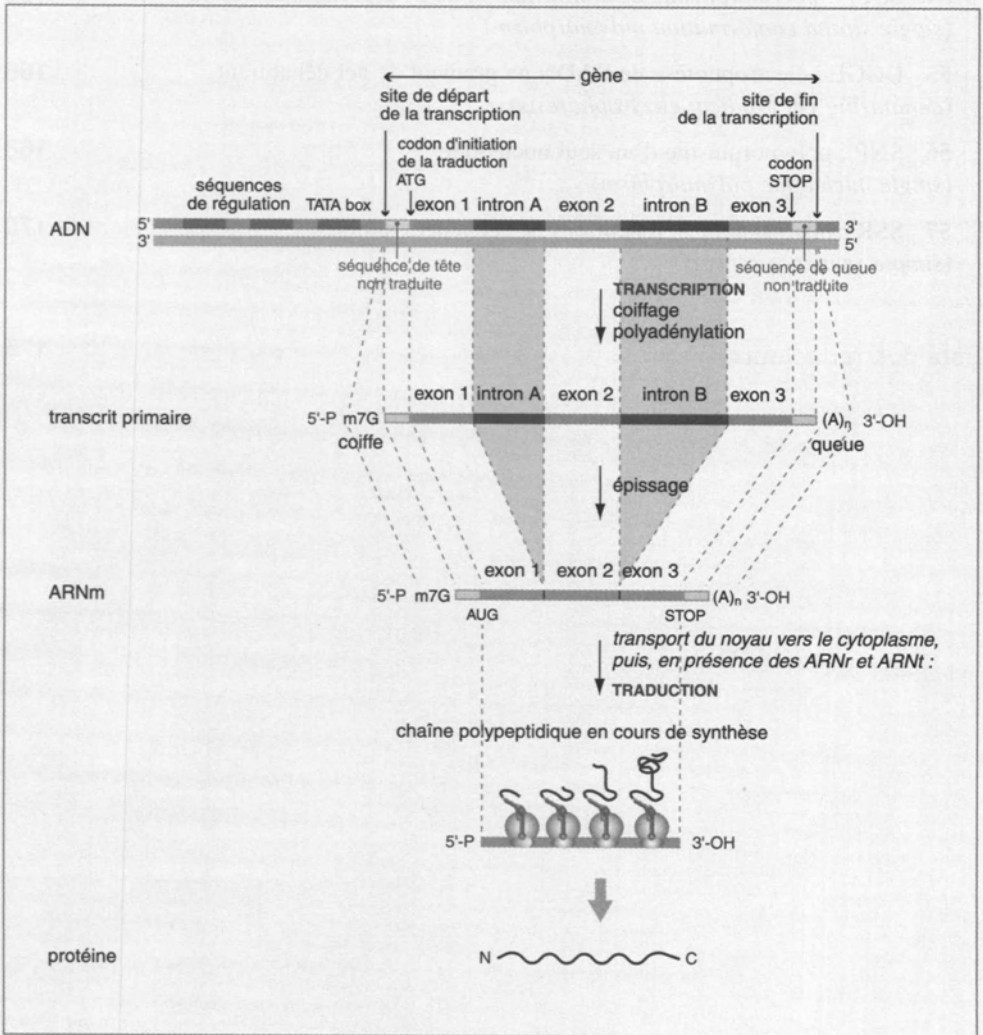
47. Marqueurs génétiques moléculaires	143
48. Cartes génétique et physique	148
49. PFGE : électrophorèse en champs pulsés (<i>pulse field electrophoresis</i>)	152
50. RFLP : polymorphisme de longueur des fragments de restriction (<i>restriction fragment length polymorphism</i>)	154
51. RAPD : polymorphisme d'ADN par amplification aléatoire (<i>random amplified polymorphic DNA</i>)	156
52. AFLP : polymorphisme de longueur de fragments amplifiés (<i>amplified fragment length polymorphism</i>)	158
53. Les rétromarqueurs	160
54. SSCP : polymorphisme de conformation de l'ADN monocaténaire (<i>single strand conformation polymorphism</i>)	164
55. DGGE : électrophorèse de l'ADN en gradient de gel dénaturant (<i>denaturing gel gradient electrophoresis</i>)	166
56. SNP : polymorphisme d'un seul nucléotide (<i>single nucleotide polymorphism</i>)	168
57. SSR : microsatellites, répétitions de séquences simples (<i>simple sequence repeats</i>)	170

Liste des rédacteurs	175
----------------------	-----

FICHE 1

STRUCTURE ET EXPRESSION D'UN GÈNE EUCARYOTE
CODANT UN ARNm ET UNE PROTÉINE

Un gène eucaryote à ARNm comporte une séquence codante bordée de séquences de régulation. Ces dernières servent de signaux de début ou de fin de transcription du gène par l'ARN polymérase II. Certaines de ces séquences (p. ex. la TATA box du promoteur) sont reconnues par des protéines appelées "facteurs de transcription généraux" car elles assistent cette enzyme dans les étapes d'initiation et d'élongation de la transcription.



D'autres séquences d'ADN¹ sont reconnues par des "facteurs de transcription spécifiques" qui modulent l'expression des gènes dans l'espace (selon le type de cellule), dans le temps (au cours du développement) et/ou sous l'effet de stress. L'agencement de ces protéines sur l'ADN déclenchera ou inhibera la transcription du gène.

La séquence codante est constituée d'exons et d'introns. Ces deux types de séquences sont transcrites (transcrit primaire), mais les introns sont éliminés lors de l'épissage. L'ARN est d'abord modifié en 5'-P (addition d'une coiffe) et en 3'-OH (addition de nucléotides à adénine), puis épissé (élimination des introns) avant d'être transporté dans le cytoplasme. Là, les ribosomes se fixent sur l'ARNm (ARN messenger) et, par l'intermédiaire des ARNt (ARN de transfert), l'ARNm est traduit en le polypeptide correspondant.

1. Elles peuvent se trouver en aval et/ou en amont de la séquence codante.

PARAMÈTRES DE DESCRIPTION D'UN GÉNOME

Paramètres¹

- Quantité d'ADN par noyau 2C : de l'ordre du picogramme ($1 \text{ pg} = 10^{-12} \text{ g}$).
- Longueur totale de l'ADN : exprimée en paire de bases (pb), en kilopb (kpb), en mégabases ($1 \text{ Mb} = 10^6 \text{ pb}$).
- Pourcentage en G + C : pourcentage de bases cytosine et guanine dans une séquence d'ADN connue.
- Diamètre de l'ADN : $20 \text{ \AA}$. Un morceau d'ADN de $1 \text{ \AA}$ de longueur a une masse de $193/6 \cdot 10^{23} \text{ g}$.
- 1 paire de bases a une masse de $660/\text{N}$, soit $660/6 \cdot 10^{23} \text{ g}$.

Grandeur de différents génomes¹

- Virus (ADN ou ARN) : de 10^3 à 10^5 pb , soit de 1 à $100 \text{ }\mu\text{m}$ déroulé.
- Bactéries : 10^6 à 10^7 pb , soit environ 1 mm .
- Levure : $1,2 \cdot 10^7 \text{ pb}$.
- Plantes supérieures :
 - noyau : de 10^8 à 10^{10} pb , soit environ 1 m . La ploïdie est généralement de 2 et l'hérédité de type mendélienne.
 - mitochondries : de $1,8$ à $6 \cdot 10^5 \text{ pb}$. La longueur est très variable. La ploïdie est non déterminée et l'hérédité est de type cytoplasmique, maternelle.
 - chloroplastes : $1,5 \cdot 10^5 \text{ pb}$, soit $50 \text{ }\mu\text{m}$. La ploïdie est de l'ordre de $10\times$ par chloroplaste et il y a de 10 à 100 chloroplastes par cellule. L'hérédité est de type cytoplasmique, maternelle.
- Homme :
 - noyau : $3 \cdot 10^9 \text{ pb}$, soit 1 m .
 - mitochondries : 16 à $20 \cdot 10^3 \text{ pb}$.

Les trois types de séquences du génome nucléaire¹

- Séquences "uniques" : une copie ou quelques copies par génome haploïde (on parle alors d'une petite famille multigénique, p. ex. les gènes codant pour les protéines de réserve des Graminées). Ces séquences comprennent des gènes de protéines et des transposons.

- Séquences moyennement répétées : de 1 000 à 100 000 copies par génome. Elles comprennent les gènes codant pour les ARN ribosomiques (ARNr), ARN de transfert (ARNt), éventuellement des gènes de protéines.
- Séquences hautement répétées : de 100 000 à 1 000 000 de copies. Ces séquences ne sont pas des gènes et peuvent être ou non transcrites. Leur rôle est peu connu (p. ex. l'ADN satellite).

Données actuelles sur les génomes

De nombreux génomes procaryotes et eucaryotes ont été séquencés ou sont en cours de séquençage (voir Fiche 3). Le tableau suivant donne quelques exemples de paramètres physiques de génomes.

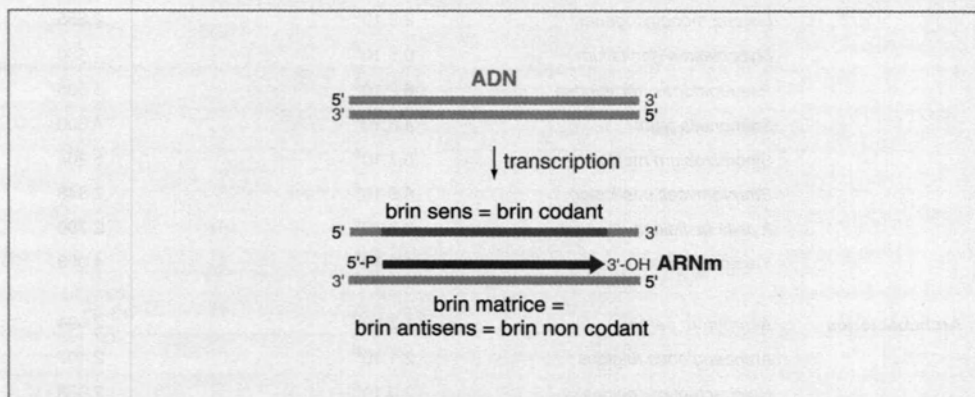
		Taille (pb)	Nombre de gènes
Eubactéries	<i>Agrobacterium tumefaciens</i>	4,8.10 ⁶	4 554
	<i>Bacillus subtilis</i>	4,2.10 ⁶	4 100
	<i>Buchnera species</i>	0,6.10 ⁶	564
	<i>Enterococcus faecalis</i>	3,2.10 ⁶	3 334
	<i>Escherichia coli</i>	4,6.10 ⁶	4 288
	<i>Haemophilus influenzae</i>	1,8.10 ⁶	1 709
	<i>Helicobacter pylori</i>	1,6.10 ⁶	1 566
	<i>Lactococcus lactis</i>	2,3.10 ⁶	2 566
	<i>Listeria monocytogenes</i>	2,9.10 ⁶	2 855
	<i>Mycoplasma genitalium</i>	0,5.10 ⁶	480
	<i>Pseudomonas aeruginosa</i>	6,2.10 ⁶	5 565
	<i>Salmonella typhi</i>	4,8.10 ⁶	4 600
	<i>Sinorhizobium meliloti</i>	6,7.10 ⁶	6 204
	<i>Streptomyces coelicolor</i>	8,6.10 ⁶	7 848
	<i>Xylella fastidiosa</i>	2,6.10 ⁶	2 766
	<i>Yersinia pestis</i>	4,6.10 ⁶	4 008
Archéobactéries	<i>Aeropyrum pernix</i>	1,7.10 ⁶	2 694
	<i>Archaeoglobus fulgidus</i>	2,1.10 ⁶	2 407
	<i>Halobacterium species</i>	2,0.10 ⁶	2 058
	<i>Methanobacterium thermoautotrophicum</i>	1,7.10 ⁶	1 869
	<i>Methanococcus jannaschii</i>	1,6.10 ⁶	1 715
	<i>Pyrobaculum aerophilum</i>	2,2.10 ⁶	2 605
	<i>Pyrococcus abyssi</i>	1,7.10 ⁶	1 765
	<i>Sulfolobus solfataricus</i>	2,9.10 ⁶	2 977
	<i>Thermoplasma acidophilum</i>	1,5.10 ⁶	1 478

		Taille (pb)	Nombre de chromosomes	Nombre de gènes (estimation)
Eucaryotes	<i>Arabidopsis thaliana</i>	125.10 ⁶	5	25 500
	<i>Caenorhabditis elegans</i>	97.10 ⁶	6	19 000
	<i>Drosophila melanogaster</i>	140.10 ⁶	4	13 600
	<i>Encephalitozoon cuniculi</i>	2.9.10 ⁶	11	1 997
	<i>Homo sapiens</i>	3 100.10 ⁶	21	35 000
	<i>Oryza sativa</i>	420.10 ⁶	12	40 000
	<i>Saccharomyces cerevisiae</i>	12.10 ⁶	16	6 100

Brin codant, brin sens : quelques définitions

L'ADN est constitué de deux brins antiparallèles et complémentaires. Lors de sa transcription en ARN, seul un des deux brins est lu et copié par l'ARN polymérase. Le brin d'ARN obtenu est donc complémentaire du brin d'ADN matrice qui a servi de copie. Par définition, ce brin d'ADN matrice est appelé antisens ; l'autre brin, qui a la même séquence que l'ARNm, est le brin sens.

L'ARNm est ensuite lu et décrypté en protéine : c'est la traduction. Par convention, la séquence d'ADN identique à l'ARN (donc le brin sens) est appelé brin codant ; l'écriture d'un gène sur le papier correspond au brin codant dans son orientation 5'-P vers l'extrémité 3'-OH.



1. D'après Francis Quétier

SÉQUENÇAGE DE GÉNOMES ENTIERS

La connaissance de la structure d'un génome dans son entier peut passer par son séquençage. Cependant, la taille des génomes étant de plusieurs millions de bases (ou mégabases, voir Fiche 2), il est nécessaire de coupler les approches de biologie moléculaire et d'informatique pour pouvoir gérer ces millions de données.

On distingue deux méthodes de séquençage des génomes entiers. Dans les deux cas, l'ADN génomique est tout d'abord fragmenté par des méthodes soit enzymatiques (enzymes de restriction), soit physiques (ultrasons) :

- la première méthode de séquençage, dite par ordonnancement hiérarchique, consiste à classer les fragments génomiques obtenus avant de les séquencer ;
- la seconde méthode, dite globale (ou *whole-genome shotgun*), néglige la phase préalable de tri et propose de séquencer dans le désordre chacun des fragments génomiques. La puissance de la bioinformatique permet dans un second temps de réordonner les fragments génomiques par chevauchement de leurs séquences communes.

La principale différence est que l'ordonnancement hiérarchique essaie d'aligner un jeu de clones de grande taille (~ 100 kb) alors que, dans la méthode globale, le génome entier est réduit en fragments de petite taille, séquencés puis alignés.

Ordonnancement hiérarchique

L'ADN génomique est, après extraction, fragmenté, généralement par des procédés de sonication, en fragments maniables (c'est-à-dire clonables dans des grands vecteurs) de 50 à 200 kb, puis cloné dans un vecteur adapté comme les chromosomes artificiels bactériens (ou BACs, voir Fiche 8). Le nombre de clones doit permettre une redondance (une représentation) de 5 à 10 fois le génome total. Le chevauchement et l'ordonnancement des clones est alors réalisé soit par hybridation de sondes spécifiques (marqueurs moléculaires), soit par analyse des profils de restriction, soit plus fréquemment par un ordonnancement après séquençage et hybridation des extrémités des BACs (500 à 600 bp). Après avoir effectué cet ordonnancement des clones (donc des fragments génomiques), les clones sélectionnés sont individuellement fragmentés, séquencés et assemblés par informatique (alignements).

Les avantages de l'ordonnancement hiérarchique sont de deux ordres : (i) une plus grande facilité d'assemblage des séquences grâce aux données des cartes physiques, génétiques et aux chevauchements des BACs ; (ii) une ouverture vers un travail en collaboration entre plusieurs laboratoires, chacun séquençant une région de chromosome. L'inconvénient majeur est la présence, à la fin, de portions d'ADN absentes de l'analyse à cause généralement des difficultés de clonage dans *E. coli* de certaines séquences répétées très fréquentes.

Méthode globale

Il s'agit d'une méthode de séquençage "au tout venant" d'ADN génomique fragmenté, mise au point par le groupe privé Celera Genomics sur des génomes bactériens, puis sur le génome de la *Drosophile* et enfin sur les génomes de l'Homme et de la Souris. Les étapes sont les suivantes : de deux à trois banques de fragments d'ADN aléatoires de tailles différentes sont réalisées. De nombreux clones sont séquencés puis assemblés. La séquence totale est réalisée à force de recouvrements et d'assemblages.

Les avantages par rapport au séquençage hiérarchique sont la rapidité et le faible coût. L'inconvénient reste le problème d'assemblage, notamment en raison de la présence importante de séquences répétées, particulièrement dans les génomes de mammifères.

Séquençage et assemblage

Pour les deux approches décrites précédemment, le séquençage et l'assemblage des fragments se font de la même manière. Le principe de séquençage repose sur la méthode de Sanger (voir Fiche 23). Le séquençage d'un génome entier nécessite des séquenceurs d'ADN particulièrement performants, ainsi que l'assistance de robots capables de réaliser des milliers de pipettages et réactions enzymatiques d'une manière automatisée.

Après la récupération des données brutes du séquençage, il faut assembler les séquences afin de reconstituer la séquence initiale. Ainsi, toutes les séquences obtenues doivent être ordonnées les unes par rapport aux autres par la mise en évidence de recouvrements (extrémités chevauchantes), c'est-à-dire des régions terminales présentant un enchaînement de nucléotides identiques (aux erreurs de séquençage près). Lorsque deux séquences se chevauchent, elles sont contiguës et forment ce qu'on appelle un "contig" qui désigne deux ou plusieurs fragments chevauchants. Le but du projet est d'aboutir (par recouvrements successifs de l'ensemble des contigs) à un contig unique correspondant à la séquence totale de la molécule d'ADN analysée (un chromosome). Le taux d'erreur peut approcher 1 %, sachant que certaines portions (comme les zones répétées) sont plus délicates à séquencer. L'obtention d'une séquence qui couvre 100 % d'un génome est particulièrement coûteuse. C'est pour ces raisons que l'on qualifie de "brouillon" (*working draft*) certains génomes en voie d'achèvement de séquençage.

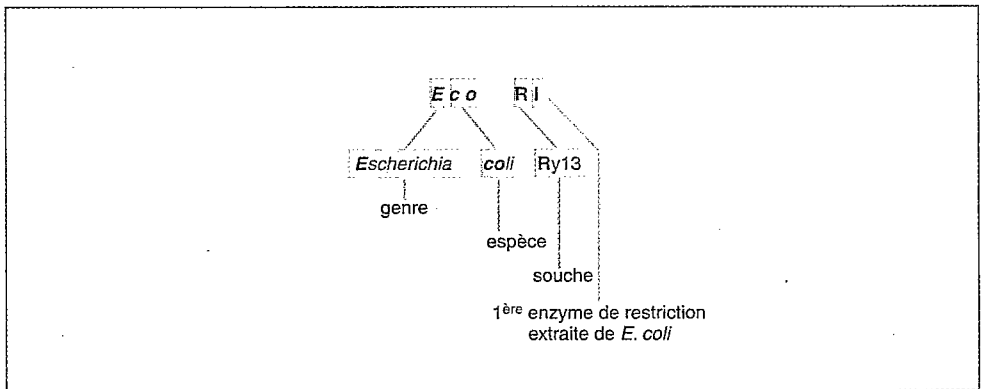
Annotation

L'annotation est l'étape finale : elle est basée sur un travail de bioinformatique à partir des séquences obtenues et alignées. Elle consiste à : (i) localiser les régions codantes (gènes) ; (ii) déterminer la direction de la transcription ; (iii) définir les protéines encodées avec des indications de structure et de fonction ; (iv) identifier les familles de gènes (paralogie¹).

1. Les gènes qui présentent un taux élevé de similarité et qui dérivent donc vraisemblablement d'un même gène ancestral sont qualifiés de paralogues.

ENZYMES DE RESTRICTION

Les enzymes de restriction sont d'origine bactérienne. Elles ont la particularité de couper les molécules d'ADN bicaténaire (ou double brin) à des sites spécifiques de la séquence : ce sont des endonucléases. Chaque enzyme de restriction reconnaît et coupe une séquence nucléotidique donnée. Le nom des enzymes de restriction provient du nom de genre et d'espèce de la bactérie dont elles ont été isolées.



Il existe trois types d'enzymes de restriction. Les types I et III sont des protéines complexes coupant l'ADN bicaténaire en dehors de leur site de reconnaissance. Les enzymes de restriction de type II sont les outils indispensables au génie génétique. Elles reconnaissent une séquence spécifique de 4, 6 ou 8 pb et coupent à l'intérieur de cette séquence, appelée site de restriction. Une particularité de ces sites est qu'ils sont généralement palindromiques, c'est-à-dire que la même séquence est lue sur les deux brins, mais en sens inverses l'un de l'autre.

De très nombreuses enzymes de restriction sont disponibles dans le commerce. Elles sont généralement fournies avec leur tampon de réaction et la température optimale d'activité est indiquée. Il est possible de digérer un même fragment d'ADN par deux enzymes de restriction différentes. Si elles fonctionnent dans des conditions analogues (concentration en sels, pH, température...), les deux digestions se déroulent simultanément. Sinon, le fragment d'ADN doit être digéré par les deux enzymes successivement, en changeant entre les réactions enzymatiques les conditions de l'expérience.

Les enzymes de restriction utilisées coupent dans la séquence selon deux modes :

- coupure franche : sont obtenues des extrémités franches par coupure au même endroit sur les deux brins. Exemple : l'enzyme *Hae* III isolée de *Haemophilus aegyptius* coupe l'ADN double brin au niveau de la séquence GG/CC.