

Les gènes chevauchants

Les gènes sont généralement des entités distinctes, éloignées les unes des autres par des milliers de paires de bases. On a récemment décrit des gènes qui se chevauchent, à la fois chez les procaryotes et chez les eucaryotes, et qui se distribuent en quatre classes structurales. I : les gènes insérés dans les introns d'autres gènes (gène *dunce* de la drosophile) ; II : les gènes superposés (*N-myc* et *N-cym*) ; III : les gènes qui ne se chevauchent qu'en 3' (*c-erbA-1* et *ear-1*) ; IV : les promoteurs bidirectionnels (*locus surfeit* chez la souris). Dans tous les cas, la nouvelle organisation structurale peut être interprétée comme un moyen de régulation fine de l'expression des gènes, indépendante de la régulation des promoteurs par les *enhancers* ou *silencers* mais reposant sur la proximité de deux gènes partenaires. Cette régulation génique fait appel à la synthèse d'ARN antisens (*EB4-PSV*) ou à l'*editing* d'un ARN (*bFGF*), permettant la dégradation des ARN, ou au contrôle de l'épissage (*c-erbA-1* et *ear-1*).

Édith Bonnelye
Vincent Laudet

En règle générale, les gènes sont distants de plusieurs milliers, voire de dizaines de milliers de paires de bases et sont donc structurellement indépendants les uns des autres. Ils sont donc perçus comme des unités génétiques distinctes et séparées dans le génome. Cependant, l'existence nouvellement mise en évidence de gènes chevauchants permet, non pas d'infirmer, mais de nuancer quelque peu cette règle générale. En effet, dans le cas des gènes chevauchants, une partie de la séquence nucléique d'un gène, codante ou non codante, est utilisée par un autre gène.

Cette organisation de gènes a été montrée d'abord chez les virus. Leur capsid ne pouvant contenir qu'une

quantité limitée d'acides nucléiques, les virus doivent coder pour un maximum d'information dans un minimum de place. Cette restriction de taille aurait donc contraint quelques virus à développer différentes stratégies d'économie comme l'épissage alternatif, le décalage de cadre de lecture ou le chevauchement de gènes. Or, récemment, ce type de structure a également été mis en évidence chez les eucaryotes qui possèdent, eux, un génome de grande taille. Prenons comme exemple le génome humain : il comprend près de 3 milliards de nucléotides où sont organisés environ 50 000 gènes dont la taille moyenne est de 30 kpb. Les eucaryotes n'ont donc apparemment pas besoin de compacter leur information géné-

ADRESSE

V. Laudet : *animateur de groupe*. E. Bonnelye : *étudiante*. Cnrs URA 1160, oncologie moléculaire, Institut Pasteur de Lille, 1, rue Calmette 59019 Lille Cedex, France.

RÉFÉRENCES

1. Chen CN, Malone T, Beckendorf SK, Davis RL. At least two genes reside within a large intron of the *dunce* gene of *Drosophila*. *Nature* 1987 ; 329 : 721-4.
2. Furia M, D'Avino PP, Crispi S, Artiaco D, Polito LC. Dense cluster of genes is located at the ecdysone-regulated 3C puff of *Drosophila melanogaster*. *J Mol Biol* 1993 ; 231 : 531-8.
3. Zhong Y, Wu CF. Altered synaptic plasticity in *Drosophila* memory mutants with a defective cyclic AMP cascade. *Science* 1991 ; 251 : 198-201.
4. Delgado R, Hidalgo P, Diaz F, Latorre R, Labarca P. A cyclic AMP-activated K⁺ channel in *Drosophila* larval muscle is persistently activated in *dunce*. *Proc Natl Acad Sci USA* 1991 ; 88 : 557-60.
5. Qiu Y, Davis RL. Genetic dissection of the learning/memory gene *dunce* of *Drosophila melanogaster*. *Genes Dev* 1993 ; 7 : 1447-58.
6. Liu J, Maxwell ES. Mouse U14 snRNA is encoded in an intron of the mouse cognate hsc70 heat shock gene. *Nucleic Acids Res* 1990 ; 18 : 6565-71.
7. Tycowski KT, Shu MD, Steitz JA. A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Genes Dev* 1993 ; 7 : 1176-90.
8. Fragapane P, Prislei S, Michienzi A, Caffarelli E, Bozzoni I. A novel small nucleolar RNA (U16) is encoded inside a ribosomal protein intron and originates by processing of the pre-mRNA. *EMBO J* 1993 ; 12 : 2921-8.
9. Armstrong BC, Krystal GW. Isolation and characterization of complementary DNA for *N-cym*, a gene encoded by the DNA strand opposite to *N-myc*. *Cell Growth Diff* 1992 ; 3 : 385-90.
10. Schwab M, Alitalo K, Klempnauer KH, Varmus HE, Bishop JM, Gilbert F, Brodeur G, Goldstein M, Trent J. Amplified DNA with limited homology to *MYC* cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. *Nature* 1983 ; 305 : 245-8.
11. Laudet V, Bègue A, Henry-Duthoit C, Joubel A, Martin P, Stéhelin D, Saule S. Genomic organization of the human thyroid hormone receptor α (*c-erbA-1*) gene. *Nucleic Acids Res* 1991 ; 19 : 1105-12.
12. Izumo S, Mahdavi V. Thyroid hormone receptor α isoforms generated by alternative splicing differentially activate myosin HC gene transcription. *Nature* 1988 ; 334 : 539-42.

tique. Il serait donc intéressant de savoir quel(s) événement(s) ont pu amener certains gènes à adopter ces positions de chevauchement, qui entraînent des complexités supplémentaires pour leur expression. D'un point de vue structural, différents cas de chevauchement ont été identifiés et peuvent être répartis en quatre classes : les gènes insérés dans un intron (I), les gènes superposés (II), les gènes qui se chevauchent uniquement en 3' (III) et les promoteurs bidirectionnels (IV).

Classe I : les gènes insérés dans un intron

Un gène est constitué d'exons et d'introns. Mais il arrive parfois qu'au sein même d'un intron d'un gène, se cache un autre gène.

Le gène *dunce* chez *Drosophila melanogaster* est un bon exemple permettant d'illustrer ce type d'organisation [1] (figure 1). Ce gène est localisé sur le chromosome X en 3C11-12. Cette région donne naissance à une structure boursoufflée (en *puff*) qui n'est autre qu'un renflement de la chromatine dont le volume dépend du taux de transcription. L'apparition de cette structure se produit à la suite de l'augmentation de la concentration de la 20 hydroxyecdysone, une hormone stéroïde, qui intervient dans la métamorphose [2].

Le gène *dunce* code pour une AMPc phosphodiésterase (PDE) et est exprimé dans certaines parties du cerveau impliquées dans la mémoire et dans les phénomènes d'olfaction (au niveau des *mushroom bodies*, localisés dans le cortex dorsal et postérieur de chaque hémisphère), chez les larves et les adultes. Des mutations du gène *dunce* entraînent des difficultés dans les processus d'apprentissage et de mémoire. La fonction de certains types de synapses au niveau des jonctions neuromusculaires est compromise [3] et que certains canaux K⁺ dépendants de l'AMPc dans des muscles larvaires sont en permanence activés [4]. Il intervient également dans la fertilité des femelles [5]. En effet, des mutations du gène *dunce* entraînent des phénomènes de stérilité chez celles-ci, l'activité PDE étant nécessaire dans les cellules soma-

tiques pour le dépôt des œufs et dans les cellules germinales pour le développement du zygote.

Le gène *dunce* est composé de 19 exons répartis sur 148 kpb d'ADN génomique [5]. Il code pour 10 ARN dont les tailles sont comprises entre 4,2 kb et 9,6 kb qui ont chacun leur rôle spécifique.

Outre le gène *dunce*, le locus comprend six autres gènes insérés dans deux de ses introns. Les exons « 0,9 » et « 1 » du gène *dunce* sont séparés par un intron de 40 kb dans lequel se trouvent les gènes *ng-1* (*nested gene-1*), *ng-2*, *ng-3* et *ng-4*, tandis que les exons 2 et 3 sont, eux, séparés par un intron de 9 kb où sont insérés les gènes *Sgs-4* (*salivary gland secretion protein*) et *Pig-1* (*pre-intermolt gene-1*) [5]. Ces six gènes s'étendent sur 11 kpb d'ADN, ce qui représente un nombre important de gènes pour un si petit espace. À l'intérieur de cette région, les six gènes sont orientés différemment. *Sgs-4* et *ng-1* possèdent la même orientation que le gène *dunce*. Les quatre autres se trouvent dans l'orientation inverse. Tous les six possèdent la même structure. Ils sont petits, sans intron, et présentent à leur extrémité N-terminale un peptide signal présomptif suggérant qu'ils codent pour des protéines destinées à être sécrétées. Quelques données fonctionnelles sont connues sur les rôles respectifs de ces gènes. A l'exception du gène *ng-4*, les cinq autres gènes sont fortement exprimés pendant la vie larvaire et ont une spécificité de tissus identique puisqu'ils sont tous transcrits au niveau des glandes salivaires. On sait que les glandes salivaires de *Drosophila melanogaster* sont le lieu de synthèse de polypeptides de gluE. Ces polypeptides s'accumulent dans le cytoplasme des cellules au cours du troisième stade larvaire, et sont sécrétés à la fin de la vie larvaire. Nous savons que le gène *Sgs-4* code pour une protéine de gluE ; quant aux gènes *ng-1*, *ng-2* et *ng-3*, il se pourrait qu'ils codent, eux aussi, pour des protéines de glu. *ng-4*, lui, est transcrit en un ARN de 0,4kb polyadénylé qui, contrairement aux autres, est très peu exprimé et semble jouer un rôle différent si on se réfère à son spectre d'expression. Ainsi, par exemple, il n'est pas exprimé au troisième stade larvaire [2].

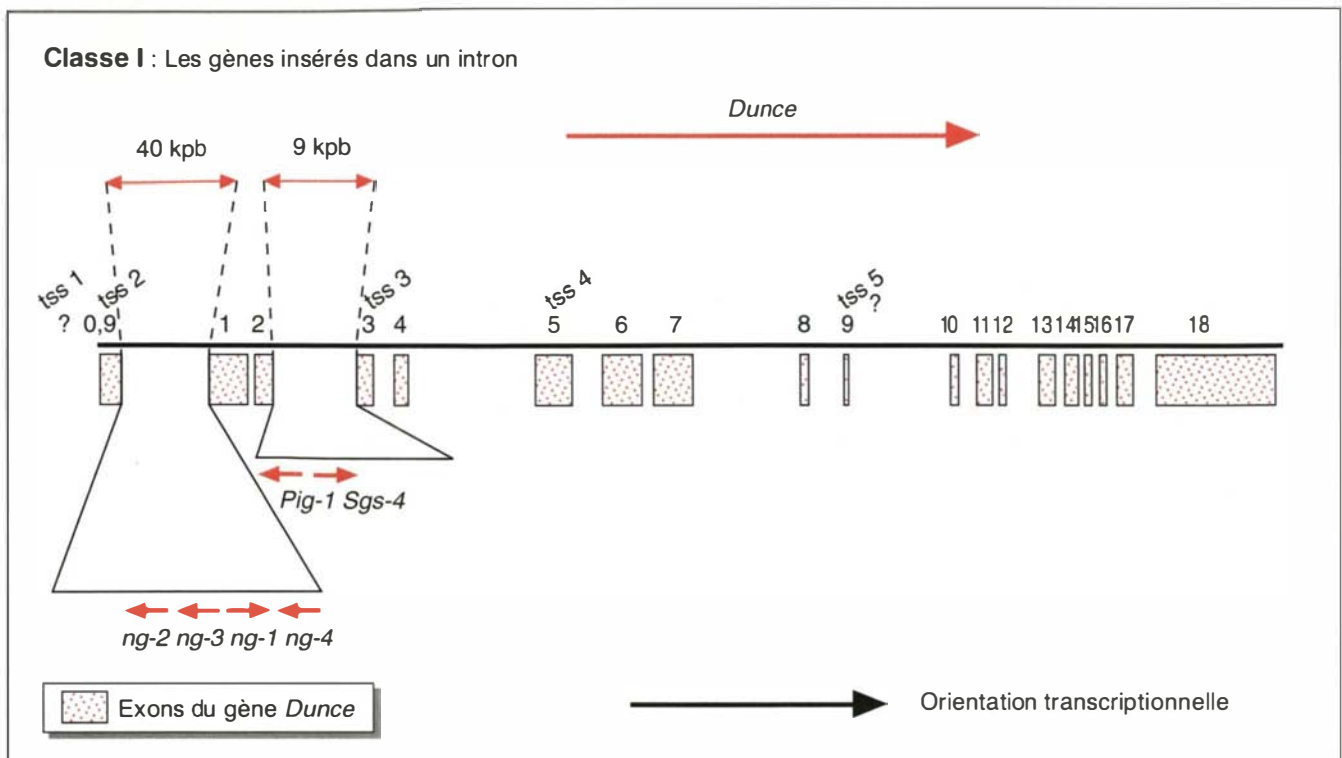


Figure 1. **Organisation du gène *Dunce***. Dix ARN sont transcrits, qui débutent aux cinq sites d'initiation de la transcription appelés tss 1 à 5 (transcription start sites). Il comprend, de plus, six gènes insérés, soit entre ses exons « 0,9 » et « 1 » (les quatre gènes *ng-1*, *ng-2*, *ng-3* et *ng-4*), soit entre ses exons 2 et 3 (les gènes *Pig-1* et *Sgs-4*).

Les gènes *Sgs-4* et *Pig-1* ne sont séparés que par 840 pb. Ils sont tous deux exprimés dans les glandes salivaires, mais pas exactement au même moment. Leur proximité et leur orientation inverse suggèrent qu'ils sont réglés par un promoteur bidirectionnel et donc qu'ils possèdent des éléments de régulation en commun. Il semblerait que ce soit aussi le cas des gènes *ng-1* et *ng-3* qui sont séparés de 300 pb. Leur expression au cours du développement ainsi que leur spécificité d'expression tissulaire sont identiques.

Bien que ces gènes possèdent une structure similaire, aucune analogie de séquence n'a été détectée entre eux, à l'exception des gènes *ng-1* et *ng-2*. Ces similitudes sont essentiellement localisées au niveau de leur extrémité 5' non codante et de leur région codante et s'élèvent respectivement à 95 % et 91,4 %, suggérant que *ng-1* et *ng-2* sont homologues, issus d'une duplication d'un même gène ancestral.

En résumé, nous avons à faire à une structure compacte où les gènes *ng-1*, 2, 3, *Pig-1* et *Sgs-4* présentent de fortes similarités structurales et ont tous une expression spécifique dans les glandes salivaires. Seul le gène *ng-4* est différent. Il est intéressant de remarquer qu'il ne semble pas y avoir de fonction commune entre ces six gènes et le gène *dunce*. En fait, la signification biologique de cette structure reste encore à élucider...

Des introns qui codent pour des petits ARN nucléolaires ou ARNSNO

Il existe des petits ARN de la famille U dont la localisation est nucléolaire (ARNSNO). Ils s'insèrent dans les introns de certains gènes et interviennent dans l'épissage des ARN ribosomiques transcrits à partir de la polymérase I. Contrairement aux gènes chevauchants de la classe I, ces ARN n'ont pas leur propre promoteur et sont produits à la suite de

la maturation du pré-messager dont ils sont issus. Il en est ainsi de *U14*, localisé dans un intron du gène *hsc 70*, qui code pour une protéine de choc thermique chez la souris [6]. Une copie de *U15*, localisée dans le premier intron d'un gène codant pour la protéine ribosomique S3, a été également récemment mise en évidence [7]. De même, *U16* est localisé dans le troisième intron d'un gène codant pour la protéine ribosomique L1 chez le xénope [8]. Au vu des données précédentes, il est intéressant de souligner, dans le cas de *U15* et *U16*, que les gènes dont ils sont issus codent pour des protéines ribosomiques. Cela implique une régulation coordonnée des protéines exprimées à partir du messager et de ces petits ARNSNO qui interviennent pour la plupart dans la genèse des ribosomes.

Classe II : les gènes superposés

Cette deuxième classe comprend des

RÉFÉRENCES

13. Koenig RJ, Lazar MA, Hodin RA, Brent GA, Larsen PR, Chin WW, Moore DD. Inhibition of thyroid hormone action by a non-hormone binding c-ErbA protein generated by alternative mRNA splicing. *Nature* 1989 ; 337 : 659-61.
 14. Lazar MA, Hodin RA, Darling DS, Chin WW. A novel member of the thyroid/steroid hormone receptor family is encoded by the opposite strand of the rat *c-erbA α* transcriptional unit. *Mol Cell Biol* 1989 ; 9 : 1128-36.
 15. Miyajima N, Horiuchi R, Shibuya Y, Fukushige SI, Matsubara KI, Toyoshima K, Yamamoto T. Two *erbA* homologs encoding proteins with different T3 binding capacities are transcribed from opposite DNA strands of the same genetic locus. *Cell* 1989 ; 57 : 31-9.
 16. Segraves WA, Hogness DS. The E75 ecdysone-inducible gene responsible for the 75B early puff in *Drosophila* encodes two new members of the steroid receptor superfamily. *Genes Dev* 1990 ; 4 : 204-19.
 17. Lazar MA, Hodin RA, Cardona G, Chin WW. Gene expression from the *c-erbA α /Rev-erbA α* genomic locus. *J Biol Chem* 1990 ; 265 : 12859-63.
 18. Huxley C, Fried M. The mouse surfeit locus contains a cluster of six genes associated with four CpG-rich islands in 32 kilobases of genomic DNA. *Mol Cell Biol* 1990 ; 10 : 605-14.
 19. Williams T, Fried M. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. *Nature* 1986 ; 322 : 275-9.
 20. Lennard AC, Fried M. The bidirectional promoter of the divergently transcribed mouse *Surf-1* and *Surf-2* genes. *Mol Cell Biol* 1991 ; 11 : 1281-94.
 21. West RWJ, Yocum RR, Ptashne M. *Saccharomyces cerevisiae Gal1-Gal10* divergent promoter region : location and function of the upstream activating sequence UASG. *Mol Cell Biol* 1984 ; 4 : 2467-78.
 22. Burbelo PD, Martin GR, Yamada Y. $\alpha 1$ (IV) and $\alpha 2$ (IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer. *Proc Natl Acad Sci USA* 1988 ; 85 : 9679-82.
 23. Collart D, Romain PL, Huebner K, Pockwinse S, Pilapil S, Cannizzaro LA, Lian JB, Croce CM, Stein JL, Stein GS. A human histone H2B.1 variant gene, located on chromosome 1, utilizes alternative 3' end processing. *J Cell Biochem* 1992 ; 50 : 374-85.
 24. Kimelman D, Kirschner MW. An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in *Xenopus* oocytes. *Cell* 1989 ; 59 : 687-96.
- gènes qui se chevauchent sur presque toute leur longueur. Leur organisation respective en exons et en introns s'imbrique totalement. On connaît peu de gènes possédant cette structure. L'exemple des gènes *N-myc* et *N-cym* est une bonne illustration de ce cas de figure [9] (figure 2). Le gène *N-myc* code pour un facteur de transcription appartenant à la famille Myc, elle-même regroupée dans une superfamille de facteurs de transcription qui présentent des motifs de dimérisation et de reconnaissance de l'ADN semblables, appelés bHLH (*basic helix loop helix*). Le gène *N-myc* intervient dans la pathogénicité d'un certain nombre de tumeurs humaines. Ainsi, on le retrouve amplifié et fortement exprimé dans les neuroblastomes chez l'enfant et dans le cancer du poumon à petites cellules [10].
- Il a été démontré qu'un autre gène nommé *N-cym*, codant pour une protéine nucléaire de 109 acides aminés, était transcrit sur le brin opposé du gène *N-myc*. Ces deux gènes sont constitués de trois exons chacun. L'imbrication des gènes est telle que l'exon I du gène *N-cym* est localisé au niveau de l'intron I de *N-myc*. L'exon II chevauche l'exon IB de *N-myc* et l'exon III est localisé en 5' du gène *N-myc* (figure 2).
- L'étude de la structure secondaire de *N-Cym* a permis de localiser, en N-terminal, deux hélices α séparées par une boucle, qui font penser à une structure HLH et une région fortement basique adjacente à la seconde hélice. Cette région basique s'est révélée riche en arginine et en lysine (30 % des 43 acides aminés en C-terminal sont des arginines et des lysines). Cette structure est bien connue comme signature de certains facteurs de transcription mais pour l'instant aucune donnée fonctionnelle ne permet d'affirmer que *N-cym* code effectivement pour un facteur de transcription.
- N-cym*, comme *N-myc*, est fortement exprimé dans les neuroblastomes et dans les cancers du poumon à petites cellules. Il semble donc que l'expression de ces deux gènes soit corrélée dans les lignées cellulaires tumorales en conditions de croissance. On retrouve cette corrélation dans des lignées de néphro-
- blastome différenciées par l'acide rétinoïque. L'acide rétinoïque diminue l'expression du gène *N-myc*. Après quatre jours de traitement par l'acide rétinoïque, on observe une décroissance parallèle de l'expression de *N-myc* et *N-cym*. Les extrémités 5' des deux unités transcriptionnelles se chevauchant, on peut imaginer qu'elles partagent certains éléments régulateurs importants pour leur réponse à l'acide rétinoïque. Cette expression corrélée suggère également que ces protéines puissent avoir des relations fonctionnelles pendant le développement normal et l'oncogénèse, le chevauchement des gènes permettant l'action simultanée de facteurs distincts. Dans l'exemple ci-dessus, la corégulation est en fait observée dans des circonstances biologiques particulières. Il est donc intéressant de savoir ce qu'il en est lors du développement fœtal normal au cours duquel *N-myc* est exprimé. *N-cym* est exprimé dans le cerveau fœtal, les poumons, le foie et les reins alors que *N-myc* n'est présent que dans le cerveau fœtal et à une faible concentration. Ainsi, *N-cym* n'est pas uniquement présent dans les cellules tumorales, et doit donc avoir un rôle fonctionnel au cours du développement fœtal normal. L'étude de l'expression des gènes au cours du développement ne permet pas de tirer de conclusion quant à la corrélation de leur expression. Elle permet seulement de suggérer que la situation est plus compliquée qu'on ne l'avait imaginée...

Classe III : les gènes qui se chevauchent uniquement en 3'

Les gènes de cette classe se chevauchent au niveau de leur extrémité 3'. C'est le cas du locus *c-erbA-1/ear-1*. Celui-ci fait intervenir le gène *c-erbA-1* qui code pour un récepteur des hormones thyroïdiennes et le gène *ear-1* (*rev-erbx*) qui code pour un récepteur orphelin (son ligand, s'il existe, n'a pas encore été identifié). Les gènes *c-erbA-1* et *ear-1* appartiennent à la même famille de gènes, celle des récepteurs nucléaires. On se trouve donc en présence d'un

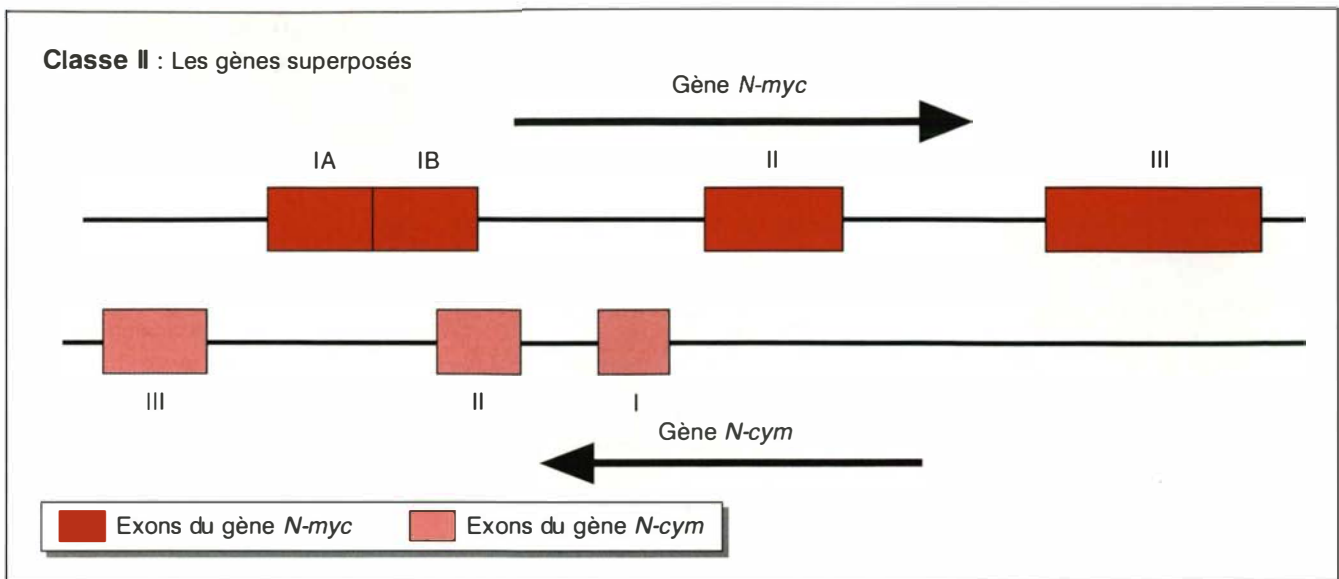


Figure 2. Description du locus chevauchant N-myc/N-cym. L'exon I du gène N-cym est localisé au niveau de l'intron I de N-myc. L'exon II chevauche l'exon IB du gène N-myc et l'exon III est localisé en 5' du gène N-myc.

chevauchement où les deux protagonistes sont structurellement proches. Chez les mammifères, le gène *c-erbA-1* code pour deux isoformes majeures à partir de deux ARNm de taille différente [11] (figure 3) : un ARN de 5 kb, correspondant aux exons 1 à 9 qui code pour un produit appelé TR α 1 et qui est le récepteur normal des hormones thyroïdiennes, et un ARN de 2,6 kb, qui code pour un produit appelé TR α 2. Cet ARN correspond aux exons 1 à 8, aux 127 premiers nucléotides de l'exon 9 (où un site d'épissage alternatif est reconnu) et enfin à l'exon 10. La forme TR α 2 n'est pas capable de fixer la T3 et se comporte comme un inhibiteur de la forme TR α 1 car les deux isoformes ont les mêmes sites de fixation à l'ADN [12, 13]. TR α 1 et TR α 2 sont en compétition pour la fixation sur les TRE (*thyroid responsive element*) localisés dans le promoteur des gènes cibles.

À cet épissage alternatif, qui permet de produire à la fois un récepteur et son inhibiteur, vient s'ajouter le chevauchement en 3' de *c-erbA-1* avec le gène *ear-1* [11, 14, 15]. Ce chevauchement implique 1 kpb, en comptant les régions non codantes, au niveau de l'exon 10 de *c-erbA-1* et de l'exon F (pour *Final*) de *ear-1*. Donc, seul l'exon

10 de *c-erbA-1*, spécifique de l'isoforme TR α 2, est impliqué dans ce chevauchement.

La localisation de l'expression de *ear-1* peut nous permettre d'avancer une hypothèse quant à sa fonction. En effet, *ear-1* est très fortement exprimé dans les tissus musculaires striés ainsi que dans le tissu adipeux brun [14]. Ces deux tissus sont les sites majeurs de la thermogénèse chez les mammifères. On peut donc émettre l'hypothèse que *ear-1* joue un rôle dans le processus de thermogénèse, qui est réglé par les hormones thyroïdiennes. Cela pourrait permettre de comprendre l'intérêt d'une liaison structurale avec le gène *c-erbA-1*.

Il existe chez la drosophile un homologue de *ear-1*, le gène *E75*, qui est induit par l'ecdysone [16] et qui semble intervenir dans la métamorphose. Le fait que le gène *E75* soit l'homologue de *ear-1* chez la drosophile suggère que celui-ci est un gène très ancien, dont la fonction est importante.

Il semble que cette structure chevauchante, ainsi que l'épissage alternatif de *c-erbA-1*, soient conservés chez les deux mammifères étudiés jusqu'à présent, le rat et l'homme [14, 17], suggérant que cette structure géno-

mique inhabituelle a une fonction biologique importante, peut-être en relation avec la régulation de l'expression des deux gènes (*voir plus loin*).

Classe IV : les promoteurs bidirectionnels

Cette classe de gènes représente plus de la moitié des gènes chevauchants décrits dans la littérature. Les parties transcrites des gènes, dans ce cas, n'interviennent pas dans le chevauchement. En effet, seuls des éléments régulateurs de la transcription sont communs aux deux gènes qui sont transcrits en orientation inverse. Le locus *Surfeit*, qui illustrera cette dernière catégorie, est relativement complexe car il ne comprend pas uniquement un promoteur bidirectionnel, mais aussi un chevauchement de gènes en 3' entre les gènes *Surf-2* et *Surf-4* [18]. Ce locus regroupe six gènes (*Surf-1* à *Surf-6*) chez la souris. Les extrémités 3' des gènes *Surf-1* et *Surf-3* (qui code pour la protéine ribosomique RpL7a) sont séparées de 70pb alors que celles des gènes *Surf-2* et *Surf-4* (qui semble coder pour une protéine de membrane) se chevauchent sur 133pb [19]. Plus récem-

RÉFÉRENCES

25. Hildebrandt M, Nellen W. Differential antisense transcription from the dictyostelium EB4 gene locus : implications on antisense-mediated regulation of mRNA stability. *Cell* 1992 ; 69 : 197-204.
26. Hélène C, Saison-Behmoaras E. La stratégie antisens : nouvelles approches thérapeutiques. *médecine/sciences* 1994 ; 10 : 253-73.
27. Maniak M, Nellen W. Evidence for a feedback regulated back-up promoter which controls permanent expression of a dictyostelium gene. *Nucleic Acids Res* 1990 ; 18 : 5375-80.
28. Knecht DA, Loomis WF. Antisense RNA inactivation of myosin heavy chain gene expression in *Dictyostelium Discoideum*. *Science* 1987 ; 236 : 1081-5.
29. Portman DS, Dreyfuss G. RNA annealing activities in HeLa nuclei. *EMBO J* 1994 ; 13 : 213-21.
30. Celano P, Berchtold CM, Kizer DL, Weeraratna A, Nelkin BD, Baylin SB, Casero RA Jr. Characterization of an endogenous RNA transcript with homology to the antisense strand of the human *c-MYC* gene. *J Biol Chem* 1992 ; 267 : 15092-6.
31. Forrest D, Sjöberg M, Vennström B. Contrasting developmental and tissue-specific expression of α and β thyroid hormone receptor genes. *EMBO J* 1990 ; 9 : 1519-28.
32. Williams T, Yon J, Huxley C, Fried M. The mouse surfeit locus contains a very tight cluster of four « housekeeping » genes that is conserved through evolution. *Proc Natl Acad Sci USA* 1988 ; 85 : 3527-30.
33. Colombo P, Yon J, Garson K, Fried M. Conservation of the organization of five tightly clustered genes over 600 million years of divergent evolution. *Proc Natl Acad Sci USA* 1992 ; 89 : 6358-62.
34. Henikoff S, Keene MA, Fichtel K, Friston JW. Gene within a gene : nested drosophila genes encode unrelated proteins on opposite DNA strands. *Cell* 1986 ; 44 : 33-42.
35. Clark DV, Henikoff S. Unusual organization features of the drosophila gart locus are not conserved within diptera. *J Mol Evol* 1992 ; 35 : 51-9.
- ment, deux autres gènes, *Surf-5* et *Surf-6*, localisés en 5' de *Surf-3* ont été identifiés. Ils sont tous les deux orientés dans la même direction, opposée à celle de *Surf-3* [18]. Une des caractéristiques importantes de ce locus est que les extrémités 5' de ces différents gènes sont associées à des îlots riches en GC.
- Les extrémités 5' des gènes *Surf-1* et *Surf-2* ne sont séparées que par 73 pb très riches en GC qui contiennent les sites d'initiation de la transcription des deux gènes : trois pour *Surf-2* (majeur, intermédiaire et mineur) et deux pour *Surf-1*. Les rôles fonctionnels de ces sites multiples d'initiation restent jusqu'à présent inconnus. Une région promotrice a été mise en évidence entre ces deux gènes. Cinq sites intitulés Su1, Su2, Su3, Su'x' et Su'y' ont été identifiés dans cette région promotrice et sont capables de fixer des facteurs de transcription [20]. Ils interviennent pour le choix du site d'initiation ainsi que dans la précision et l'efficacité de l'expression des deux gènes. Des délétions dans ces sites ont été réalisées et ont permis de montrer que certains d'entre eux interviennent de façon concomitante dans l'expression des deux gènes. Ainsi, une délétion du site Su2 inhibe la transcription de *Surf-1* à partir de son site d'initiation majeur ainsi que la transcription du gène *Surf-2*. Cela montre que le site Su2 est capable de fixer un facteur de transcription intervenant dans la transcription des gènes *Surf-1* et *Surf-2*, d'où l'appellation de promoteur bidirectionnel. Les différents effets des délétions des cinq sites sont reportés dans le *Tableau I*.
- L'étude des facteurs susceptibles de se fixer sur ces sites [20] a permis de montrer que les séquences du site Su2 sont homologues aux séquences reconnues par la protéine PEA3 qui appartient à la famille des facteurs de transcription Ets.
- L'usage alternatif des multiples sites d'initiation peut être perçu comme une possibilité d'alternance dans le contrôle par des éléments régulateurs. Ainsi, selon l'élément régulateur, l'initiation respective des deux gènes varie en fonction des sites d'initiation activés. Cette alternance dans le choix des sites permettrait de contribuer à la maintenance de l'expression des deux gènes, par exemple dans différents types cellulaires ou encore dans les différents états de la cellule. Les sites Su2, Su3 et Su'y' sont des éléments régulateurs qui interviennent dans la régulation des deux gènes, ce qui ne veut pas dire que lorsque le facteur se fixe sur son site, il active toujours simultanément la transcription des deux gènes. Plusieurs cas de figure sont envisageables. En effet, soit ce facteur active la transcription des deux gènes, soit il active uniquement celle de l'un des deux gènes, soit encore il inhibe totalement leur expression. Plusieurs niveaux de régulation sont possibles si l'on considère tous les sites de fixation de facteur de transcription présents dans ce promoteur. Il est envisageable que la transcription des deux gènes varie en fonction d'une combinaison des facteurs de transcription, permettant une régulation très fine de leur expression.
- Ce partage d'éléments régulateurs entre *Surf-1* et *Surf-2* suggère l'existence d'une relation fonctionnelle entre ces deux gènes. Ce ne serait pas un cas unique. En effet, de nombreux exemples sont connus, comme celui des gènes *Gall* et *Gall10* chez *Saccharomyces cerevisiae* [21], ou encore celui des gènes codant pour les formes $\alpha 1(IV)$ et $\alpha 2(IV)$ du collagène [22], et pour des protéines d'histones (*H2B1* et *H2A*) [23]. La transcription bidirectionnelle apparaît donc comme un mécanisme permettant de coordonner de façon positive, ou négative, l'expression de deux gènes.

Des régulations nouvelles par ARN antisens, editing et contrôle de l'épissage

La découverte de l'existence de gènes chevauchants a permis de mettre en évidence de nouveaux types de régulation de l'expression des gènes, fondés sur la proximité structurale des unités transcriptionnelles. Ainsi, outre la régulation transcriptionnelle « classique », qui se passe essentiellement au niveau des promoteurs, il existe des régulations, différentes selon les cas, fondées, par exemple, sur l'hybridation entre transcrits complémentaires ou sur

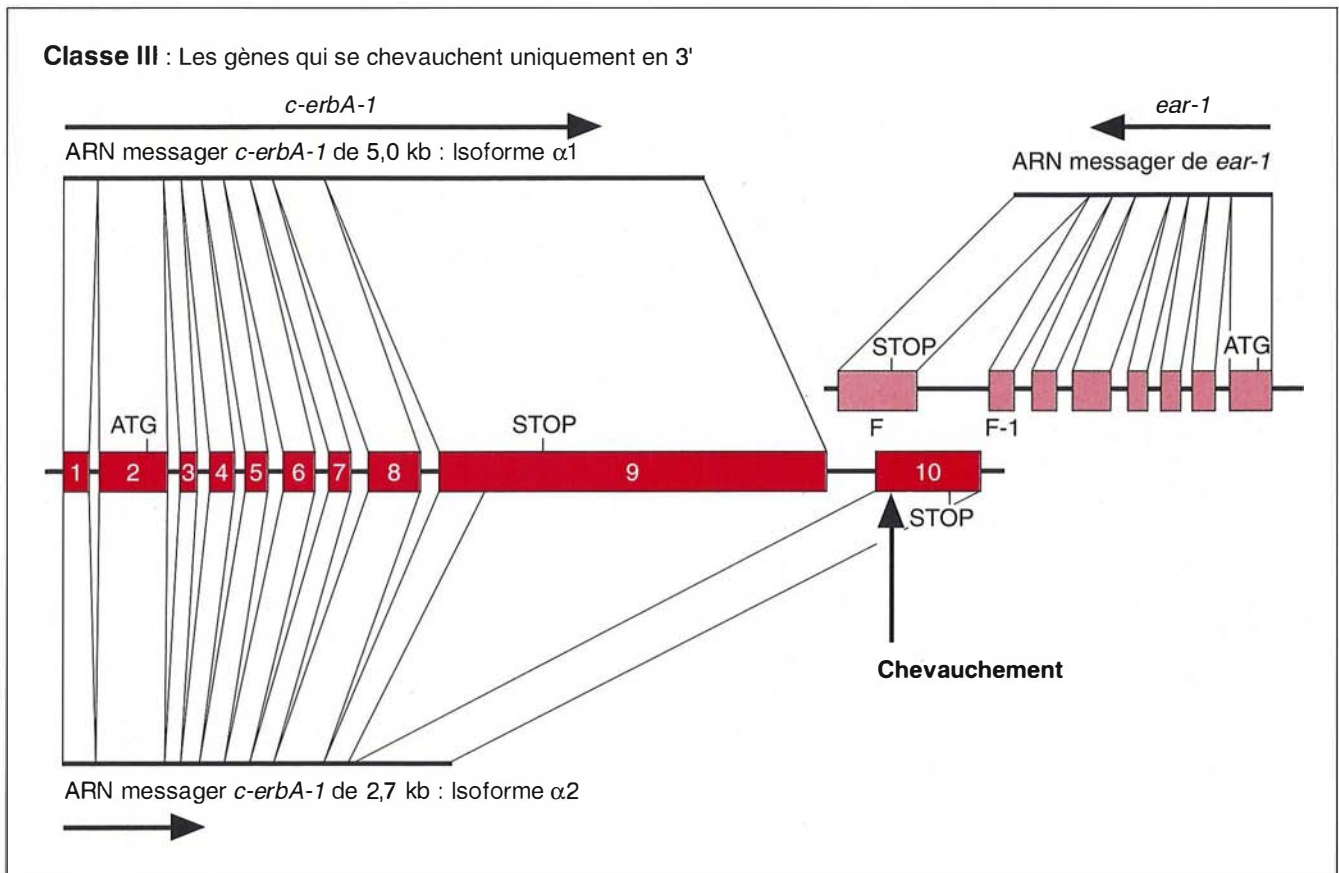


Figure 3. **Organisation génomique du locus chevauchant *c-erbA-1/ear-1*.** Le gène *c-erbA-1* code pour deux isoformes $TR\alpha 1$ et $TR\alpha 2$, grâce à un épissage alternatif. Le gène *ear-1*, placé en orientation transcriptionnelle inverse, chevauche le gène *c-erbA-1* à son extrémité 3'.

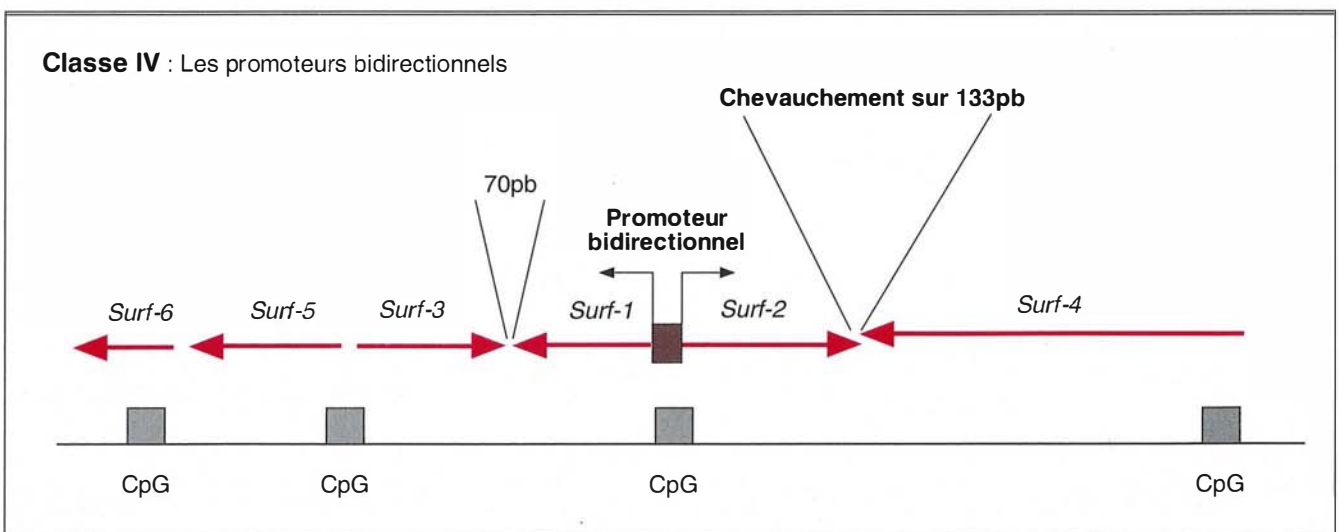


Figure 4. **Organisation du locus Surf.** Le promoteur bidirectionnel est localisé entre les gènes Surf-1 et Surf-2.

l'interférence entre les gènes pendant la transcription [24, 25].

La régulation de gènes par ARN antisens

On sait que, dans certains systèmes, l'introduction artificielle d'un ARN antisens peut inhiber l'expression du gène en provoquant la dégradation de son ARNm endogène [26] (figure 5). Ainsi, Maniack et Nellen démontrent une déstabilisation de l'ARN endogène proportionnelle au taux d'expression de l'ARN antisens correspondant [27, 28]. Ces auteurs ont établi un système artificiel de régulation en intégrant un ARN antisens de la myosine, afin d'étudier le rôle de cette protéine dans des cellules non musculaires. L'hybridation de l'antisens avec l'ARNm endogène induit effectivement la dégradation de ce dernier. Cela a permis de montrer que l'élimination de la myosine dans ces cellules non musculaires n'était pas létale, mais que cette protéine jouait un rôle important au cours du développement.

Ce type de régulation post-transcriptionnelle est également possible *in vivo*. Il a été mis en évidence pour le gène *EB4-PSV* qui code pour une protéine associée à la membrane des préspores de *Dictyostelium discoideum* [25]. Au cours du développement précoce de cette amibe sociale, un transcrit antisens du gène *EB4-PSV* a été identifié. Le site d'initiation de ce transcrit a été localisé à l'intérieur de la région codante et, plus précisément, dans l'exon 3 de *EB4-PSV*. Au cours du développement, l'ARNm de *EB4-PSV* s'accumule, alors que l'antisens est très peu présent. Lors de la désagrégation après la maturation des spores, la transcription de l'antisens débute et il finit par s'accumuler en grande quantité. On constate alors une diminution du taux d'ARN *EB4-PSV*. Pour expliquer ce phénomène, l'hypothèse a été avancée qu'un excès d'ARN antisens entraînait la formation d'hybrides entre les deux ARN et conduisait à leur dégradation. Pour confirmer cette hypothèse, l'équipe de Nellen a tenté d'inhiber la transcription de

Tableau I						
DIAGRAMME DES RÉSULTATS D'EXPRESSION <i>IN VIVO</i> APRÈS MUTATIONS DES DIFFÉRENTS SITES SU'X', SU1, SU2, SU3, SU'Y'						
Gène	Sites d'initiation	Su'x'	Su1	Su2	Su3	Su'y'
Surf-1	majeur	—	—	↓	↓	—
	intermédiaire	—	↓	—	↓	↓
Surf-2	majeur	—	—	↓	↓	↓
	intermédiaire et mineur	—	—	↓	↓	—

(D'après [20].)

l'ARN antisens et de vérifier si l'ARNm *EB4-PSV* sens restait stable dans ces conditions. Des molécules inhibitrices de la transcription, telles que la daunomycine ou la nogalomyicine, ont été utilisées après déstabilisation de l'ARN endogène par une surexpression du transcrit antisens : le processus de dégradation s'interrompait alors rapidement et les ARNm *EB4-PSV* sens résiduels restaient stables. Ainsi, même si le contrôle de la stabilité par des protéines en *trans* n'est pas à exclure pour la régulation des ARN *EB4-PSV*, les résultats de Hildebrandt et Nellen démontrent l'existence d'un mécanisme régulateur prédominant qui est dû à un ARN antisens endogène. De plus, une RNase spécifique des ARN double-brins a été identifiée dans le cytoplasme où se forment les molécules hybrides [25]. Ce type de RNase a également été mis en évidence chez les eucaryotes, et plus particulièrement dans les cellules Hela [29]. Le gène *c-MYC* humain serait également soumis à ce type de régulation [30].

La régulation des gènes par *editing*

C'est le cas chez le xénope pour le gène codant pour le bFGF (*basic fibroblasts growth factor*) (Tableau II) [24]. Dans cet exemple, le gène

qui se trouve en orientation inverse de bFGF entraîne des modifications de l'expression de l'ARNm transcrit à partir du gène *bFGF* pendant la maturation de l'ovocyte. Des études récentes ont montré l'existence d'une activité cytoplasmique qui permet de convertir la moitié des résidus adénine en résidus inosine et cela uniquement dans la région chevauchante. Il s'agirait, en fait, d'une sorte d'*editing* (*m/s n°7, vol. 8, p. 764*). Cette modification induirait la dégradation de l'ARN du *bFGF* pendant la méiose ou conduirait à la synthèse de protéines non fonctionnelles. La protéine bFGF semble être uniquement nécessaire pendant l'ovogenèse, ce qui expliquerait la destruction de son ARN après maturation au cours de la méiose [24]. Cet ARN antisens semble donc intervenir dans la régulation de la stabilité de l'ARNm *bFGF*.

La régulation des gènes par contrôle de l'épissage

Un autre moyen de régulation, lié à un chevauchement de gènes a été mis en évidence pour le locus *c-erbA-1/ear-1*. Comme nous l'avons vu, le gène *c-erbA-1*, qui est chevauché en 3' par *ear-1*, code pour deux isoformes TR α 1 et TR α 2, grâce à un épissage alternatif (figure 3). L'isofor-

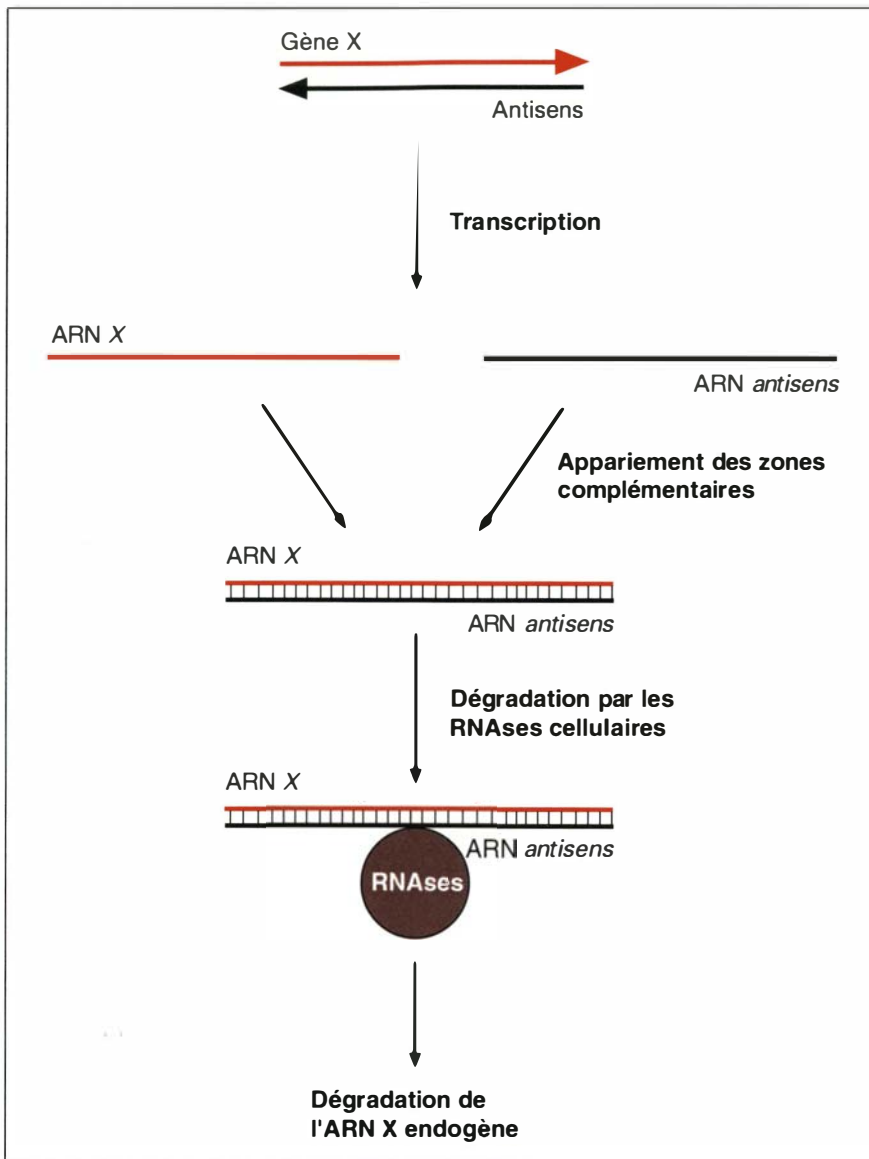


Figure 5. **Régulation de l'expression d'un gène « X » par un ARN antisens.** L'hybridation de l'ARN X et de l'ARN antisens rend l'ARN double-brin beaucoup plus vulnérable aux RNases. Elle entraîne donc la dégradation de l'ARN X, réglant ainsi l'expression de ce gène.

me TR α 2 comprend l'exon 10 qui est chevauchant avec l'exon F de *ear-1*. Or, lorsque l'expression de *ear-1* est augmentée, celle de TR α 2 diminue [17] et celle de TR α 1 augmente. Tout se passe donc comme si le contrôle de l'épissage qui conduit à la formation de TR α 1 et TR α 2 dépendait de l'expression du gène *ear-1*. Le chevauchement semble être ici un moyen de régulation du rapport entre les deux isoformes TR α 1 et TR α 2 de *c-erbA-1*. Cependant, les

mécanismes précis de cette régulation restent encore à découvrir.

Chevauchement de gènes et évolution

L'étude des chevauchements de gènes au cours de l'évolution nous conduit à déterminer leur processus de formation et à enrichir ainsi les connaissances sur les mouvements des gènes.

Ce que l'on connaît de l'évolution





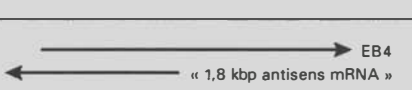
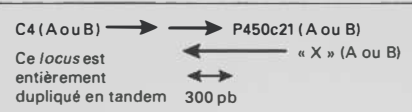

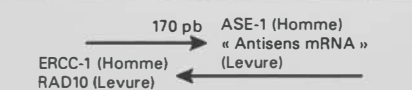

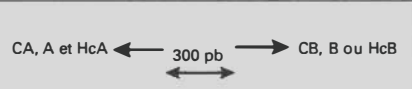
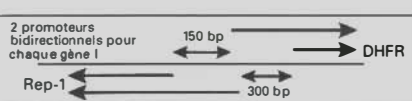
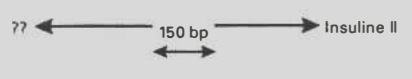


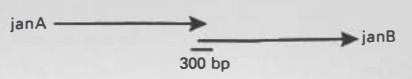
des récepteurs nucléaires permet de formuler des hypothèses quant à la genèse du locus *c-erbA-1/ear-1*. On sait en effet que *c-erbA-1* et *ear-1* sont des gènes anciens. Nous avons vu qu'il existe un homologue de *ear-1* chez la drosophile : c'est le gène *E75*. *c-erbA-1* est présent chez tous les poissons et les amphibiens, mais seule l'isoforme TR α 1 qui code pour le récepteur normal y a été trouvée jusqu'à présent. En fait, l'isoforme TR α 2 semble spécifique des mammifères ; l'équipe de Vennström [31] l'a recherchée sans succès chez le poulet. On peut donc considérer TR α 2 (et donc l'exon 10 de *c-erbA-1*) comme une acquisition récente. Cet exon 10 et l'épissage alternatif qui en découle pourraient avoir été mis en place lors d'une translocation du gène *ear-1* dans le gène *c-erbA-1* (ou l'inverse). Il y aurait donc clairement un lien entre l'épissage alternatif et le chevauchement. Il reste à montrer quand cet événement s'est produit...

Un deuxième exemple d'étude de genèse de locus chevauchant concerne le locus *Surfeit*. Les regroupements géniques chez les vertébrés peuvent être de deux types : l'un est constitué d'un ensemble de gènes qui apparaissent après duplication d'un gène ancestral (gène de la globine, gènes à homéoboîte) ; l'autre comprend des gènes qui sont liés fonctionnellement, mais qui ne possèdent aucune homologie de séquences nucléiques ou de structures protéiques. Le locus *Surfeit* appartient à la deuxième catégorie, puisque aucune homologie n'existe entre les six gènes qui forment le locus *Surfeit*. Ces gènes sont tous exprimés de façon ubiquitaire et chacun possède des propriétés caractéristiques des gènes d'expression constitutive.

Il n'est tout de même pas exclu que le hasard ait rassemblé ces différents gènes qui, dans ce cas, resteraient ensemble non pour des raisons fonctionnelles mais simplement parce que leur séparation est impossible à cause des faibles distances qui les séparent. Lors de l'étude, au cours de l'évolution, de la conservation du locus *Surfeit* (d'abord décrit chez la souris), les auteurs se sont penchés sur son existence chez le poulet [32, 33]. Ils se sont aperçus

Tableau II

CLASSIFICATION ET STRUCTURES D'UNE QUINZAINE DE *LOCI* CHEVAUCHANTS REPRÉSENTATIFS

	Gènes	Fonction Espèces	Structure simplifiée	Références
I	NF-1 EVI2 NF1-c2	Neurofibromatose Inconnue Homme		Wallace <i>et al.</i> , <i>Science</i> 1990 ; 249 : 181.
	GnRH SH	Gonadolibérine Inconnue Homme		Adelman <i>et al.</i> , <i>Science</i> 1987 ; 235 : 1514.
II	bFGF antisens-FGF	Facteur de croissance des fibroblastes Inconnue Xénope, Bœuf		Kimelman, Kirschner, <i>Cell</i> 1989 ; 59 : 687.
	HAP3 « 3 kb mRNA »	Régulateur de la fonction respiratoire Inconnue <i>Saccharomyces cerevisiae</i>		Hahn <i>et al.</i> , <i>Mol Cell Biol</i> 1988 ; 8 : 655.
	EB4 « 1,8 kbp antisens »	Sporulation Inconnue <i>Dictyostelium discoïdum</i>		Hildebrandt et Nellen, <i>Cell</i> 1992 ; 69 : 197.
III	C4 P450c21 « X »	Composant du complément Stéroïde 21 hydroxylase Inconnue Homme		Morel <i>et al. Proc Natl Acad Sci USA</i> 1989 ; 8 : 6582.
	Ddc « X »	Dopa décarboxylase Inconnue <i>Drosophile</i>		Spencer <i>et al.</i> , <i>Nature</i> 1986 ; 322 : 279.
	ERCC-1/ RAD10 ASE-1	Réparation de l'ADN Inconnue <i>Saccharomyces</i> → <i>Homme</i>		Van Duin <i>et al.</i> , <i>Mol Cell Biol</i> 1989 ; 9 : 1794.
IV	HTF9 : Gène A et gène C	Îlot CpG de fonction inconnue <i>Souris</i>		Somma <i>et al.</i> , <i>Nucleic Acids Res</i> 1991 ; 19 : 2817.
	CA et CB A et B HcA et HcB	Protéines du chorion (3 loci différents, tous sur le même modèle) <i>Bombyx</i>		Spoerel <i>et al.</i> , <i>J Mol Biol</i> 1993 ; 230 : 151.
	DHFR Rep-1	Dihydrofolate réductase Réparation de l'ADN <i>Souris</i>		Linton <i>et al.</i> , <i>Mol Cell Biol</i> 1989 ; 9 : 3058.
V (?)	Insuline II ??	Hormone Inconnue <i>Rat</i>		Efrat, Hanahan, <i>Mol Cell Biol</i> 1987 ; 7 : 192.
	H2B.1 H2A	Histones Homme		Collart <i>et al.</i> , <i>J Cell Biochem</i> 1992 ; 50 : 374.
	PCNA « X »	Réplication, prolifération Inconnue Homme		Rizzo <i>et al.</i> , <i>Exp Cell Res</i> 1990 ; 188 : 286.
	janA janB	Détermination du sexe <i>Souris</i>		Yanicostas <i>et al.</i> , <i>Mol Cell Biol</i> 1989 ; 2526.

Les flèches représentent le sens transcriptionnel des gènes.

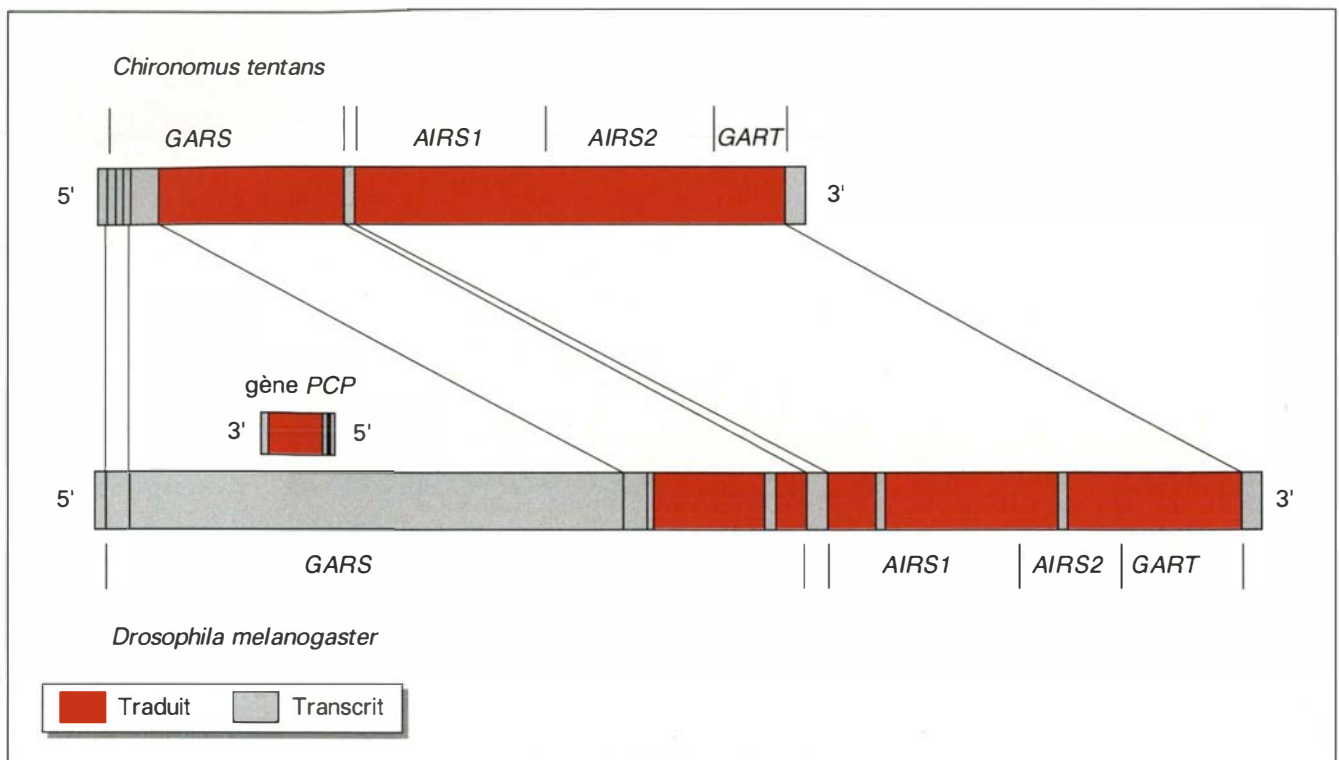


Figure 6. **Comparaison de l'organisation du locus Gart de Chironomus tentans et de Drosophila melanogaster.** Les gènes du locus Gart codent pour trois enzymes qui interviennent dans la voie de la biosynthèse des purines : glycinamide ribonucléotide synthétase (GARS), amino-imidazole ribonucléotide synthétase (AIRS), et glycinamide ribonucléotide transformylase (GART). Le gène PCP qui code pour une protéine de cuticule sur le brin opposé est inséré complètement dans un intron du gène GARS chez Drosophila melanogaster.

que le gène *Surf-3*, qui code pour une protéine ribosomique, est très conservé chez le poulet. On sait que ce gène appartient à une famille multigénique chez les mammifères (15-30 copies). Un grand nombre de ces copies se trouvent être des pseudogènes. Cependant, il semble n'exister qu'une seule copie de *Surf-3* chez les oiseaux, les amphibiens, les insectes, et certaines levures et deux copies chez *Saccharomyces cerevisiae*. Il y a donc d'abord eu mise en place du locus *Surfeit* avec une seule copie de *Surf-3*, puis, chez les mammifères, duplication de *Surf-3* pour donner plusieurs loci. Il apparaît donc très intéressant de savoir pourquoi brusquement le gène *Surf-3* s'est trouvé dupliqué chez les mammifères. Les autres gènes du locus *Surfeit* ne semblent exister que sous une simple copie et ne codent pas pour une protéine ribosomique.

En fait, au moins quatre autres gènes du locus *Surfeit* semblent être conservés chez les oiseaux : *Surf-1*, 2, 4 et 5. Ils ont, avec *Surf-3*, la même organisation topographique chez le poulet que chez la souris. Les distances qui séparent ces différents gènes chez ces deux espèces semblent être conservées. Les îlots riches en GC associés aux extrémités 5' des gènes *Surf* dans le locus murin sont eux aussi conservés chez les oiseaux.

L'existence d'un promoteur bidirectionnel permet de comprendre la conservation des gènes *Surf-1* et *Surf-2*. De plus, la conservation topographique des gènes *Surf-1* et 3, et 2 et 4, qui se chevauchent à leur extrémité 3', fait envisager pour ces gènes un rôle physiologique et une régulation par ARN antisens. En effet, des événements peuvent intervenir à différents moments du cycle cellulaire

ou à des stades différents de la différenciation, qui provoqueraient l'augmentation de l'expression d'un gène *Surf* et induirait ainsi la diminution de l'expression du gène *Surf* voisin sans faire intervenir une régulation transcriptionnelle. De même, la conservation des îlots riches en GC montre que l'organisation est importante pour l'expression des différents gènes.

Le troisième exemple d'étude de la genèse d'un locus chevauchant concerne le locus *Gart* chez *Drosophila melanogaster*. Ce locus appartient à la classe I (figure 6) et contient, en orientation inverse dans son premier intron, un gène appelé *PCP* (*pupal cuticule protein*). Un homologue du gène *Gart*, appelé *ADE5,7* a été mis en évidence chez *Saccharomyces cerevisiae* [34]. Chez cette espèce, la partie 5' du gène *Gart* de *Drosophila melanogaster* n'exis-

te pas. On suppose donc qu'une extrémité 5' a été rajoutée au cours de l'évolution conduisant à l'apparition de l'intron où est inséré le gène *PCP*. Justement, chez une autre espèce, le diptère *Chironomus tentans*, l'emplacement de cet intron est conservé, mais il est beaucoup trop petit (237pb) pour contenir le gène *PCP*, constitué de 4142 pb chez *Drosophila melanogaster* [35]. L'intron qui contenait le gène *PCP* n'existe pas : la structure en gènes imbriqués n'est donc pas conservée chez *Chironomus tentans*. On peut alors supposer que, spécifiquement dans la lignée de *Drosophila melanogaster*, il y aurait eu transposition du gène *PCP* dans l'intron en 5' du gène *Gart*.

Ces différents exemples nous permettent de nous interroger sur les mécanismes ayant conduit à l'apparition de ces gènes à configuration complexe, ainsi que sur les complexités supplémentaires que ces structures entraînent au niveau de la régulation. On a parfois l'impression que le jeu n'en vaut pas la chandelle. Ces structures résultent probablement, au départ, d'événements fortuits (translocation, duplication, « mouvement divers » de l'ADN) sélectionnés parce que les conséquences étaient neutres ou favorables. Dans un certain nombre de cas, il semble bien que la mise en place d'une structure chevauchante soit un excellent moyen d'adapter des mécanismes de régulation de gènes plus fins et plus précis, différents de ceux utilisés couramment à partir des promoteurs. Jusqu'à présent, trois modes de régulation ont été mis en évidence : l'utilisation d'ARN antisens, l'*editing* et le contrôle de l'épissage. La poursuite de l'étude de ces structures devrait nous permettre de trouver d'autres modes de régulation originaux de l'expression des gènes ■

Summary

Overlapping genes

Genes are generally structurally independent and separated from each other by several thousands base pairs of nucleotides. Recently several overlapping genes have been described. This new genetic organization was initially found in viral genomes the size of which is submitted to a strong pressure; however, an increasing number of overlapping genes is now found in both prokaryotic and eukaryotic genomes. The overlapping genes can be divided into four structural classes : (1) genes inserted into introns of other genes. For example, the gene cluster present in two introns of the *drosophila Dunce* gene ; (2) superimposed genes in which the coding sequences of both genes are completely overlapping. An example of this class is the *N-myc locus* which contains in the opposite strand the *N-cym* gene ; (3) 3' overlapping genes such as the *c-erbA-1/ear-1* genes, members of the nuclear hormone

receptor superfamily. In this case, the two 3' exons of the genes are encoded by the same 1 kbp DNA fragment ; (4) the bidirectional promoters. For this latter class of overlapping genes, there is no overlap of the coding regions but only of their regulatory regions. This last class is exemplified by the complex mouse *Surfeit locus* in which two genes are regulated by a common 73 bp promoter. In all of these cases, the unusual gene organization may be viewed as a new way of regulating gene expression not based on promoter(s) and enhancer(s) but on the vicinity of the two partners. Data demonstrating the influence of the overlapping regions on gene expression have been recently published, confirming that the overlapping gene may produce an antisense-RNA which could hybridize to the transcript of its partner gene and induce its degradation (*EB4-PSV*). The study of overlapping genes is also interesting from an evolutionary point of view since it allows to trace back the various events which can move genes from one place to another in the genome.

Remerciements

Nous tenons à remercier Dominique Stéhelin pour le soutien apporté tout au long de notre travail. Nous remercions Jean Coll, Yvan de Launoit, Catherine Hänni, Véronique Fafeur et Guillaume Adelmant pour la lecture du manuscrit. Merci également à tous les membres de l'équipe Endocrinos pour leurs discussions et leur gentillesse.

TIRÉS À PART

V. Laudet