m/s DOSSIER

médecine/sciences 1992 ; 8 : 966-71

La carte à pas de géant

J. David Grausz, Sylvie Paulien

Trois surprises de taille attendaient les participants à la conférence sur le séquençage et la cartographie du génome, édition 1992, qui s'est tenue à Cold Spring Harbor. Les laboratoires européens, bien représentés, étaient dans de nombreux cas les acteurs des progrès réalisés.

La première grande surprise : la construction de *contigs* atteignant la taille d'un chromosome humain

Le chromosome 21q, 42 Mb (Ilia Chumakov *et al.** dans l'équipe de Daniel Cohen, au CEPH, Paris et au GÉNETHON, Évry, France) et 40 % du chromosome Y (la partie euchromatique de ses 42 Mb (Doug Vollrath *et al.**, *Whitehead Institute*, Cambridge, MA, USA) ont été couverts par des clones de YAC chevauchants. Les deux laboratoires utilisent la même approche fondée sur la technique de *top-down* PCR qui permet l'identification de STS (2 à 4 par Mb) et la sélection des YAC correspondants. Au CEPH à Paris, Ilia Chumakov *et al.** ont utilisé la toute nouvelle approche des *Alu*-PCR pour présélectionner 220 YAC spécifiques du chromosome 21 à partir de la banque génomique humaine totale (en grande partie grâce aux méga-YAC développés par Pierre Ougen *et al.**, au CEPH à Paris, dont la taille des insertions est de l'ordre du mégabase) *(figure 1)* [1].

L'équipe de David Page (*Whitehead Institute*) (S. Foote *et al.** [2]) a, quant à elle, réussi à localiser physiquement 160 STS correspondant à un ensemble de 60 délétions constitutives du chromosome Y. Ils ont ensuite identifié 200 YAC (d'une taille moyenne de 600 kb) préparés à partir de l'ADN d'un individu XYYYY, et ils ont construit des *contigs* en faisant, en moyenne, une série de 25 PCR par YAC selon la très efficace stratégie de mise en ordre des clones en trois dimensions.

La deuxième grande surprise a été le séquençage rapide des ADNc que l'équipe de Craig Venter appelle des EST *(expressed sequence tags)***

Craig Venter et son équipe (NIH, Bethesda, MD, USA) (Mark Adams *et al.**), ont ainsi démontré la puissance de cette méthode qui consiste à séquencer partiellement 5 000 ADNc dérivés de messagers exprimés au niveau du cerveau, 4 000 d'entre eux n'ayant pas d'homologie avec des séquences déjà connues. Il s'agit manifestement d'une surestimation qui nécessite l'élimination des séquences provenant d'un même clone d'ADNc, et de fragments en amont et en aval d'un gène connu. Comme nous le verrons plus en détail, une estimation réelle de la fréquence de ces gènes « orphelins » a pu être faite chez le nématode *C. elegans* (John Sulston *et al.*, Cambridge, GB), et bien que ce pourcentage soit inférieur

*** GLOSSAIRE ***

***Contig** : carte physique d'une région du génome établie grâce à l'arrangement de fragments clonés partiellement chevauchants.*

EST (expressed sequence tag) *: séquence partielle d'une séquence transcrite, c'est-à-dire avant tout d'un ADNc. Un EST permet, comme un STS, de disposer d'une sonde pour un site particulier du génome. De plus, ce site, dans ce cas, est un gène exprimé.*

PCR (polymerase chain reaction) *: amplification enzymatique* in vitro *exponentielle d'une séquence d'ADN entre deux amorces.*

***Alu-PCR** : amplification par PCR en utilisant comme amorces des oligonucléotides spécifiques de séquences répétées Alu, disséminées dans le génome et typiques de l'espèce humaine. Cela permet de disposer de sondes humaines constituées par les séquences uniques situées entre deux motifs Alu.*

RFLP (restriction fragment length polymorphism) *: polymorphismes de certains sites de restriction tels qu'une enzyme de restriction donnée va engendrer, selon les individus, des fragments d'ADN de tailles variables, reconnus par une même sonde d'ADN.*

SSLP (simple sequence length polymorphism) *: polymorphismes génétiques de séquences simples répétées, entourées de séquences uniques : ces régions sont également appelées « micro-satellites ». Une PCR des amorces complémentaires des séquences uniques entourant une répétition de type $(CA)_n$ permet de déterminer directement la valeur, polymorphique, de n.*

STS (sequence tag site) *: site sur l'ADN défini par sa séquence. Une sonde spécifique de ce site peut aisément être produite par PCR avec deux amorces, généralement espacées d'environ 200 pb.*

YAC (yeast artificial chromosome) *: chromosome artificiel de levure, permettant de cloner des grands fragments d'ADN.*

* *Indique les abstracts publiés à Cold Spring Harbor,* Genome Mapping and Sequencing, *6-10 mai 1992.*

** *EST dans le cas de Craig Venter, le terme d'EST n'est pas tout à fait approprié — les véritables* sequence tags *doivent :*
être localisées dans leur région chromosomique alors que seulement 180 des 5 000 séquences décrites ci-dessus l'ont été,
être exprimées et on ne sait pas si la majorité sont exprimées. En effet, certaines correspondraient à des artefacts de clonage plutôt qu'à de réels ADNc (Tableau I, [7]).
Le travail de C. Venter est complémentaire des travaux bien documentés et précis des groupes anglais (John Sulston pour les nématodes, Ross Sibson et Sidney Brenner pour l'homme), français (Charles Auffray) et japonais (Kenichi Matsubara et Nobuo Nomura). Un effort de nomenclature est actuellement en cours.

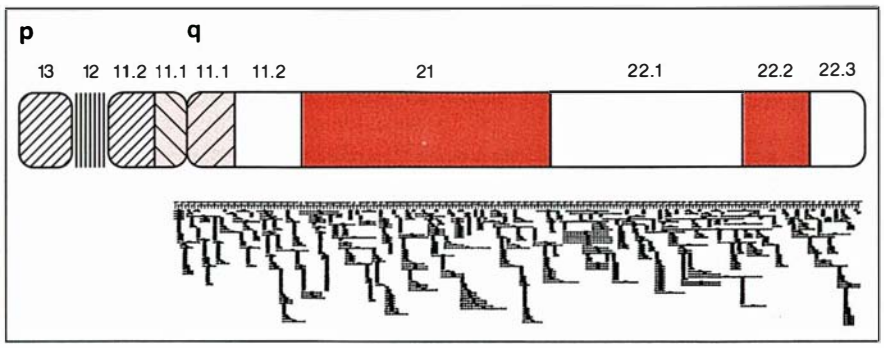
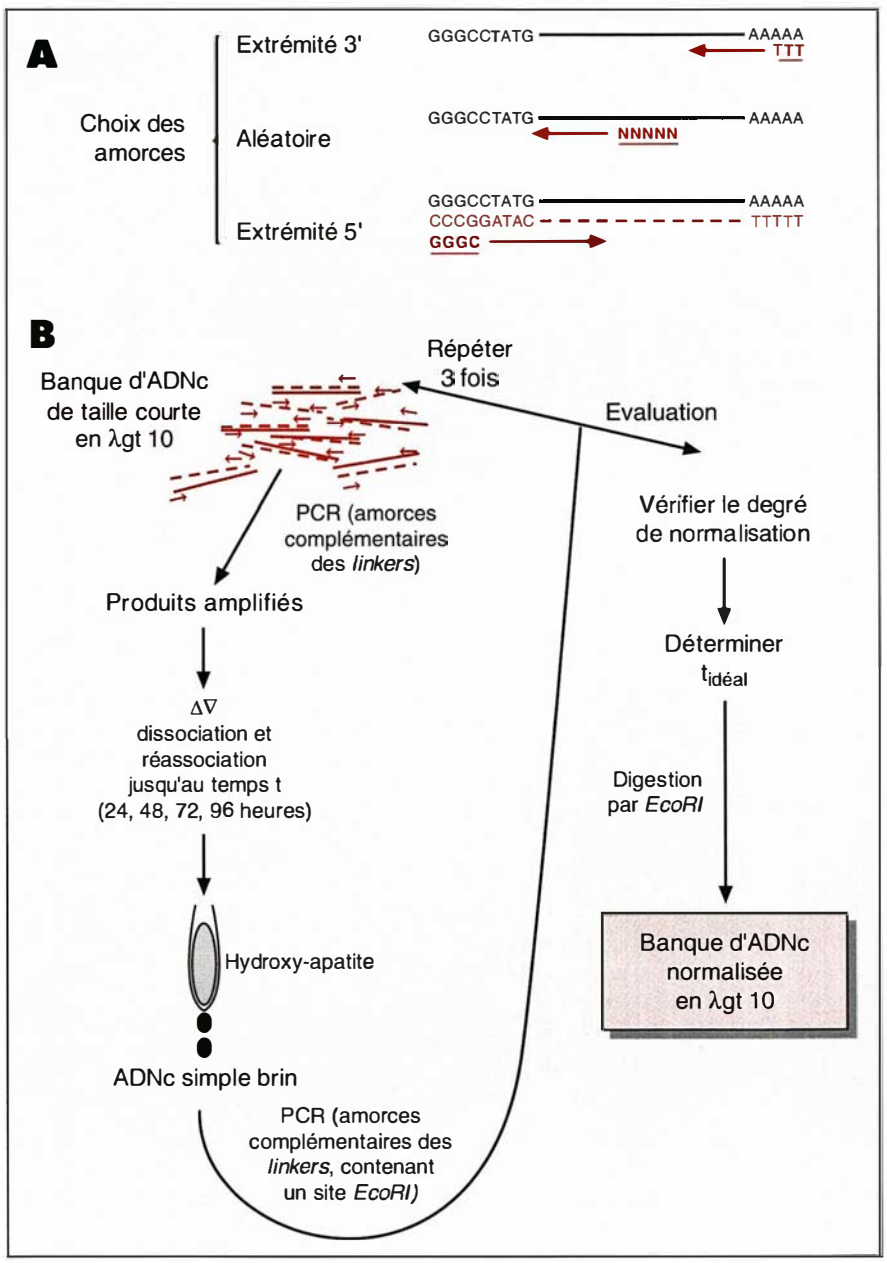


Figure 1. **Contig *du bras long du chromosome 21*** *[1].*



aux 60 % obtenus chez ce même organisme en séquençant au hasard des clones d'ADNc, il pourrait pourtant correspondre à 50 % ou plus de nouvelles séquences obtenues. De plus, 200 séquences obtenues à partir d'une banque d'ADNc de cerveau humain ont été assignées à leur chromosome d'origine (Michael H. Polymeropoulos *et al.** de l'ATCC *american type culture collection*). Une seule provient du chromosome 21, alors qu'un nombre impressionnant d'entre elles (50 % ou plus) est localisé sur le chromosome 1.

Deux points méthodologiques importants doivent être discutés par rapport au séquençage d'ADNc *(figure 2)* et concernant la banque utilisée comme source de transcrits :

- il peut s'agir d'une banque préparée soit à partir d'amorces oligo-dT, s'hybridant avec les extensions poly-A des messagers, soit d'amorces aléatoires ;
- la banque peut être normalisée, en égalisant la représentation des clones correspondant à des ARNm abondants, ou bien en reflétant l'abondance réelle de ces messagers dans les tissus.

La troisième grande surprise qui a révolutionné le séquençage à grande échelle

C'est le succès remporté sur les 315 kb du chromosome III de *S. Cerevisiae* (CEE, projet *BRIDGE*, C. J. Herbert *et al.**, Gif-sur-Yvette, France [4]), ainsi que les rapides progrès réalisés sur trois autres chromo-

Figure 2. ***Banques d'ADNc pour le séquençage.*** *L'originalité de l'information et des séquences varie en fonction des banques d'ADNc (amorces utilisées et degré de normalisation préalable).* ***A.*** *Différentes procédures d'obtention des ADNc double brins.* ***B.*** *Normalisation d'une banque d'ADNc, c'est-à-dire appauvrissement en séquences très abondantes. Des ADNc double brins sont amplifiés par PCR, puis dissociés et rehybridés dans des conditions contrôlées et pendant des temps variables. Les séquences abondantes se rehybrident le plus vite, redonnant des double brins qui sont retenus sur la colonne d'hydroxyapatite. Les simple brins non retenus sont amplifiés et la procédure est répétée.*

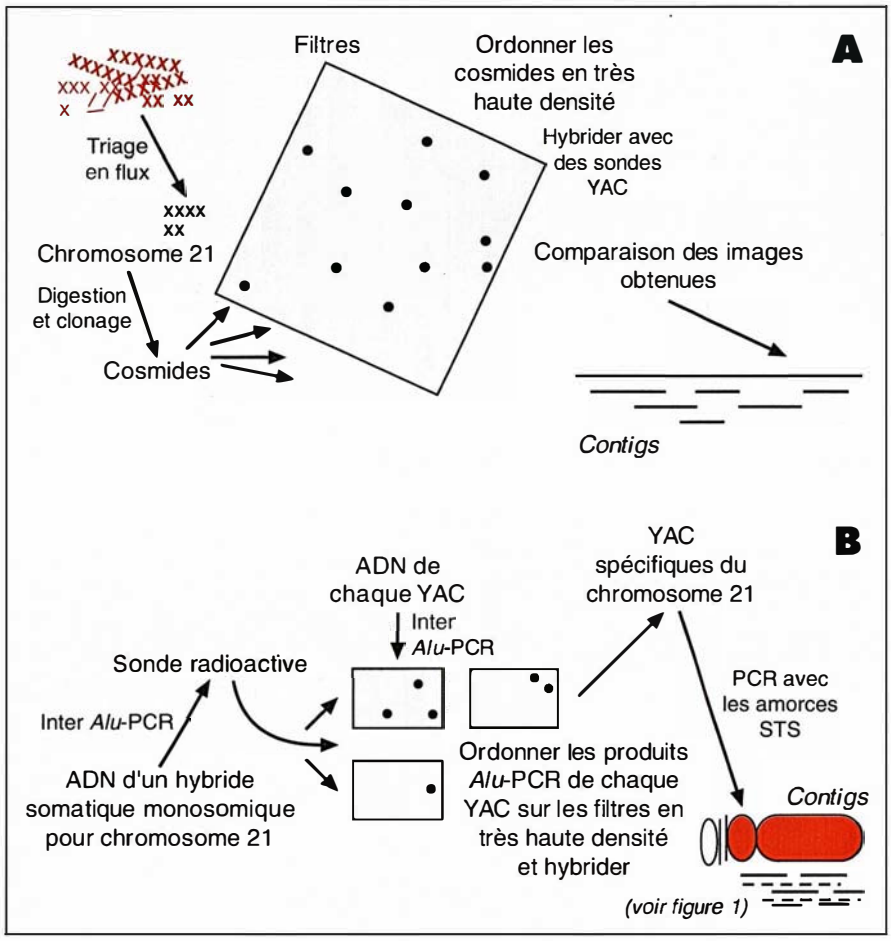


Figure 3. ***Comparaison des méthodes d'obtention de* contigs *du chromosome 21q où la préférence va à la PCR. A.*** *Organisation de* contigs *de cosmides par hybridation avec des YAC (groupe de Lehrach).* ***B.*** Contigs *de YAC par utilisation de l'*Alu*-PCR et des STS (groupe de Daniel Cohen).*

Tableau I

CORRESPONDANCE ENTRE SÉQUENCE ET GÈNES CONNUS (D'APRÈS [5])

	Gènes	Pourcentage connu	Référence
Génome			
C. elegans chromosome III	32	44	[3]
S. cerevisiae chromosome III	182	33	[4]
chromosome IX	46	33	[5]
	Clones	**Pourcentage connu**	**Référence**
ADNc			
C. elegans	1 194	29	[6]
	422	25	[7]
H. sapiens (cerveau)	5 000 (~60 % codants) [9, 10]	33	[8]

somes de la levure. Quatre-vingt pour cent des phases ouvertes de lecture, récemment découvertes grâce aux travaux réalisés sur les ADNc, correspondent à des gènes n'appartenant pas à une famille connue. Cependant, cette apparente absence de similitude pourrait refléter une autre propriété suggérée par les nouvelles séquences obtenues, c'est-à-dire l'existence de séquences spécifiques de l'espèce. La comparaison existe aussi au niveau des motifs kinase, *leucine zipper*, etc., et des structures secondaires et/ou tertiaires, hélices α, feuillets β, etc., des protéines [9, 10] qui ne sont pas toujours prises en compte quand on fait une comparaison des séquences d'acides aminés.

Le choix des méthodes de sélection des clones, soit par hybridation, soit par PCR, reste sujet à une controverse

Dans le cas de la cartographie physique du chromosome 21, la sélection à l'aide de la PCR apparaît manifestement d'une efficacité supérieure, car à la fois plus rapide et moins contraignante au niveau technique. En dépit d'un investissement technique impressionnant et de leur avance importante dans ce domaine, Hans Lehrach et son équipe (Mark T. Ross *et al.**, ICRF, Londres, GB) ont utilisé la méthode d'hybridation avec des filtres de YAC de haute densité, mais n'ont pas réussi à construire un *contig* du chromosome 21 aussi complet que celui réalisé par l'équipe de Daniel Cohen *(figure 3)* grâce aux deux techniques *Alu*-PCR et STS (par PCR également).

Cependant, la technique d'hybridation permet d'établir une meilleure corrélation entre les séquences exprimées et les séquences génomiques *(figure 4)*. Premièrement, elle permet d'éliminer les séquences répétées, par exemple en ajoutant de l'ADN hau-

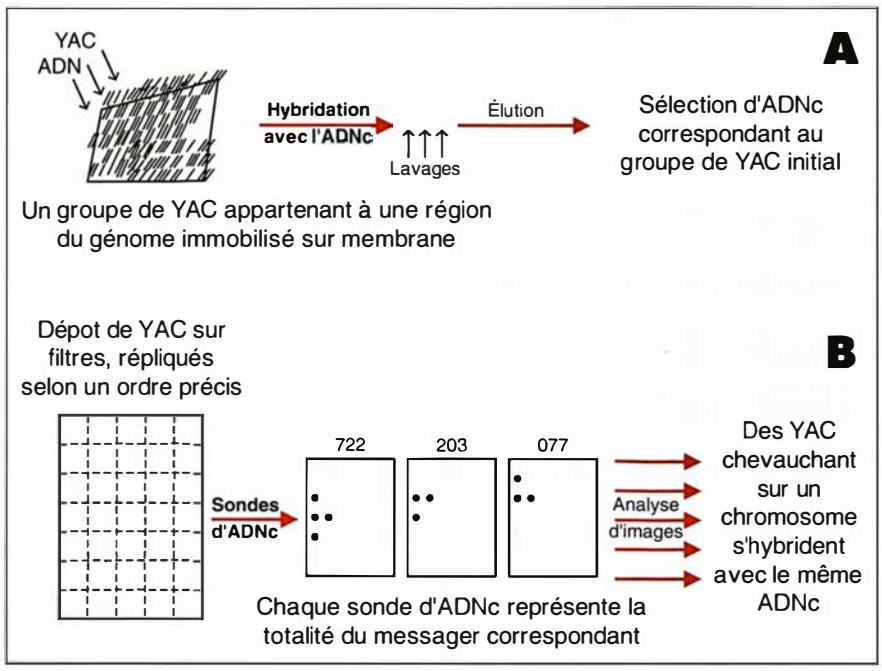


Figure 4. ***Deux circonstances où la préférence va à l'hybridation*** *(établissement de la relation entre séquences exprimées et séquences codées).* ***A. Direction génome vers l'ADNc.*** *Les clones de YAC immobilisés sur filtres peuvent être utilisés pour sélectionner un ensemble de clones d'ADNc situés sur une même région du génome (équipe de Sherman Weissman [Parimoo et al.,* Das Gupta *et al,* New Haven, CT, USA*]).* ***B. Direction ADNc vers le génome.*** *Afin de hiérarchiser le génome du nématode, R. Wilson* et al. *ont ordonné 958 clones de YAC sur filtres (appelés des* polytene filters*) représentant deux fois le génome de* C. elegans *et hybridés avec un clone d'ADNc [6]. Cette méthode a permis à ces auteurs, ainsi qu'à l'équipe de Craig Venter [7], de déterminer la localisation des clones. Ces derniers avaient accès à l'information que contient la séquence d'environ 350 bases, mais ils ont choisi de préparer une sonde en utilisant des amorces aléatoires le long de l'ADNc (dont la longueur moyenne est de 1,5 kb), puis de l'hybrider.*

tement répétitif au tampon d'hybridation. Deuxièmement, elle évite de se positionner obligatoirement dans un exon, les séquences d'ADN complémentaires s'hybridant malgré la présence d'introns. Troisièmement, elle permet de sélectionner les séquences exprimées qui sont en nombre limité. Une méthode similaire fondée sur la technique de la PCR aurait l'inconvénient de dévoiler à la fois des séquences homologues légitimes et des séquences illégitimes ayant une certaine similitude de séquence.

L'exploitation des banques de YAC, en vue de cartographier le génome à l'échelle du mégabase

C'est une source de renseignement sur la structure même de certaines régions chromosomiques qui se sont avérées jusqu'à présent difficiles à cloner. Cela a été précédemment démontré dans le cas des méga-YAC du CEPH à Paris (Pierre Ougen *et al.**), et se révèle vrai pour chaque secteur de la cartographie physique. Les YAC sont manifestement plus stables que les cosmides, mais des améliorations ont été apportées aux techniques de clonage dans le phage P1 (N. S. Shepherd *et al.**, Wilmington, DE, USA) et les cosmides améliorés (BAC *vectors*, Shizuya *et al.**, Pasadena, CA, USA) dont la taille maximale des insertions est de l'ordre de 150 et 300 kb respectivement. Le chimérisme, qui limitait l'utilisation des YAC, paraît concerner la majorité des clones localisés dans certaines régions chromosomiques (Gillian P. Bates *et al.**, ICRF, Londres, région 4p16.3 associée à la maladie de Huntington). Les résultats démontrent que certaines régions du génome sont instables (lorsqu'on les clone dans des levures, mais probablement aussi *in vivo*), et d'autres fragments d'ADN ne peuvent être clonés (en raison de délétions constantes, bien que dérivant d'une partie relativement stable d'une région instable). De telles observations confirment l'intérêt d'utiliser dans une même proportion les deux techniques de clonage dans les YAC et les cosmides/phages (BAC ou P1) dont les contraintes sont complémentaires.

Les différents objectifs choisis en vue de cartographier le génome

Les équipes qui utilisent la technique des YAC ont pour objectif de cloner des insertions de plus grandes tailles, celui des équipes qui utilisent la technique d'hybridation *in situ* est d'atteindre une résolution beaucoup plus fine à l'aide de fragments plus petits. L'objectif commun et ambitieux, cartographier l'ensemble du génome humain, est aujourd'hui un projet tout à fait réalisable grâce aux des nombreuses et différentes techniques utilisées. La cartographie physique du génome à grande échelle est une réalité, et les techniques utilisées actuellement permettent de faire la relation entre les distances génétiques, de l'ordre de quelques dizaines à quelques centaines de mégabases par chromosome, et les distances physiques qui sont de l'ordre de quelques centaines de bases, c'est-à-dire établir

Les récents efforts développés pour séquencer les clones d'ADNc

Alors que l'ARNm doit être intact, seulement 350 bases environ seront séquencées. Ces séquences peuvent être choisies en fonction de leur position aux extrémités 5', aux extrémités 3' non codantes, ou entre ces deux extrémités. Chacune de ces approches a ses avantages. Une fois les séquences leader *et* signal *éliminées, les séquences situées aux extrémités 5', qui codent pour une protéine, peuvent être comparées aux séquences contenues dans une banque de données. Malheureusement ces séquences 5' couvrent probablement plusieurs exons. Les séquences qui correspondent aux extrémités 3' non codantes sont très probablement le produit d'un seul exon. Une fois les séquences répétées (soit environ 15 % des séquences humaines situées aux extrémités 3' d'après le groupe de Craig Venter) et les séquences poly-A éliminées, la séquence ainsi obtenue permet de synthétiser des amorces pour déterminer par PCR la position d'un marqueur (encore appelé* sequence tagged site *ou STS). Ce marqueur peut être utilisé pour identifier les YAC correspondants et facilite ainsi la construction de* contigs. *La séquence codante située en 5' est caractéristique d'une famille de messagers ; elle ne permet pas toujours de différencier les représentants d'une même famille de gènes, ni, par conséquent, de leur attribuer une position unique sur un chromosome. Une séquence codante interne située entre ces deux extrémités présente les mêmes inconvénients qu'une séquence située en 5', mais elle a l'avantage de pouvoir être utilisée avec la partie 5' ou 3' du même clone afin d'identifier toutes les séquences codantes pour des gènes d'une même famille.*

Après avoir orienté les insertions des clones d'ADNc et choisi la direction dans laquelle se fera le séquençage, il faut réfléchir aux moyens nécessaires pour normaliser la banque. En ce qui concerne les organismes qui ont été choisis comme modèles, le séquençage de leur ADNc n'a nécessité aucune normalisation (comme W. R. Crombie et al. *l'ont démontré dans le cas du séquençage du nématode [7]). En fait, même la banque d'ADNc de cerveau humain dont nous avons parlé précédemment n'était pas normalisée. Seules les séquences ribosomiques et mitochondriales avaient été préalablement éliminées. 24 % des séquences des clones du nématode sont redondantes contre seulement 15 % des séquences exprimées au niveau du cerveau humain. Cependant — des études plus approfondies, entreprises, d'une part, par R. Waterston* et al.* *[6], qui ont répertorié tous les ARNm du nématode moyennement exprimés, et, d'autre part, par Okubu* et al. *(Osaka, Japon), qui ont recherché les séquences d'ADNc inconnues — nécessitent soit l'élimination de messagers dont les séquences sont déjà connues, soit le séquençage d'une banque normalisée. La normalisation fragmente obligatoirement l'ADN, ce qui peut être un problème lorsqu'on veut éviter de séquencer deux fois le même ADNc.*

une relation entre les *contigs* et les séquences *(voir encadré)*.

Un tel projet est réalisable en grande partie grâce à l'automatisation et à l'informatique

La construction de *contigs* couvrant une grande distance sur un chromosome, la comparaison de séquences d'ADNc dans les banques de données et le stockage de ces séquences nécessitent la conception de nouveaux programmes informatiques très complexes.

• Les nouvelles techniques d'empreinte génétique *(fingerprinting)* devraient permettre de cartographier l'ensemble du génome (humain) grâce à des modèles sophistiqués adaptés aux programmes (Daniel Cohen *et al.**, *voir m/s n° 8, vol. 8, p. 881* pour les développements les plus récents).

• Les nombreuses possibilités offertes par le projet GRAIL du *Department of Energy* (Richard J. Mural *et al.**, Oak Ridge, TN, USA) et l'équipe dirigée par David States au NCBI (*National Library of Medicine*, Bethesda, MD, USA), permettent de faire des comparaisons de séquences du type ADNc et des recherches d'exons.

RÉFÉRENCES

1. Chumakov I, Rigault P, Guillou S, *et al.* Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992 ; 359 : 380-6.

2. Foote S, Vollrath D, Hilton D, *et al.* The human Y-chromosome : overlapping DNA clones spanning the euchromatic region. *Science* 1992 ; 258 : 60-6.

3. Sulston J, Du Z, Thomas K, *et al.* The *C. elegans* genome sequencing. *Nature* 1992 ; 356 : 37-41.

4. Oliver SG, Van der Aart QJM, Agostini-Carbone ML, *et al.* The complete DNA sequence of yeast chromosome III. *Nature* 1992 ; 357 : 38-46.

5. Chothia C. One thousand families for the molecular biologist. *Nature* 1992 ; 357 : 543-4.

6. Waterston R, Martin C, Craxon M, *et al.* A survey of expressed gene in *C. elegans*. *Nature Genet* 1992 ; 1 : 114-23.

7. McCombie WR, Adams MD, Kelley JM *et al. C. elegans* ESTs identify gene families and potential disease homologies. *Nature Genet* 1992 ; 1 : 124-31.

8. Adams MD, Dubnick M, Kerlavage AR, *et al.* Sequence identification of 2375 human brain genes. *Nature* 1992 ; 356 : 632-4.

9. Bürglin TR, Barnes TM. Introns in sequence tags. *Nature* 1992 ; 357 : 367.

10. Adams MD, Fields C, Venter JC. Reply. *Nature* 1992 ; 357 : 367-8.

Carte génétique de haute résolution de la souris

La vraie vedette de la conférence de Cold Spring Harbor a été le programme de cartographie du génome humain, mais les efforts qui ont été développés pour séquencer le génome d'organismes modèles, en particulier celui de la souris, méritent d'être commentés. L'équipe d'Eric Lander (N. C. Dracopoli et al.* *du Whitehead Institute) a réalisé une carte génétique de haute résolution de la souris. Celle-ci contient environ mille marqueurs du type SSLP* (simple sequence length polymorphisms), *révélés par PCR et intégrés à des cartes précédemment construites à l'aide de marqueurs du type RFLP. Ces marqueurs sont hautement informatifs (polymorphiques dans 90 % des croisements entre individus de groupes différents,* M. musculus *avec* M. spretus, *et 50 % des croisements entre individus d'un même groupe). Ils sont utilisés avec des YAC contenant de grandes insertions (600 kb) pour la construction d'une carte physique de la souris. L'ensemble est ingénieusement installé autour d'une machine à PCR améliorée (sur laquelle peut être placée une douzaine de microplaques à la fois), toutes les amorces choisies ayant une température d'hybridation et des conditions d'amplification cyclique identiques. Le programme réalisé par Steve Lincoln et Eric Lander permet de détecter 80 % des erreurs effectuées lors de l'entrée des données. L'élaboration du projet et son automatisation, de la conception à la réalisation, sont tout à fait impressionnantes. Le* scoop *de cette conférence est la prochaine mise en place d'une collaboration entre cette brillante équipe et le CEPH à Paris afin de réaliser la charpente de la carte physique du génome humain.*

• Avec le temps, la banque de données de séquences *Genome Data Base* devient de plus en plus sophistiquée (Peter Pearson et son équipe, K. A. Brandt *et al.**, Baltimore, MD, USA).

• Un choix impressionnant de possibilités d'interrogation a été présenté à Cold Spring Harbor. Il existe au moins cinq systèmes différents d'interrogation sur tout type d'ordinateur possible. La base de données Chrominfo (Prakash Nadkarni, Steeve Reeders, New Haven, CT, USA) permet d'optimaliser au maximum l'utilisation d'un Macintosh. Mais, il existe aussi des programmes sur PC (George S. Michaels au NIH, Bethesda, MD USA), des stations de travail DEC, Sony, Sparc, etc.

• Alors que le projet de cartographie du génome du nématode semble être le plus avancé, le département informatique correspondant ne manque pas de sophistication. J.-P. Thierry-Mieg et R. Durbin (Montpellier, France et Cambridge, GB), qui dirigent ce département, ont développé une nouvelle base de données facile d'accès, ACEDB, qui devient un modèle du genre pour les autres équipes impliquées dans le séquençage du génome.

Conclusions

Cet exposé a été involontairement écrit à la manière d'une symphonie de Brahms dont le thème se répète chaque fois avec plus d'intensité et plus de vigueur, alors même que vous pensez qu'elle va s'achever. Et pourtant cet exposé est loin d'être exhaustif. Toutes les fois où nous arrivons à une conclusion, une information importante nous vient à l'esprit, et un nouveau mouvement — ou seulement quelques mesures supplémentaires — s'impose. Peut-être cela reflète-t-il l'impression générale qui ressort de cette conférence : loin d'être achevées, les variations les plus récentes, les plus provocatrices et innovatrices, donneront naissance à l'ensemble de la séquence du génome ■

J. David Grausz
docteur ès sciences
S. Paulien
docteur ès sciences
Centre d'étude du polymorphisme humain, 27, rue Juliette-Dodu, 75010 Paris, France.

TIRÉS A PART

J. D. Grausz.

Ce dossier est le compte rendu du congrès *Genome mapping and Sequencing*, Cold Spring Harbor, 1992.