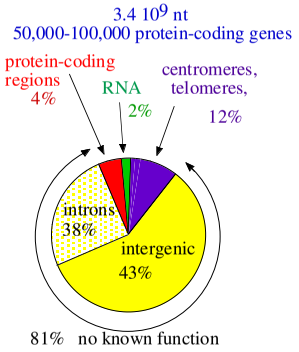
Searching for functional regions (coding or non-coding) in mammalian genomes

- Human genome sequence:
  - 1300 Mb (38 %) available in GenBank (November 22 1999)
  - Draft sequence 90% complete in 2000
  - Finished sequence in 2003
  - Experimental characterization of all human genes. How many years ?
- Organization of the human genome
- Prediction of functional elements by computer analysis of genomic sequences
  - State of the art
  - Success and pitfalls of different approaches
- Prediction of function by homology
  - Orthology/paralogy

Functional elements in the human genome

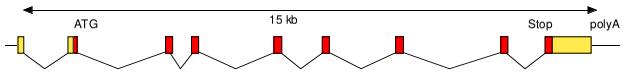


Untranslated RNAs: Xist, H19, His-1, bic, etc.

Regulatory elements: promoters, enhancers, etc.

Repeated sequences (SINES, LINES, HERV, etc.) : 40% of the human genome

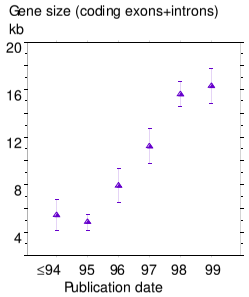
Structure of human protein genes

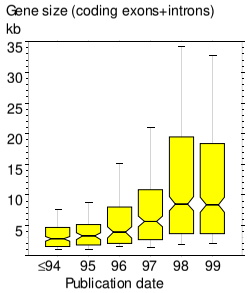


- 1396 complete human genes (exons + introns) from GenBank
- Average size (25%, 75%)
  - Gene: 15 kb ± 23 kb (4, 16) (10% > 35 kb)
  - CDS: 1300 nt ± 1200 (600, 1500)
  - Exon (coding): 200 nt ± 180 (110, 200)
  - Intron: 1800 nt ± 3000 (500, 2000)
  - 5'UTR: 210 nt (Pesole et al. 1999)
  - 3'UTR: 740 nt (Pesole et al. 1999)
- Intron/exon
  - Number of introns: 6 ±3 introns / kb CDS
  - Introns / (introns + CDS): 80%
  - 5' introns in 15% of genes (more ?), 3 'introns very rare
- Alternative splicing in more than 30% of human genes (Hanke et al. 1999)

Structure of human protein genes

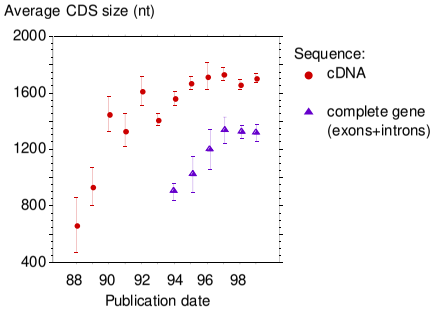
- GenBank: bias towards short genes
- 1396 complete human genes (exons + introns)





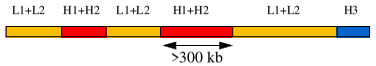
Structure of human protein genes

- GenBank: bias towards short genes
- 1396 complete human genes (exons + introns)
- 9268 complete human mRNA



Isochore organization of the human genome

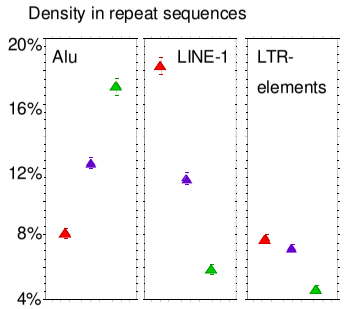
Bernardi et al. 1985



isochore	%C+G	% total genomic DNA
L1+L2 :	33%-44%	62 %
H1+H2 :	44%-51%	31%
H3 :	51%-60%	3-5%

- **Insertion of repeated sequences** (A. Smit 1996)
- **Recombination frequency** (Eyre-Walker 1993)
- **Chromosome banding** (Saccone, 1993)
- **Replication timing** (Bernardi, 1998)
- **Gene density** (Mouchiroud, 1991)
- **Gene expression ??** -> No
- **Gene structure** (Duret, 1995)

Isochores and insertion of repeat sequences



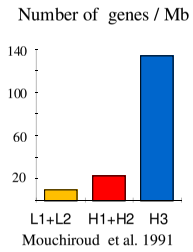
G+C content of genomic sequence:

- G+C < 39%
- G+C 39%-47%
- G+C > 47%

4419 human genomic sequences > 50 kb

Isochores and gene density

isochore	% total genomic DNA	%total genes
L1+L2 :	62 %	31%
H1+H2 :	31%	39%
H3 :	3-5%	30%



MHC locus (3.6 Mb) (The MHC sequencing consortium 1999)
Class I, class II (H1-H2 isochores): 20 genes/Mb, many pseudogenes
Class III (H3 isochore): 84 genes/Mb, no pseudogene

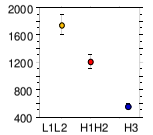
Class II boundaries correlate with switching of replication timing

Isochores and introns length

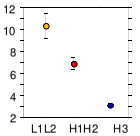
Duret, Mouchiroud and Gautier, 1995

- 760 complete human genes
- L1L2: intron G+C content < 46%
- H1H2: intron G+C content 46-54%
- H3: intron G+C content >54%

Average intron length (bp)



Gene compaction (intron length/coding region length)



Prediction of functional elements (1)

- Ab initio methods
  - Ruled-based or statistical methods
  - e.g.: protein genes prediction, promoter prediction, …
  - Very useful but ...
    - Limits in sensibility/specificity
    - No method available for many functional elements (non-coding RNA genes, regulatory elements, …)
- Large scale transcriptome projects: ESTs, full-length cDNA
  - Identification of transcribed genes (protein or non-coding RNA)
  - Information on alternative splicing, polyadenylation (Hanke et al. 1999, Gautheret et al. 1998), expression pattern
  - Very useful but ...
    - Problems with genes expressed at low level, narrow tissue distribution, stage-specific expression, …
    - Limited tissue sampling
    - Artifacts in ESTs (introns, partially matured RNA, …)
    - Limited to polyadenylated RNA

Prediction of functional elements (2)

- Comparative sequence analysis (phylogenetic footprinting)
  - Function => selective pressure

Corollary

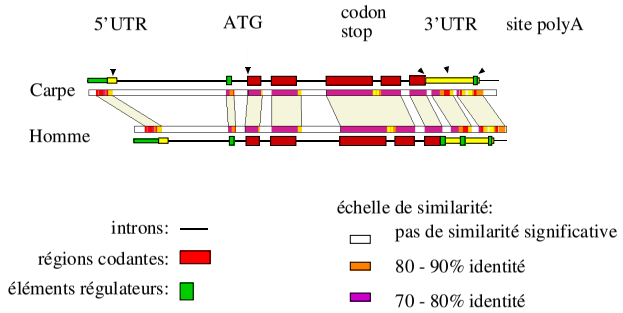
  - Sequence conservation = selective pressure = function

provided the number of aligned homologous sequences represents enough evolutionary time for the accumulation of mutations at the less constrained (presumably selectively neutral) base positions.

- Evolutionary rate in non-functional DNA: ~ 0.3% / My (± 0.069)

Man/Mouse:	~ 80 Myrs	46-58% identity
Mammals/Birds:	~ 300 Myr	26-28% identity
Random sequences		25% identity

Analyse comparative des gènes de β-actine de l'homme et de la carpe



Phylogenetic footprinting

- Advantages
  - Works for all kinds of functional elements (transcribed or not, coding or not) as far as the information is in the primary sequence
  - Does not require any *a priori* knowledge of the functional elements
- Limits
  - Absence of evolutionary conservation does not mean absence of function
  - No efficient method to detect unknown conserved secondary structure in RNA
  - Function, but what function ?
  - Depends on the sequencing status of other genomes
    - Human, mouse, fugu, C. elegans, drosophila, yeast, A. thaliana
  - Number of sequences to compare : > 200 Myrs of evolution
    - Mammals/birds: 310 Myrs
    - Human + mouse + bovine : 240 Myrs

Prédiction de fonction par homologie ?

- Similarité entre séquences homologie
- Homologie structure conservée
- Structure conservée fonction conserv ée

- Oui, mais …
  1. Fonction: concept flou
     - activité biochimique identique ? e.g. même ligand pour un récepteur, même substrat pour une enzyme, même gènes cibles pour un facteur de transcription.
     - distribution tissulaire ? (isoformes tissu-spécifiques).
     - compartimentalisation cellulaire: cytoplasme, mitochondrie, etc.
  1. Protéines homologues de fonction différentes
     - Protéines homologues ligands (activateur/répresseur) d 'un même récepteur
     - Recrutement pour une fonction totalement différente: τ -cristalline / α-énolase
  1. Orthologie/paralogie
  1. Évolution modulaire

Prédiction de fonction par homologie ?

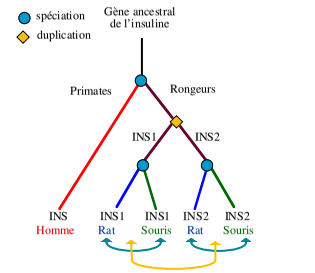
```
MZEORFG: 1     ILNSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLXDNLTLWTSDTNEDGGDE 59
                I N+P++AC LAKQAFD+AI+ELD+L E+SYKDSTLIMQLL DNLTLWTSD  ++   E
BOV1433P: 186 IQNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGE 244

 Score = 87.4 bits (213), Expect = 1e-17
 Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA            MAM       26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA, complete cds.
ACCESSION   J03868

LOCUS       MZEORFG       187 bp    mRNA            PLN       31-MAY-1994
DEFINITION  Zea mays putative brain specific 14-3-3 protein, tau protein
            homolog  mRNA, partial cds.
```

Orthologie/paralogie



Homologie: deux gènes sont homologues si ils ont un ancêtre commun

Orthologie: deux gènes sont orthologues si ils ont divergé à la suite d'un évènement de spéciation

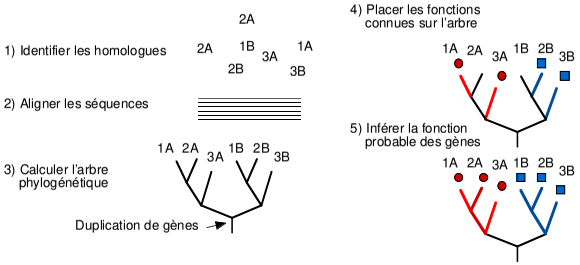
Paralogie: deux gènes sont paralogues si ils ont divergé à la suite d'un évènement de duplication

Orthologie ≠ équivalence fonctionnelle

Diversification fonctionnelle par duplication de gènes au cours de l'évolution des métazoaires

- Iwabe et al. MBE 13:483-493 (1996):
- Analyse phylogénétique de 25 familles de gènes
- isoformes compartimentalisées: duplications avant la divergence animaux/champignons
- paralogues avec activités différentes: duplication avant la divergence vertébrés/arthropode
- isoformes tissu-spécifiques: duplication avant la divergence poissons/tétrapodes

Approche phylogénétique pour la prédiction de fonction



Orthologie/paralogie: abus de langage

- Fitch (1970) Syst. Zool. 19:99-113:

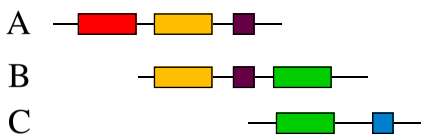
"Where the homology is the result of gene duplication [...] the genes should be called paralogous. Where the homology is the result of speciation [...] the genes should be called orthologous."

- Koonin (1996) TIG, PNAS,Curr. Opin. Genet. Dev.

"By definition, orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species. By contrast, paralogs are homologous genes that have evolved by duplication and code for proteins with similar, but not identical functions".

- Corrigé dans *Science* 1997

Evolution modulaire



Prédiction de régions régulatrices

- Méthodes *ab initio*
  - Prédiction de promoteurs
  - Îlots CpG
- Approche comparative

Prédiction de promoteurs eucaryotes

- Combinaison de sites de fixation de facteur de transcription (ordre, orientation, distance)
- Motifs courts, dégénérés
  - Difficile de distinguer les vrais sites des faux positifs:
  - Motif à 4 bases: ≈1/256 pb (1/128 pb sur les deux brins)
- Boîtes TATA, CAAT , GC: absents dans beaucoup de promoteurs
- Banques de données de sites de fixation de facteurs de transcription (TRANSFAC), de promoteurs caractérisés expérimentalement (EPD)
- PromoterScan (Prestridge 1995): Mesure de la densité en sites potentiels de fixation de facteurs de transcription de long de la séquence (pondération en fonction de la fréquence des sites dans ou en dehors des vrais promoteurs)

Prédiction de promoteurs: sensibilité, spécificité

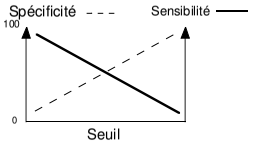
- Sensibilité: fraction des promoteurs qui sont trouvés par le logiciel

$$sensibilité = \frac{vrais_positifs}{vrais_positifs + faux_négatifs}$$

  - PromoterScan: sensibilité = 70% (promoteurs à boîte TATA)
- Spécificité: fraction des vrais promoteurs parmi ceux qui ont été prédits

$$spécificité = \frac{vrais_positifs}{vrais_positifs + faux_positifs}$$

  - PromoterScan: spécificité = 20%
  - Un faux positif / 10 kb



- Génome humain: ≈100 000 gènes, ≈1 promoteur/30 kb

Prédiction de promoteurs eucaryotes: recherches en cours

- Prise en compte de l'orientation relative et des distances entre sites de fixation de facteurs de transcription
  - COMPEL (Kolchanov 1998): banque de données d'éléments composites
  - FastM : recherche dans une séquence génomique d'une combinaison de deux sites de fixation de facteurs de transcription à une distance définie l'un de l'autre
- Recherche de corrélations entre sites
  - Prospector (Wemer 2000)
    - Sensibilité: 50%
    - Spécificité: 80%

http://www.gsf.de/biodv/index.html

- Combinaison recherche *ab initio* / approche comparative: recherche de sites potentiels parmi les régions conservées

Îlots CpG

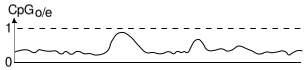
- Génome de vertébrés :
  - méthylation des C dans les dinucléotides 5 '-CG-3 '(CpG)
- Me-C fortement mutable -> T

```
5 '-CG- 3 '        5 '-TG-3 '   ou   5 '-CA-3 '
3 '-GC- 5 '        3 '-AC-5 '        3 '-GT-5 '
```

- Génome des vertébrés: globalement dépourvu en CpG (excès de TG, CA)

$$CpG_{o/e} = \frac{Nombre_de_CpG_observé}{Nombre_de_CpG_attendu} = 0.25$$

- Certaines régions (200 nt à plusieurs kb) échappent à la méthylation
  - Pas de déplétion en CpG: $CpG_{o/e}$ proche de 1
  - Riche en G+C
  - Îlot CpG:
    - Longueur > 500 nt
    - CpGo/e > 0.6
    - G+C > 50%

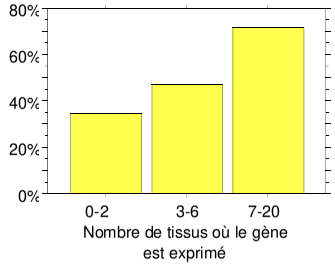


Îlots CpG: associé aux régions promotrices ?

- Bird (1986), Gardiner-Garden (1987) Larsen (1992) ref
  - 40% des gènes tissu-spécifiques possèdent un îlot CpG en 5 '
  - 100% des gènes ' housekeeping ' possèdent un îlot CpG en 5 '
- Rechercher des îlots CpG pour prédire des régions promotrices ?
  - Sensibilité: 40-100%
  - Spécificité ?? (Quelle fraction des îlots CpG correspond effectivement à des régions promotrices ?)
- Ponger (1999): comparaison des îlot CpG qui recouvre ou non le site d 'initiation de la transcription

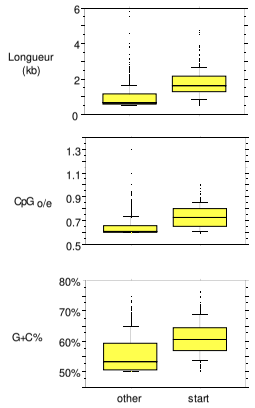
Fréquence des gènes humains avec un îlot CpG recouvrant le site d 'initiation de la transcription

- 800 gènes humains avec promoteur décrit
- Mesure de la distribution tissulaire à l 'aide d 'EST (20 tissus)



Comparaison des îlots CpG recouvrant ou non le site d 'initiation de la transcription

1. 272 îlots start CpG recouvrant le site d 'initiation de la transcription
1. 1078 îlots CpG en dehors d 'un promoteur connu



Recherche de régions régulatrices par analyse comparative (empreintes phylogénétiques)

- Goodman et al. 1988: régulation de l'expression des gènes du cluster β-globine au cours du développement
  - Alignement de séquences orthologues de 6 mammifères (> 270 Ma d'évolution)
  - 13 empreintes phylogénétiques: ≥ 6 nt, conservation 100%
  - Analyse par retard de bande sur gel:
  - 12/13 (92%) correspondent à des sites de fixation de protéines
- 1996: 35 empreintes phylogénétiques avec protéines fixatrices identifiées
- Enhancers de gènes HOX (Fugu/souris) (Aparicio et al. 1995)
- enhancer TCR α (homme/souris) (Luo, 1998)
- promoteur COX5B (11 primates) (Bachman, 1996)
- promoteur uPAR (homme/souris) (Soravia, 1995)

Large scale phylogenetic footprinting

Non-coding sequences **:** 325,247 sequences 145 Mb

everything except protein-coding regions and structural RNA genes (rRNA, tRNA, snRNA, scRNA)

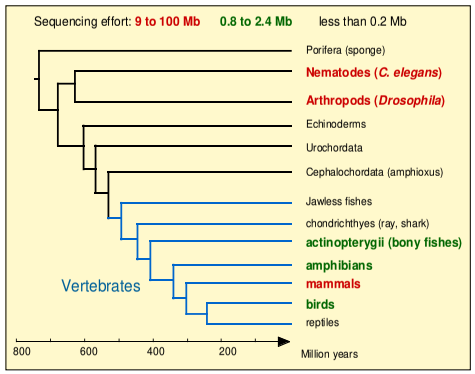
Introns, 5' and 3' untranslated regions, intergenic sequences

Filtering of microsatellite repeats and cloning vectors: **XBLAST**

Similarity search: **BLASTN + LFASTA**

Vertebrates, insects, nematode

Metazoan Genome Projects



Sequence Similarities

1- Identification of new genes

protein-genes, RNA-genes: *intronic snoRNA genes*

2- Retroviral elements, retrotransposons

3- Low complexity sequences:

GC-rich, AT-rich, cryptic microsatellites

4- Artefacts:

annotation errors, sample contamination (*sponge insulin, ascidian RNA, chicken TGFB1*)

5- 326 highly conserved regions (HCRs)

- do not code for proteins
- do not correspond to any known structural RNA

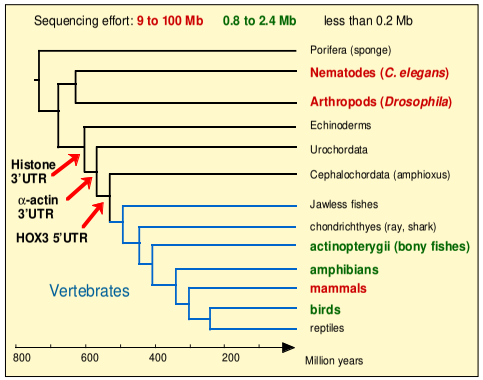
326 Highly Conserved Regions (HCRs)

- > 70% identity over 50 to 2000 nt after more than 300 Myrs
- Unique sequences
- Generally specific of only one gene
- Longest HCR:

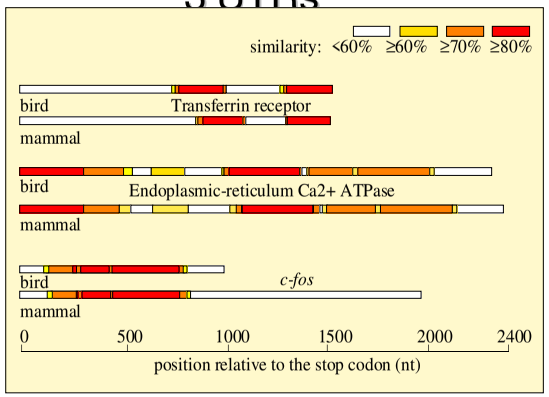
  84% identity over 1930 nt after 300 Myrs

  3'UTR deltaEF1 transcription factor
- Oldest HCRs: 500 to 600 Myrs
- No HCR between vertebrates and insects or nematode

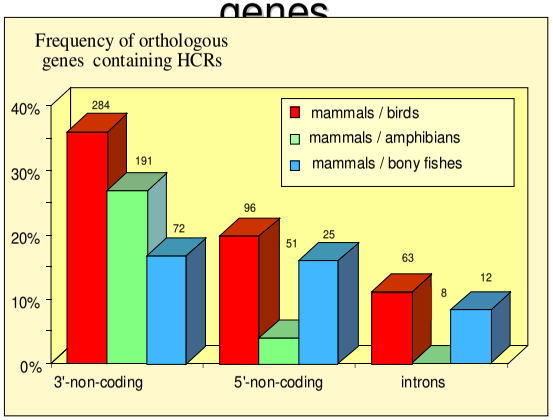
Oldest HCRs



Conservation pattern in 3'UTRs



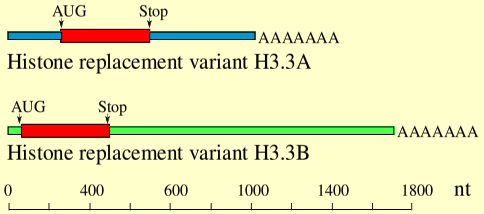
Distribution of HCRs within genes



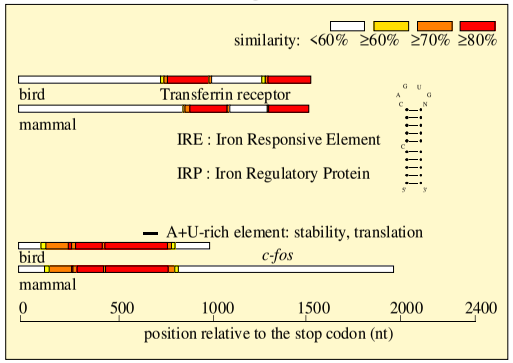
HCRs and multigenic families

Histone replacement variant H3.3A and H3.3B, Calmodulins

- several genes coding for a same protein
- non-coding sequences are distinct, and conserved

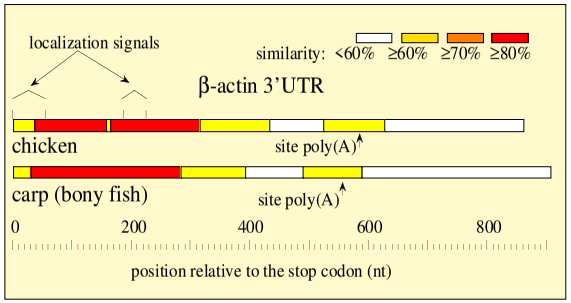


Function of 3'HCRs: mRNA stability, translation



Function of 3'HCRs: mRNA subcellular localization

Myosin heavy chain, c-myc, vimentin, β-actin



ACUTS: compilation of Ancient Conserved UnTranslated Sequences

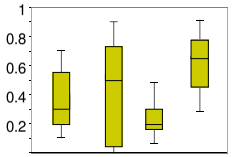
Annotated multiple alignments:

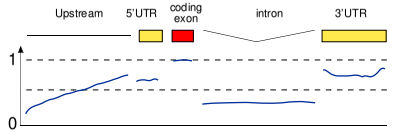
- age of the conserved element
- gene function
- function of the conserved element
- bibliographic references (MEDLINE)
- sequences available from different species (EMBL)
- description of sequence features

http://pbil.univ-lyon1.fr/acuts/ACUTS.html

Comparaison des régions non-codantes de 77 gènes orthologues homme/souris (Jareborg et al. 1999)

Fraction des régions non-codantes conservées entre homme et souris





Prédiction *ab initio* de gènes eucaryotes

- Prédiction d 'exons codants
  - Recherche de phases ouvertes de lecture (ORF: open reading frame)
    - Taille moyenne des exons: ± 150 nt
  - Statistiques sur les nucléotides, usage des codons
    - Périodicité d'ordre 3, fréquence d 'hexamères
    - Modèles de Markov cachés
  - Signaux d 'épissage
    - Profils, modèles de Markov cachés, réseau neuronaux
- Construction d 'un modèle de gène protéique
  - Combinaison d 'exons de phases compatibles (pondération en fonction des scores de chaque exon potentiel)
  - Recherche de limites de gènes
    - Exons terminaux (5 ', 3 ')
    - Promoteur
    - Signal de polyadénylation
- Epissage alternatif ?? Exons non codants ?? Gène transcrits non codants (Xist, …) ??

Prédiction de gènes eucaryotes: qualité de la prédiction

- Comparaison des différents logiciels: sensibilité/spécificité
  - Sn: sensibilité Sp: spécificité par exon (sn_e, sp_e) ou par nucéotide (sn_e, sp_e)
  - Jeu de données Burset-Guigo (1996): 570 gènes de vertébrés

	Sn_e	Sp_e	Sn_n	Sp_n
GenScan	0.78	0.81	0.93	0.93
FGENES 1.6	0.83	0.82	0.92	0.93
Grail2	0.36	0.43	0.72	0.87

  - Jeu de données Salamov et al (1998): 660 gènes humains

	Sn_e	Sp_e	Sn_n	Sp_n
GenScan	0.70	0.71	0.92	0.90
FGENES 1.6	0.77	0.77	0.90	0.91

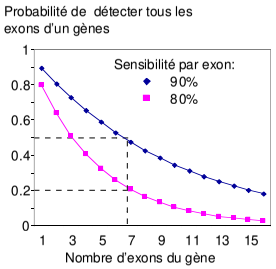
Prédiction de gènes eucaryotes: qualité de la prédiction

- Comparaison des différents logiciels: sensibilité/spécificité
  - Sn: sensibilité Sp: spécificité par exon (sn_e, sp_e) ou par nucéotide (sn_e, sp_e)
  - Locus BRCA2 (1.4 Mb, chrom. 13q) (Sanger Centre 1999): région "difficile" pour les logiciels de prédiction. 159 exons

	Sn_e	Sp_e	Sn_n	Sp_n
GenScan	0.66	0.36	0.81	0.44
FGENES 1.6	0.69	0.57	0.79	0.66
FGENES 1.6 masked	0.69	0.65	0.79	0.74
GenScan+FGENES	0.61	0.82	0.67	0.90

Prédiction de gènes protéiques complets

- *C. elegans*: la plupart des ' gènes ' annotés sont seulement des prédictions
- Peut-on utiliser ces méthodes pour annoter les séquences génomique humaines ?



  - + les faux positifs !

Un peu d ’optimisme

- Fraction de la longueur des gènes correctement prédits:

70-80%

- Probabilité que deux exons potentiels consécutifs soient réels (et donc positifs en RT-PCR)

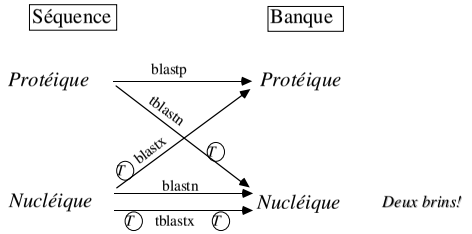
0.5

Prédiction de gènes eucaryotes (suite)

- Utilisation des EST
  - Alignement séquence génomique / cDNA (EST): SIM4 (alignement de séquences nucléiques très fortement similaires avec ‘ épissage ’)
- Approche comparative
  - Comparaison d ’une séquence génomique avec des gènes déjà caractérisés dans d ’autres espèces (WISE2: alignement ADN/protéine avec épissage)
  - Comparaison de séquences génomiques (non-annotées) homologues
    - Locus mnd2 (homme souris) (Jang et al. 1999): >80 kb
    - Prédiction d ’exons internes basée sur la conservation de séquence
      ORF ≥ 80 nt
      Séquence protéique ≥ 70% similarité
      Séquence ADN ≥50% identité
      GT AG conservés
      => détection de tous les exons internes du gène D6Mm5e
    - Généralisation de la méthode (Guigo 2000). Sensibilité ? Spécificité ?

Stratégies de recherche de similarités: ADN ou protéine ?

- Limites des recherches de similarité au niveau ADN
  - Alphabet réduit (4 lettres)
  - Dégénérescence du code génétique
- Mais … tout n'est pas codant
  - régions régulatrices, ARN structuraux, ...



Traitement du bruit de fond: filtres et masques

- *Séquences de faible complexité (protéines, ADN):*

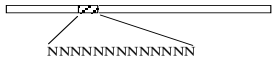
40% des protéines — ADN: microsatellites
15% du total des résidus — exemple: CACACACACACACACA
Ala, Gly, Pro, Ser, Glu, Gln
logiciels de filtrage: SEG, XNU, DUST

```
RSPPR--KPQGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPP
. :::    : :: : :  ::::: :  :: :.:  :: : :::::
QGPPRPGNQQCPPPQGG--PQGPPRP--GNQQRP--PPQGGPQGPP
```

(filtré par défaut par BLAST)

- *Séquences abondantes*

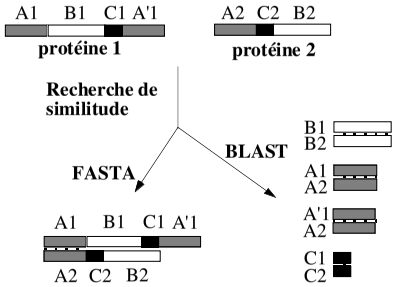
3000 Immunoglobulines dans GenBank
10^6 Alu, 10^5 L1 dans le génome humain
logiciels de masquage: XBLAST, RepeatMasker



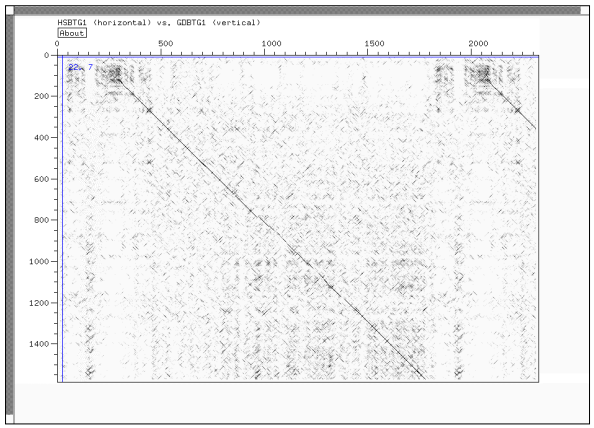
Quelle approche adopter ?

- Recherche rapide de similarité dans les banques
  - algorithme (BLAST, FASTA)
  - matrices de substitution, pondération des gaps
  - stratégie de recherche (nucléique, protéique)
  - traitement du bruit de fond
  - complétude des banques de données, répéter la recherche régulièrement
- Comparaison d'une séquence génomique à un cDNA (EST)
  - BLASTN < LFASTA << SIM4 (épissage)
- Comparaison d'une séquence génomique à une protéine
  - BLASTX, FASTX << WISE2 (épissage)
- Comparaison de deux séquences génomiques
  - ADN: BLASTN < LFASTA << Smith-Waterman (SIM, LALIGN)
  - Protéique: TBLASTX

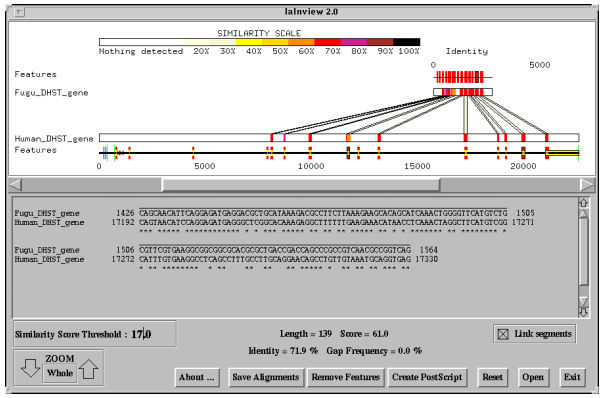
Alignement par bloc ou alignement global : comparaison BLAST / FASTA



Représentation graphique des similarités entre séquences: DOTTER



Représentation graphique des similarités entre séquences: LALNVIEW



Représentation graphique des similarités entre séquences: PIP-maker

Limitation des comparaisons deux à deux (BLAST, FASTA, ...)

```
Seq A       CGRRLILFMLATCGECDTDSSE … HICCIKQCDVQDIIRVCC
            ::  :       :::           ::   :       :
Insuline    CGSHLVEALYLVCGERGFFYTP … EQCCTSICSLYQLENYCN
              :::   :     :  :           ::  :  :
Seq B       YQSHLLIVLLAITLECFFSDRK … KRQWISIFDLQTLRPMTA
```

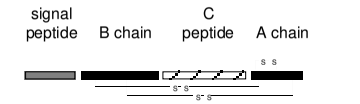
Comparaisons 2 à 2:

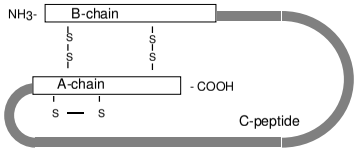
Insuline / Seq A : 25% d'identité
Insuline / Seq B : 25% d'identité

Alignement de séquences de la famille des insulines

```
                                B-chain                                A-chain
INSL4     Q14641  ELRGCGPRFGKHLLSYCPMPEKTFTTTPGG...[x]58 ....SGRHRFDPFCCEVICDDGTSVKLCT
INSL3     P51460  REKLCGHHFVRALVRVCGGPRWSTEA.....[x]51 ....AAATNPARYCCLSGCTQQDLLTLCPY
RLN1      P04808  VIKLCGRELVRAQIAICGMSTWS.....[x]109 ....PYVALFEKCCLIGCTKRSLAKYC
BBXA      P26732  VHTYCGRHLARTLADLCWEAGVD.....[x]25 ........GIVDECCLRPCSVDVLLSYC
BBXB      P26733  ARTYCGRHLADTLADLCF--GVE.....[x]23 ........GVVDECCFRPCTLDVLLSYCG
BBXC      P26735  SQFYCGDFLARTMSILCWPDMP.....[x]25 ........GIVDECCYRPCTTDVLKLYCDKQI
BBXD      P26736  GHIYCGRYLAYKMADLCWRAGFE.....[x]25 ........GIADECCLQPCTNDVLLSYC
LIRP      P15131  VARYCGEKLSNALKLVCRGNYNTMF.....[x]58 ........GVFDECCRKSCSISELQTYCGRR
MIP I     P07223  RRGVCGSALADLVDFACSSSNQPAMV.....[x]29 ....QGTTNIVCECCMKPCTLSELRQYCP
MIP II    P25289  PRGICGSNLAGFRAFICSNQNSPSMV.....[x]44 ....QRTTNLVCECCFNYCTPDVVRKYCY
MIP III   P80090  PRGLCGSTLANMVQWLCSTYTTSSKV.....[x]30 ....ESRPSIVCECCFNQCTVQELLAYC
MIP V     P31241  PRGICGSDLADLRAFICSRRNQPAMV.....[x]44 ....QRTTNLVCECCYNVCTVDVFYEYCY
MIP VII   P91797  PRGLCGNRLARAHANLCFLLRNTYPDIFPR...[x]86 ..EVMAEPSLVCDCCYNECSVRKLATYC
ILP       P22334  AEYLCGSTLADVLSFVCGNRGYNSQP.....[x]31 ........GLVEECCYNVCDYSQLESYCNPYS
INS       P01308  NQHLCGSHLVEALYLVCGERGFFYTPKT.....[x]35 ........GIVEQCCTSICSLYQLENYCN
IGF1      P01343  PETLCGAELVDALQFVCGDRGFYF.....[x]12 ........GIVDECCFRSCDLRRLEMYCAPLK
IGF2      P01344  SETLCGGELVDTLQFVCGDRGFYF.....[x]12 ........GIVEECCFRSCDLALLETYCATPA
                     *.        .*                              **   *    .   *
```

Structure des insulines



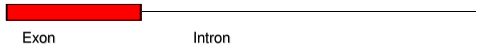


Représentation d 'un motif conservé par une *matrice de fréquences (profil)*

- Exemple: site donneur d 'épissage (vertébrés)

Matrice de fréquence (pourcentage):

Base	Position								
	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**33**	**60**	8	0	0	**49**	**71**	6	15
C	**37**	13	4	0	0	3	7	5	19
G	18	14	**81**	**100**	0	**45**	12	**84**	20
T	12	13	7	0	**100**	3	9	5	**46**
Cons.	M	A	G	G	T	R	A	G	T



Recherche d 'un motif dans une séquence à l 'aide d 'un profil

- Calcul des scores de similarité en faisant glisser une fenêtre de la longueur du motif le long de la séquence. Exemple:

```
A    33   60    8    0    0   49   71    6   15
C    37   13    4    0    0    3    7    5   19
G    18   14   81  100    0   45   12   84   20
T    12   13    7    0  100    3    9    5   46
```

```
GAAAGGTGAGTCAT...
GAAAGGTGA          S=18+60+8+0+0+45+9+84+15=239
.AAAGGTGAG         S=33+60+8+100+0+3+12+6+20=242
..AAGGTGAGT        S=33+60+81+100+100+45+71+84+46=620
...AGGTGAGTC       S=33+14+81+0+0+49+12+5+19=213
....GGTGAGTCA      …etc
```

PSI-BLAST

- Position-Specific Iterated BLAST
  - 1-recherche BLAST classique (protéine)
  - 2-construction d'une matrice de pondération (profil) avec les séquences similaires détectées
  - 3-recherche BLAST à partir de ce nouveau profil
  - 4-itération des étapes 2-3 jusqu'à convergence
- plus sensible que Smith-Waterman
- 40 fois plus rapide

Comparaison d 'une séquence génomique à un motif ou à une banque de motifs

- Banques de données de motifs protéiques
  - PROSITE
  - BLOCKS
  - PFAM
  - PRODOM
  - PRINTS
- Logiciels de comparaison d 'une séquence à une banque de motifs, ou de comparaison d 'un motif à une banque de séquences
  - Pfscan: profils
  - WISE2