DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR

UNIVERSITAT DE BARCELONA

PROGRAMA DE DOCTORAT DE BIOTECNOLOGIA

BIENNI 2002-2004

CARACTERITZACIÓ BIOINFORMÀTICA DE LA CONTRIBUCIÓ DE L'*SPLICING* ALTERNATIU A LA VARIABILITAT DEL PROTEOMA

Tesi realitzada pel llicenciat en Biologia David Talavera i Baró sota la direcció dels Drs. Modesto Orozco López i Xavier de la Cruz Montserrat per optar al títol de doctor per la Universitat de Barcelona.

David Talavera i Baró Xavier de la Cruz Montserrat Modesto Orozco López

BARCELONA, 2007

Als que m'estimo

Agraïments

Primer de tot, vull donar les gràcies al Xavier de la Cruz i al Modesto Orozco, que han estat els meus directors de tesi i que m'han acollit al seu laboratori. D'ells he intentat encomanar-me la passió per la ciència i la feina que fem i espero haver-ne aprés el rigor científic.

Al Josep Lluís i els Ivans, per fer que durant aquests quasi cinc anys, les màquines funcionessin.

A l'Adam, per solucionar-me tots els dubtes informàtics amb una immensa paciència, per les converses tot dinant, per ensenyar-me casa seva... No canviïs, xaval!

Als que vau començar al mateix temps que jo, donar-vos ànims. Abans de que us n'adoneu ja haureu enllestit tota la feina.

Als bioinformàtics, quants *Journal Clubs*! Quantes reunions al despatx del Xavier! No defalliu que el premi es a tocar.

A la resta del grup MMB, moltes gràcies a tots, els que us quedeu i els que ja heu marxat, els de Farmàcia i els del BSC.

A la Marta Closa, pels nostres dinars i cafès, sempre amanits de llargues converses. Quants cops no m'hauré sentit molt més optimista després de passar una estona entretinguda xerrant amb tu!

Als companys de la Facultat, encara que ens vegem poc, sabeu que us duc ben endins.

Als amics del poble, companys de jocs, festa i altres aventures, que ens han permés compartir-ho tot alhora que creixíem.

A la Sara, per suportar-me tal com sóc, per amagar la pròpia por i fer-me sentir segur.

I, finalment, a la meva família, per ser allà on són, per donar-me tantes coses i no demanar res a canvi, per les hores treballades i les que hem rigut junts, perquè sé com n'estaran de feliços.

Barbens/Barcelona, maig de 2007

CONTINGUT

Índex

Índex de figures

Índex de taules

ABREVIATURES

Abreviatures

AS	*Splicing* Alternatiu.
ATP	Adenosina Trifosfat
BLAST	*Basic Local Alignment Search Tool*
BLOSUM	*Blocks of Amino Acid Substitution Matrix*
BP	*Branch Point*
CDD	*Conserved Domains Database*
DNA	Àcid Desoxiribonucleic
ESE	*Exonic Splicing Enhancer*
ESS	*Exonic Splicing Silencer*
EST	*Expressed Sequence Tag*
FGF	*Fibroblast Growth Factor*
FGFR	*Fibroblast Growth Factor Receptor*
GD	Duplicació Gènica
hnRNP	*Heterogeneous ribonucleoprotein*
IG	Identitat Global
IL	Identitat Local
ISE	*Intronic Splicing Enhancer*
ISS	*Intronic Splicing Silencer*
KS	Test de Kolmogorov-Smirnov
mRNA	RNA missatger
NMD	*Non-sense Mediated Decay*
PAM	*Point Accepted Mutation*
PCR	Reacció en Cadena de la Polimerasa
PDB	*Protein Data Bank*

PSM	*Prostate-Specific Membrane*
RMA	*Robust Multiarray Averaging*
RNA	Àcid Ribonucleic
SNP	*Single Nucleotide Polymorphism*
snRNP	*Small Nuclear Ribonucleoprotein*
SPLASH	*Search at Protein Level of Alternative Splicing Homologs*
SR	*Serine/Arginine-rich proteins*
SRS	*Sequence Retrieval System*

INTRODUCCIÓ

1 Introducció

L'*splicing* alternatiu és el mecanisme que permet a un gen la codificació de diversos productes proteics. Mitjançant l'ús de diversos patrons d'*splicing*, és a dir, tallant i empalmant els exons de maneres diferents, a partir d'un gen es generen múltiples RNAs missatgers, que, quan es tradueixen, generen diferents productes proteics –isoformes de la mateixa proteïna.

1.1 Importància biològica de l'splicing alternatiu

Des del descobriment de l'estructura gènica dividida en exons i introns en el gen *hexon* d'adenovirus (Berget and Sharp, 1977; Chow et al., 1977) es va començar a postular l'existència de l'*splicing* alternatiu i les seves possibles implicacions en l'evolució de la funció gènica (Gilbert, 1978; Mount and Steitz, 1983). No obstant això, no ha estat fins que s'han encarat els projectes de seqüenciació massiva del genoma (IHGSC, 2001; Venter, 2001) que l'*splicing* alternatiu ha pres protagonisme. Ara mateix es pensa que l'*splicing* alternatiu és un mecanisme que participa en el control de l'expressió gènica dels organismes eucariotes.

1.1.1 Freqüència de l'*splicing* alternatiu

La Taula 1 mostra l'evolució que han tingut les estimacions sobre la quantitat d'*splicing* alternatiu en els genomes. Si fins a finals dels anys 90 es pensava que aquest era un fenomen molt rar que afectava pocs gens (Sharp, 1994), després les estimacions han anat augmentant, situant-se alguns cops per sobre del 50% dels gens (IHGSC, 2001; Kan et al., 2001) i arribant fins a les tres quartes parts o més dels gens multi-exònics (Johnson, 2003; Kampa et al., 2004). Les estimacions actuals més conservadores donen valors al voltant del 40% (Brett et al., 2000; Mironov et al., 1999; Modrek et al., 2001; Stolc et al., 2004). No obstant això, sembla que en la majoria de casos hi pot haver subestimacions del percentatge d'*splicing* alternatiu, ja que aquest depén de la quantitat d'ESTs analitzats (Brett et al., 2002).

Any	Estimació	Referència
1994	5%	(Sharp, 1994)
1999	35%	(Mironov et al., 1999)
2000	38%	(Brett et al., 2000)
2000	22%	(Croft et al., 2000)
2001	59%	(IHGSC, 2001)
2001	55%	(Kan et al., 2001)
2001	42%	(Modrek et al., 2001)
2003	74%	(Johnson, 2003)
2004	40%	(Stolc et al., 2004)
2004	88%	(Kampa et al., 2004)

Taula 1. Percentatge de gens amb *splicing* alternatiu. La taula mostra l'evolució de les estimacions de gens amb *splicing* alternatiu.

1.1.2 Especificitat tissular i de desenvolupament

S'ha proposat que l'*splicing* alternatiu té dues funcions: augmentar la diversitat del proteoma i ser un altre nivell de regulació en l'expressió gènica dels teixits sans, fent que cada teixit expressi les isoformes adequades a cada estadi de desenvolupament (Pan et al., 2004). A més, hi ha teixits que requereixen més plasticitat que d'altres, és a dir, per funcionar correctament necessiten expressar més diversitat de productes peptídics, ja sigui perquè estan formats per diversos tipus de cèl·lules o perquè les proteïnes han de fer diverses funcions (Blencowe, 2006). Com a exemples d'aquest punt tenim els sistemes nerviós i immunològic que semblen tenir un nivell d'*splicing* alternatiu superior a la majoria de teixits (Grabowski and Black, 2001; Modrek et al., 2001).

S'han fet diferents projectes per estudiar l'especificitat tissular. En un d'ells s'arribà a la conclusió que almenys entre un 10 i un 30% dels gens tenien alguna isoforma amb especificitat tissular (Xu et al., 2002). En un altre estudi es trobà que el sistema nerviós

central, el fetge i els testicles eren els teixits amb més isoformes específiques (Yeo et al., 2004). Tanmateix, poques isoformes tindrien especificitat tissular absoluta –en un sol teixit (Xie et al., 2002).

Existeix un control combinatori, és a dir, xarxes de processos que controlen l'*splicing* alternatiu i que permeten l'elecció de llocs d'*splicing* en base a factors reguladors específics de cada teixit (Blencowe, 2006; Smith and Valcarcel, 2000) i seqüències diana també específiques (Sugnet, 2006). Un exemple d'això seria la proteïna Nova-1 que és un regulador d'*splicing* específic del teixit neuronal que s'uneix a seqüències intròniques específiques (Jensen et al., 2000) i controla l'expressió de tota una xarxa de proteïnes relacionades amb la inhibició de la sinapsi (Ule et al., 2003).

Curiosament, s'ha vist que els gens que tenen un *splicing* alternatiu amb regulació tissular específica no es corresponen amb gens que tenen una transcripció regulada a nivell tissular. (Blencowe, 2006).

Un cas molt interessant de l'especificitat de les isoformes en el desenvolupament és la determinació sexual de *Drosophila melanogaster* (Black, 2003) (veure la Figura 1 per un esquema). La diferenciació entre mascles i femelles es fa a través d'una cascada d'activació de gens al capdamunt de la qual hi ha *sex lethal* (s*xl*). s*xl* té un *splicing* alternatiu específic de gènere i a partir d'aquí la forma activa –la femenina- reprimeix el desenvolupament de les característiques masculines, actuant com un factor regulador de l'*splicing*. Els seus gens diana són *transformer* (*tra*) i *male-specific lethal 2* (*msl2*) –què també tenen *splicing* específic de génere- i ell mateix –per mantenir la dosi en femelles. Addicionalment, Tra també és un regulador de l'*splicing* que provoca *splicings* alternatius per mascles i femelles en *doublesex* (*dsx*) i *fruitless* (*fru*). Doublesex és el responsable de la diferenciació diferencial de les cèl·lules somàtiques de mascles i femelles, mentre Fruitless té a veure amb el comportament dels mascles (Lopez, 1998).

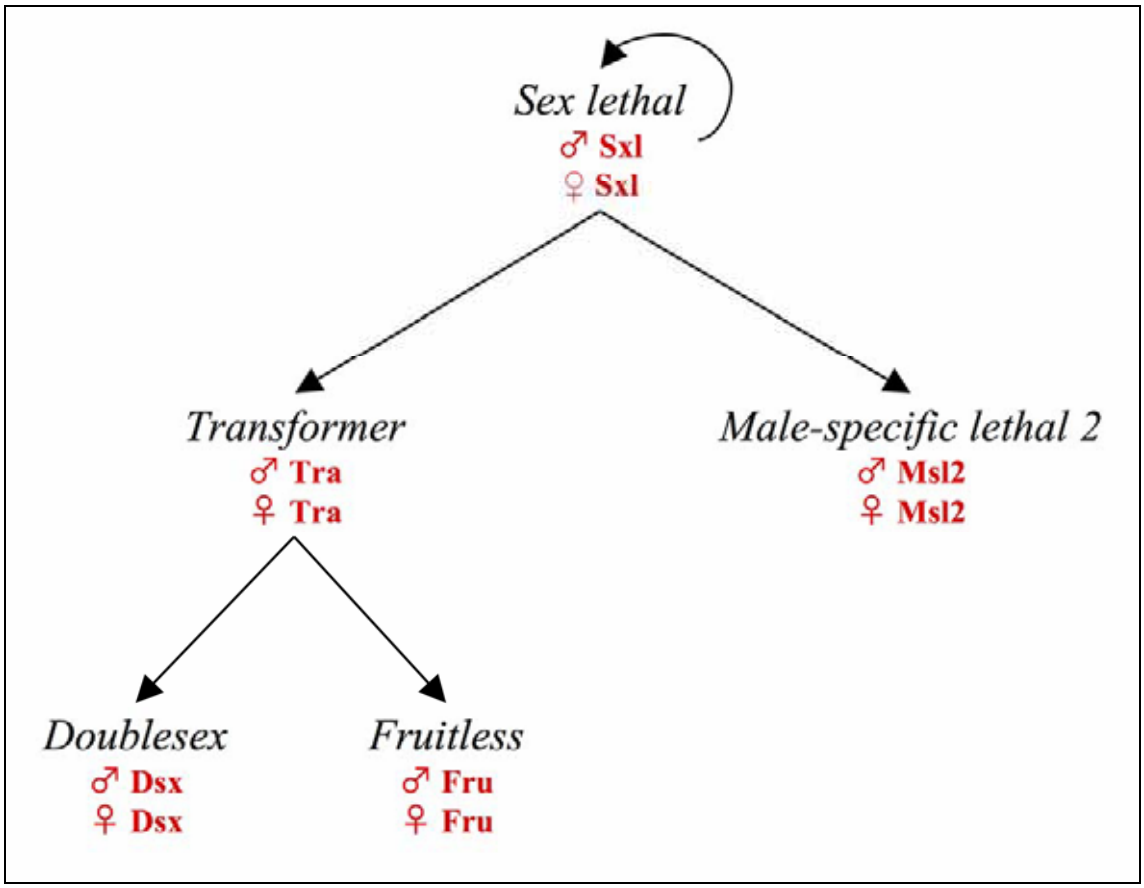

Figura 1. Esquema de la determinació sexual a *D. melanogaster*. En negre hi ha els gens i en roig, les proteïnes. Les fletxes indiquen la regulació.

Finalment, cal remarcar que no tots els *splicings* específics de teixit tenen una funcionalitat tan clara. Així, per exemple, s'han trobat RNAs que pateixen *splicing* alternatiu amb especificitat tissular i, en canvi, no són codificants (Ravasi et al., 2006).

1.1.3 Interacció amb altres processos i regulació

S'ha vist que l'*splicing* en general i l'*splicing* alternatiu en particular tenen relació amb molts altres processos, que poden arribar a regular-lo o influir en la viabilitat dels productes proteics.

A l'hora de l'elecció dels llocs d'*splicing* es pensava que funcionava el mecanisme "el primer que arriba, és servit el primer", és a dir, l'spliceosoma actuava a mesura que trobava els llocs d'*splicing*. No obstant això, ara se sap que el procés és més complex, intervenint-hi diversos factors: la maquinària de transcripció, les modificacions post-transcripcionals, l'estructura secundària de l'RNA i l'estructura de la cromatina. Així, s'ha vist que un nivell de regulació de l'*splicing* alternatiu el controlarien proteïnes

implicades en l'estructura de la cromatina (Batsche et al., 2006; Kornblihtt, 2006). Nogués i col·laboradors mostraren com l'acció d'un inhibidor de la deacetilació de les histones portava una baixa taxa d'inclusió d'exons alternatius (Nogues et al., 2002) i en un altre treball del mateix grup veieren que una alta taxa de replicació afavoria la inclusió d'aquests exons (Kadener et al., 2001). Finalment, Batsché i col·laboradors han vist com el factor modificador de la cromatina SWI/SNF, canvia el patró de fosforilació de l'RNA polimerasa II, n'alenteix el ritme i facilita la inclusió d'exons alternatius (Batsche et al., 2006). D'aquesta manera, el ritme amb el que avança l'RNA polimerasa II és molt important per determinar la inclusió o omissió dels exons (Kadener et al., 2001); per tant, els factors de transcripció que regulen l'inici o l'elongació del transcrit influeixen en la selecció dels llocs d'*splicing* (Kornblihtt, 2006). A més, s'ha vist que l'edició de l'RNA pot comportar el canvi dels llocs d'*splicing* (Laurencikiene et al., 2006). Finalment, també feia anys que se sabia que l'estructura secundària que adopta el pre-mRNA en el moment de la transcripció pot arribar a interferir l'spliceosoma (Sirand-Pugnet et al., 1995), a més, en els darrers anys s'han anat succeint els treballs experimentals en que es troba una relació entre l'*splicing* alternatiu i l'estructura secundària del pre-mRNA (Graveley, 2005; Hefferon et al., 2004); tot això ha fet pensar que d'alguna manera l'estructura secundària del pre-mRNA exerceix un paper regulador de l'*splicing* (Buratti and Baralle, 2004; Lian and Garner, 2005; Meyer and Miklos, 2005).

Un efecte modulador important que rep l'*splicing* és el mecanisme d'eliminació de transcrits defectuosos (NMD per *non-sense RNA mediated decay*), que elimina els transcrits que tenen codons terminadors prematurs (Lewis et al., 2003). S'ha vist que una part significativa dels transcrits alternatius són degradats abans de traduir-se, així produeixen una quantitat de proteïna menor que la que els tocaria per nivell d'expressió (Lewis et al., 2003). En realitat, encara no se sap massa quina seria la finalitat d'aquesta estratègia, però podria ser que l'NMD tingués un rol actiu en el control de l'expressió gènica (Green et al., 2003). No obstant això, s'ha vist que una quarta part dels esdeveniments d'*splicing* alternatiu conservats entre humans i ratolins podrien ser dianes de l'NMD (Baek and Green, 2005) i anàlisis funcionals més recents semblen indicar que els transcrits amb terminadors prematurs s'expressen en nivells molt baixos i en la majoria de teixits independentment de l'acció de l'NMD (Pan et al., 2006) i que, per tant, tenen alguna funció biològica (Neu-Yilik et al., 2004). Tanmateix, això no en

demostra la importància funcional (Xu et al., 2002).

1.1.4 Relació amb malalties

La relació de l'*splicing* alternatiu amb les malalties es pot donar per dues vies diferents: canvis patològics en el patró d'expressió de les isoformes i expressió d'isoformes aberrants, és a dir, expressió d'isoformes normals en un estadi de desenvolupament o teixit incorrecte i expressió d'isoformes anormals obtingudes a partir d'errors en l'*splicing*.

La majoria de canvis en el patró d'*splicing* són naturals i de resposta a senyals externes (Stamm, 2002), però n'hi ha d'altres que són anormals i donen com a resultat isoformes patològiques (Grabowski and Black, 2001), ja que les diferents isoformes poden tenir rols molt diferents i, per tant, els errors en l'expressió de les isoformes poden comportar fenotips patològics. Un exemple d'això és el cas de Bcl-x. Aquesta proteïna té dues isoformes produïdes per l'ús d'un donador d'*splicing* alternatiu –Bcl-x_L i Bcl-x_S. Mentre Bcl-x_L, la isoforma llarga, és antiapoptòtica, la isoforma curta és proapoptòtica. Així la viabilitat de les cèl·lules depèn de la proporció en que s'expressen aquestes isoformes (Rohrbach et al., 2005; Taylor et al., 1999).

D'altra banda, els *splicings* aberrants s'han relacionat amb moltes malalties (Caceres and Kornblihtt, 2002; Stoilov et al., 2002; Venables, 2004) i amb l'envelliment (Meshorer and Soreq, 2002). Així, s'ha vist que moltes mutacions causen malalties perquè promouen l'omissió d'un exó sencer o activen un lloc d'*splicing* críptic (Blencowe, 2006), provocant la síntesi d'isoformes patològiques. D'acord amb una hipòtesi recent, les mutacions que afecten l'*splicing* arribarien a ser el 60% de les mutacions del genoma humà que causen algun tipus de malaltia, convertint-se en la causa més freqüent de les malalties hereditàries (Lopez-Bigas et al., 2005). Els autors d'aquest estudi mostren que la longitud de la seqüència codificant i el nombre d'introns tenen una alta correlació amb la propensitat dels gens a ser causants de malalties (Lopez-Bigas et al., 2005). Segons Xing i Lee (Xing and Lee, 2006), hi ha diverses evidències que fan que l'anterior hipòtesi s'hagi de tenir en compte: primer, l'*splicing* està associat amb un pressió selectiva negativa (Bustamante et al., 2002; Lopez-Bigas et al., 2005); segon, l'*splicing* alternatiu sembla augmentar aquesta selecció negativa, afegint més restriccions (Xing and Lee, 2005); tercer, la predicció que els gens amb més exons i, en particular, aquells amb més variants d'*splicing* tinguin un risc més alt de

patir mutacions a la zona d'*splicing* causants de malalties casa amb l'experiència (Buee et al., 2000; Eng, 2004).

1.1.4.1 *Splicing* alternatiu i càncer

La relació entre l'*splicing* alternatiu i el càncer es dóna tant per l'aparició de mutacions en un gen que n'afecten l'*splicing* (Srebrow and Kornblihtt, 2006; Venables, 2004), com per defectes en els factors que el regulen (Srebrow and Kornblihtt, 2006; Venables, 2006). Entre els primers, trobaríem els gens *KIT* (Chen, 2005) i *LKB1* (Hastings et al., 2005) i, pel que fa als segons, un bon exemple seria el gen *RON* que té unes isoformes amb poder transformant (Zhou et al., 2003b) que s'expressen a causa d'una desregulació de l'*splicing* provocada per la sobreexpressió del factor SF2/ASF (Ghigna et al., 2005).

Tot i que es troben diferències pel que fa a les isoformes expressades (Kalnina et al., 2005; Roy et al., 2005) o en l'aparició d'*splicings* aberrants (Wang et al., 2003b), s'ha vist que les correlacions només existeixen amb tipus específics de càncers (Kirschbaum-Slager et al., 2004; Kirschbaum-Slager et al., 2005; Okumura et al., 2005). Per tant, no és possible trobar en l'*splicing* una causa única que ens permeti explicar el desenvolupament del càncer, sinó que sembla que cada càncer té els seus problemes d'*splicing* específics (Venables, 2006)

1.1.4.2 *Splicing* alternatiu i sistema nerviós

El sistema nerviós sembla tenir un nivell d'*splicing* alternatiu superior a la majoria de teixits (Grabowski and Black, 2001; Modrek et al., 2001). A més a més, també sembla que seria el teixit amb més isoformes específiques (Xu et al., 2002). S'han trobat diversos gens implicats en malalties del sistema nerviós per culpa d'isoformes patològiques. La llista inclou malalties neuromotores (Aerbajinai et al., 2002; Lorson et al., 1999; Munch et al., 2002; Robertson et al., 2003), síndromes neuromusculars (Brenner et al., 2003), desordres mentals associats a l'estrès (Abdel-Rahman et al., 2002; Cohen et al., 2002) i malalties neurodegeneratives (Benmoyal-Segal et al., 2005; Inestrosa et al., 1996) (Buee et al., 2000; Frederikse and Ren, 2002; Goedert et al., 1988; Hutton et al., 1998; Isoe-Wada et al., 1999; Sato et al., 1999).

1.1.5 *Splicing* alternatiu en terapèutica

La medicina i la farmàcia també han estat atentes al gran coneixement sobre l'*splicing* alternatiu que s'ha generat en els darrers anys. D'aquesta manera s'han desenvolupat

diferents aplicacions, entre les que cal destacar l'ús de les isoformes d'*splicing* alternatiu com a biomarcadors (Brinkman, 2004) en el diagnòstic de les malalties o com a diana terapèutica (Bracco and Kearsey, 2003; Hagiwara, 2005) en el seu tractament.

1.1.5.1 Variants d'*splicing* en diagnosi

L'ús com a biomarcadors es basa en l'expressió diferencial de les isoformes: en el cas del gen *PAX5*, s'ha vist que els malalts de limfoma només expressen una isoforma, mentre els individus sans n'expressen diverses (Robichaud et al., 2004); la proporció d'isoformes d'acetilcolinesterasa permet controlar l'evolució del tractament dels malalts d'Alzheimer (Darreh-Shori et al., 2004); s'han trobat diferències en les isoformes que s'expressen en les cel·lules normals i en el càncer de pit (Mills, 2005). Així, sembla clar que l'ús de microxips amb sondes específiques per a isoformes poden permetre fer diferents diagnòstics (Wang et al., 2003b; Yeakley et al., 2002).

D'altra banda, si no es té en compte la seva existència, l'*splicing* alternatiu també pot ser un problema a l'hora de fer una diagnosi. Per exemple, per diagnosticar el càncer de pròstata s'utilitza la PCR en temps real per detectar l'antigen PSM en sang, però es va veure que hi havia un cert nombre de falsos positius a causa de la detecció d'una altra isoforma. L'ús de primers específics de variants d'*splicing* ajudarà, doncs, a millorar el diagnòstic (Hisatomi et al., 2002).

1.1.5.2 Variants d'*splicing* en tractaments

La farmacogenòmica –l'estudi del genoma i els seus productes per dissenyar fàrmacs- ha d'estar amatent als possibles casos d'*splicing* alternatiu, perquè aquest pot afectar a l'eficàcia dels fàrmacs o a la seva toxicitat (Bracco and Kearsey, 2003). Per exemple, molts problemes de toxicitat dels fàrmacs són a causa de l'*splicing* alternatiu en membres de la família del citocrom P450 (Hanioka et al., 1990), els quals s'encarreguen del metabolisme dels fàrmacs.

Un altre punt molt important és l'elecció de la isoforma diana. Per exemple, les isoformes de *VEGF* tenen diferències pel que fa a la seva expressió, funció biològica i rol en el desenvolupament del càncer. Zhang i col·laboradors van proposar l'ús de tècniques d'RNA *interference* per actuar sobre les isoformes implicades en l'angiogènesi dels tumors (Zhang et al., 2003). Així mateix, influir en la proporció d'isoformes de *BCL2* es pot utilitzar com a tractament per regular l'apoptosi, però equivocar-se d'isoforma pot ser contraproduent (Akgul et al., 2004). Per exemple, en el

cas de *p53*, l'alteració de les proporcions entre les isoformes pot portar un augment de la predisposició al càncer o de l'envelliment de les cèl·lules (Mills, 2005).

1.2 Mecanismes a nivell d'àcids nucleics

Els gens eucariotes tenen distribuïda la informació codificant per les proteïnes de forma discontínua, seguint un patró d'exons i introns. Els exons contenen la informació que es traduirà en les proteïnes, mentre que els introns, que són molt més grans, no són codificants però poden contenir seqüències reguladores. A l'hora de la transcripció gènica, tant els exons com els introns són inclosos en el pre-mRNA.

L'existència de l'*splicing* alternatiu està íntimament lligada a l'estructura dels gens, és a dir, al seu patró d'exons i introns i a la forma de processar-lo. En efecte, els introns s'han d'eliminar, empalmant els exons entre ells, en el procés anomenat *splicing*. Aquest procés és realitzat de manera co-transcripcional (Goldstrohm et al., 2001) per un complex nucleoproteic anomenat spliceosoma –probablement, un antic ribozima (Roberts and Smith, 2002)- que realitza dues transesterificacions successives (Lopez, 1998). A més de la relació espacio-temporal entre els dos processos –transcripció i *splicing*-, s'ha suggerit que hi ha un control comú de les dues maquinàries (Auboeuf et al., 2002) i, fins i tot, una relació funcional entre l'RNA polimerasa II i la maquinària d'*splicing* (Hicks et al., 2006). En les següents subseccions es tracta, amb més detall, la relació existent entre l'estructura gènica i l'*splicing* alternatiu.

1.2.1 Estructura gènica i *splicing*

Els diferents organismes no tenen una estructura gènica comuna. Així, mentre els llevats tenen pocs introns i petits, els gens humans tenen, en promig, entre vuit i nou introns (Blencowe, 2006; Wong et al., 2001). D'altra banda, Clark i Thanaraj van veure que la mida dels introns depèn del contingut en guanines i citosines (Clark and Thanaraj, 2002).

Els gens tenen diversos tipus d'exons, depenent de la freqüència d'aparició en els RNAs missatgers. La majoria d'exons són constitutius –sempre apareixen a l'RNA-, mentre els alternatius es divideixen en majoritaris i minoritaris –aquests últims rarament són inclosos al transcrit (Modrek and Lee, 2003). Un fet molt important a l'hora de relacionar l'estructura gènica i l'*splicing* alternatiu és que una tercera part dels exons poden ser traduïts en més d'una pauta sense trobar senyals d'acabament –normalment,

són exons curts i rics en guanina i citosina (Clark and Thanaraj, 2002). Llavors, qualsevol canvi en el patró d'*splicing* pot comportar un canvi en la traducció dels exons posteriors (Clark and Thanaraj, 2002; Matlin et al., 2005)

Tant els exons com els introns tenen uns elements conservats que en permeten el reconeixement per la maquinària d'*splicing*. Així, la majoria de juntures d'*splicing* –els llocs donadors i acceptors d'*splicing*- estan definides per dues parelles de dinucleòtides que marquen els extrems dels introns –GU i AG-, encara que també existeixen altres possibilitats, essent GC i AG la més corrent entre aquestes (Burset et al., 2000). Per la seva banda, els introns venen definits per uns llocs d'*splicing* a 5' i 3', un nus (BP per *branch point*) i un tram de pirimidines (Goldstrohm et al., 2001) (veure la Figura 2 per un esquema). El lloc d'*splicing* 5' (donador) consens seria $^{-2}$AG↓GURAGU^{+6} –on R és una purina i els nombres indiquen les posicions respecte el lloc d'*splicing* representat amb la fletxa. El lloc acceptor d'*splicing* consens seria $^{-4}$NYAG↓RN^{+2} –on N és qualsevol nucleòtid, Y és una pirimidina i els nombres indiquen les posicions respecte el lloc d'*splicing* representat amb la fletxa. La seqüència consens del nus seria CURAY –on l'adenina que hi ha entremig és el lloc on l'extrem 5' de l'intró s'uneix. Addicionalmnet, Hiller i col·laboradors van trobar que el 30% dels acceptors d'*splicing* de gens humans tenen la seqüència NAGNAG (on N pot ser A, T, C o G) (Hiller et al., 2004), creant acceptors en tàndem, que s'han trobat experimentalment en diverses espècies (Ferranti et al., 1999; Li and Howe, 2001; Rogina and Upholt, 1995). En aquests casos, la primera part del tàndem sempre queda retinguda a l'intró, mentre la segona pot ser retinguda a l'intró o incorporada a l'exó, obtenint-se així diferents isoformes que difereixen per un residu. En el mateix treball, trobaren que almenys un 5% dels gens humans tenen *splicing* alternatiu per aquesta raó (Hiller et al., 2004).

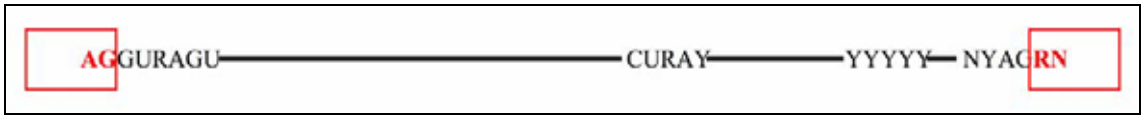

Figura 2. Esquema dels principals elements involucrats en l'*splicing*. A la figura, apareixen els dos llocs d'*splicing*, el nus i el tram de pirimidines. En vermell, la part exònica del gen; en negre la part intrònica.

1.2.2 Seqüències senyal que regulen l'*splicing* alternatiu

La maquinària d'*splicing* ha de distingir entre els llocs d'*splicing* correctes i els críptics,

que són molt abundants (Blencowe, 2006; Sun and Chasin, 2000). A més, dins dels llocs d'*splicing* correctes, n'hi ha de forts i de dèbils (Thanaraj and Clark, 2001). Per tant, no n'hi ha prou amb les seqüències dels llocs d'*splicing*, sinó que cal que hi hagi seqüències estimuladores o silenciadores de l'*splicing* per poder distingir entre exons reals o pseudoexons (Matlin et al., 2005).

Tant els exons com els introns contenen seqüències curtes –i poc conservades- que regulen l'*splicing* (Matlin et al., 2005) (veure la Figura 3 per un esquema). Així, força exons contenen seqüències –que comparteixen espai amb la seqüència codificant (Blencowe, 2000)- que n'estimulen o n'inhibeixen l'*splicing*. D'altra banda, també hi ha seqüències reguladores als introns (Hastings et al., 2001; Simard and Chabot, 2002). Aquestes, molts cops, actuen a distància. Pel que fa als motius silenciadors, s'ha vist que són especialment rics en els pseudoexons, els exons alternatius i els introns flanquejants dels exons constitutius (Blencowe, 2006; Wang et al., 2004; Zhang and Chasin, 2004) i que són necessaris per permetre que l'*splicing* es doni correctament (Sun and Chasin, 2000). Finalment, s'ha vist que els motius reguladors dels *splicings* alternatius especifics de teixit estan situats als introns propers als exons alternatius regulats (Brudno et al., 2001; Sugnet, 2006).

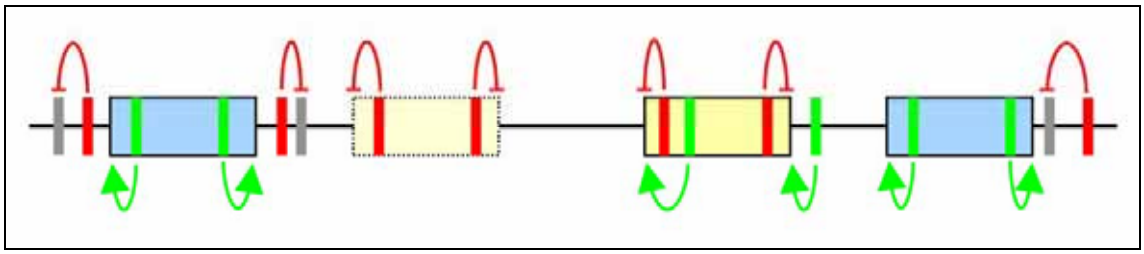

Figura 3. Esquema de la regulació de l'*splicing* alternatiu. L'esquema mostra un fragment de DNA genòmic: les capses blaves són els exons constitutius i la groga, l'alternatiu, mentre que la capsa amb línia discontínua és un pseudoexó; les barres grises mostren llocs d'*splicing* críptics, mentre que les verdes són seqüències estimuladores –ESE i ISE- i les roges, inhibidores –ESS i ISS. Les seqüències estimuladores afavoreixen la inclusió dels exons, mentre que les seqüències inhibidores impedeixen l'ús dels llocs d'*splicing*.

1.2.3 La maquinària d'*splicing*

Tota aquesta col·lecció de senyals associada al patró d'exons i introns del gen és processada per una complexa maquinària molecular anomenada spliceosoma. Aquest està format per 5 snRNPs (*small nuclear ribonucleoproteins*) –U1, U2, U4, U5 i U6- i

més de 300 proteïnes diferents (Jurica and Moore, 2003). Es creu que hi ha tantes proteïnes que en formen part per assegurar el reconeixement dels llocs d'*splicing*, acoblar l'*splicing* amb la maquinària transcripcional i introduir senyals per a processos post-transcripcionals (Nilsen, 2003). Ara, fins i tot, es creu que els spliceosomes interactuen entre ells, formant estructures supramoleculars (Azubel et al., 2006)

Durant molt temps s'havia cregut que l'spliceosoma es formava *in situ* en presència del pre-mRNA a processar, en canvi, alguns resultats dels darrers anys apunten a la formació del complex *a priori* (Stevens et al., 2002), mantenint el tema candent encara obert en el camp d'estudi (Bentley, 2005).

A l'hora de l'*splicing* hi ha diversos models per explicar el funcionament del mecanisme. El model que s'anomena "definició exònica" suggereix que la maquinària d'*splicing* busca llocs d'*splicing* prou separats i amb l'orientació correcta, quan els troba, l'exó és definit amb la unió dels snRNPs U1 i U2 i proteïnes SR (Berget, 1995). U1 s'uneix al lloc 5' i U2 s'uneix al BP, però aquest necessita l'ajuda d'U2AF per fer-ho (Lopez, 1998). Posteriorment, U4, U5 i U6 ajuden a definir el lloc 3' (Hastings and Krainer, 2001) –la definició de l'exó 3' terminal és una mica diferent (Goldstrohm et al., 2001). En canvi, es creu que els pre-mRNA amb introns curts utilitzen el model anomenat "definició intrònica" (Berget, 1995). A més, alguns dels passos que es donen per empalmar els exons requereixen la hidròlisi d'ATP i es creu que aquests són punts de control (Hastings and Krainer, 2001).

Tal com s'ha mencionat, a més de les snRNPs de l'spliceosoma, hi ha moltes altres proteïnes reguladores de l'*splicing* que s'uneixen i se separen dels transcrits a mesura que aquests maduren, estimulant o inhibint la inclusió dels exons (Smith and Valcarcel, 2000). Les proteïnes SR són la família de proteïnes reguladores de l'*splicing* més conegudes (Graveley, 2000). Contenen dominis d'unió a l'RNA i un domini amb dipèptids arginina/serina repetits (dominis RS), molt comú entre les proteïnes que regulen l'*splicing* alternatiu (Cowper et al., 2001). Les proteïnes SR s'uneixen als ESE –estimuladors exònics de l'*splicing*-, però s'ha vist que hi ha SRs que s'uneixen de manera específica al nus, quan encara no hi ha l'spliceosoma madur (Shen et al., 2004) i que després s'uneixen al lloc d'*splicing* 5' (Shen and Green, 2004). Per contra, hnRNP és el grup més conegut d'inhibidors de l'*splicing* (Krecic and Swanson, 1999) i a diferència de les proteïnes SR, les hnRNPs no tenen dominis RS (Smith and Valcarcel, 2000). Sembla que les proteïnes SR promourien *splicing*s proximals subòptims, mentre

que hnRNPs afavoririen els distals (Caceres and Kornblihtt, 2002; Smith and Valcarcel, 2000). Finalment, el control de tots aquests factors d'*splicing* alternatiu es basaria en les vies de transducció de senyals (Stamm, 2002).

1.2.4 Tipus d'*splicing* alternatiu

A nivell d'mRNA, l'*splicing* alternatiu es pot donar per una gran diversitat de mecanismes: donadors o acceptors d'*splicing* alternatius, retenció d'introns, omissió d'exons i exons excloents (veure Figura 4). A més, aquests mecanismes poden estar combinats (Smith and Valcarcel, 2000; Xing and Lee, 2006) per generar esdeveniments d'*splicing* força complexos (Zheng et al., 2005). En aquest sentit, Nagasaki i col·laboradors feren una classificació sistemàtica d'esdeveniments d'*splicing* alternatiu i n'arribaren a trobar més de 150 de diferents (Nagasaki et al., 2006). A més a més, s'ha vist que l'ús de diferents promotors també té relació amb l'*splicing* alternatiu, arribant a afectar llocs d'*splicing* distants (Kornblihtt, 2005).

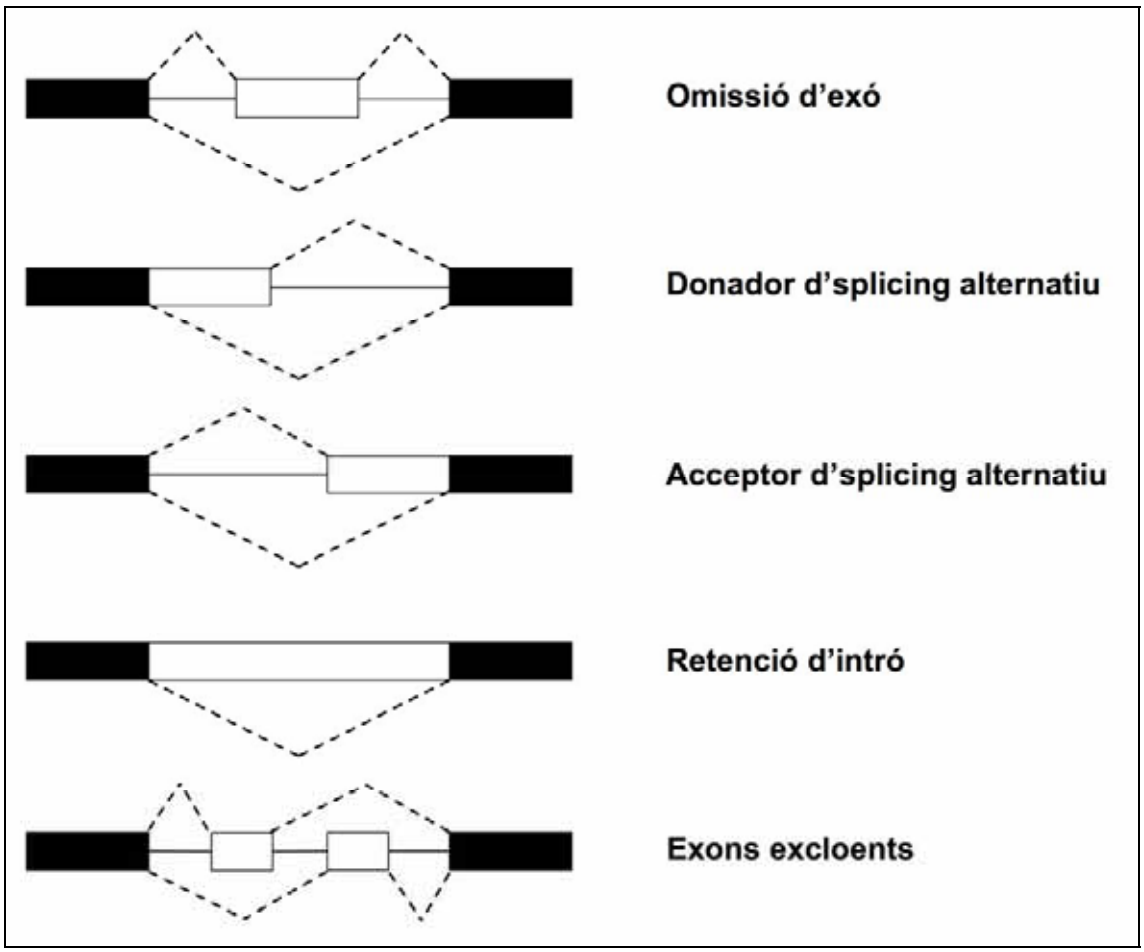

Figura 4. Mecanismes més corrents d'*splicing* alternatiu. Les capses negres representen els exons constitutius, mentre les blanques representen els alternatius. Les línies discontínues indiquen els possibles

splicings. L'omissió d'exó seria el mecanisme més freqüent (~40%). Els donadors d'*splicing* alternatiu són molt més freqüents que els acceptors d'*splicing*. La retenció d'introns es dona en menys del 3% dels casos. Hi ha un terç dels casos que corresponen als exons excloents i altres fenòmens complexos (Ast, 2004; Sugnet et al., 2004).

Normalment, l'*splicing* es dóna amb elements en cis, però s'han trobat força casos d'*splicing* en trans –empalmament d'exons de diversos pre-mRNAs- a *C. elegans* i organismes unicel·lulars, encara que sense afegir variabilitat proteica (Horiuchi and Aigaki, 2006). Tanmateix, també se n'ha trobat en organismes més complexes i, en aquests casos, implicant la part codificant. A Drosophila s'ha trobat el gen *mod(mdg4)* (Dorn et al., 2001; Horiuchi et al., 2003; Labrador et al., 2001), que està organitzat amb gens en les dues cadenes de DNA, i el gen *lola* (Horiuchi et al., 2003), que transcriu exons dels dos al·lels. A mamífers s'ha trobat trans-*splicing* intragènic (Caudevilla et al., 1998) i intergènic (Hirano and Noda, 2004; Li et al., 1999).

Addicionalment, tots aquests processats del transcrit poden donar lloc a un inici de traducció alternatiu o a un acabament alternatiu de la transcripció (Zavolan et al., 2003). D'aquesta manera, es calcula que prop del 90% dels processos d'*splicing* alternatiu estan associats a algun tipus de canvi en la proteïna final (Modrek and Lee, 2002).

1.3 Impacte de l'splicing alternatiu a nivell de proteïnes

Tot i la gran diversitat i complexitat dels mecanismes descrits anteriorment, a nivell de la seqüència peptídica tan sols es donen dos tipus d'efectes diferents: les insercions i delecions de fragments i les substitucions d'unes seqüències per unes altres (veure Figura 5) (Kondrashov and Koonin, 2001; Kondrashov and Koonin, 2003; Letunic et al., 2002; Liu and Altman, 2003).

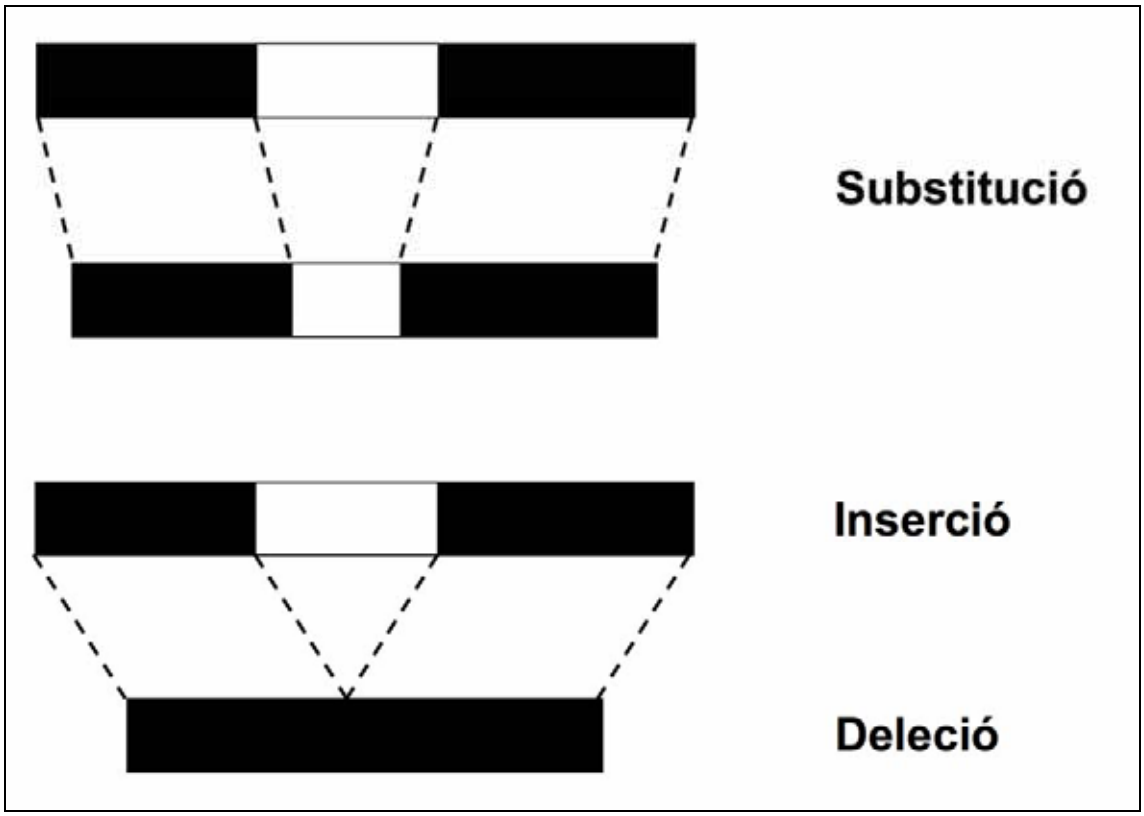

Figura 5. Efectes a nivell proteic de l'*splicing* alternatiu. Les capses negres representen la part constituent de la seqüència peptídica. En canvi, els fragments blancs són la part variant.

Òbviament, els canvis a nivell de la seqüència poden originar canvis estructurals en la proteïna, que són difícils de predir *a priori*. Així, s'ha vist que insercions petites poden produir canvis estructurals considerables, fins i tot, en parts allunyades de la proteïna (Stetefeld and Ruegg, 2005). Malauradament, no s'han fet prou estudis per interpretar, d'una manera global, l'impacte estructural i funcional dels canvis d'*splicing*. Nogensmenys, a continuació es mencionen una sèrie de resultats experimentals que ens parlen d'aquest impacte.

Així, s'ha vist que els canvis funcionals poden estar causats per canvis en la composició dels residus involucrats en interaccions o la catàlisi, com ara en el cas de l'FGFR2. FGFR2b i FGFR2c són dues isoformes dels receptors dels factors de creixement del fibroblast (FGF), que tenen diferent especificitat (Ornitz et al., 1996). S'ha vist que FGF7, FGF10 i FGF22 s'uneixen a FGFR2b, però no a FGFR2c. La cristal·lització de l'estructura de FGF10 unida a FGFR2b permeté veure que molts ponts d'hidrogen de la interacció es donen pel domini Ig-like D3 que diferencia les dues isoformes del receptor (Yeh et al., 2003), fet que explica perquè FGFR2c no pot unir-se amb aquells FGFs.

També s'han observat variacions funcionals com a conseqüència de canvis en la flexibilitat dels dominis; per exemple, les proteïnes AGX i WT1. AGX1 i AGX2 són dues isoformes d'una classe de pirofosforilases, que es diferencien per una inserció força flexible de 17 residus. La inserció provoca un canvi de l'estat oligomèric de les isoformes i de l'arquitectura del seu centre actiu (Peneff et al., 2001). Així, mentre que AGX1 és dimèrica i actua preferentment com a UDP-N-acetilgalactosamina pirofosforilasa, AGX2 ho fa millor com a UDP-N-acetilglucosamina pirofosforilasa. Per la seva banda, el gen WT1 codifica per un factor de transcripció que conté quatre dominis dits de zinc a la regió C-terminal. L'*splicing* alternatiu del gen genera isoformes amb o sense un tripèptid conservat de Lys-Thr-Ser (KTS) entre el tercer i quart dominis d'unió al DNA (Laity et al., 2000a). Això els dóna diferent flexibilitat en aquella regió de la proteïna i, conseqüentment, els dos tipus d'isoformes tenen diferent capacitat per unir-se al DNA (Laity et al., 2000b): WT1-KTS s'uneix al DNA i actua com un factor de transcripció, mentre que WT1+KTS s'uniria a l'RNA i interactuaria amb la maquinària d'*splicing*.

Altres canvis d'*splicing* poden provocar canvis d'elements d'estructura secundària, tal com li passa a Piccolo, que és una proteïna bastida (*scaffolding*) de les regions presinàptiques. Té dos dominis C_2 d'unió a Ca^{2+} (C_2A i C_2B) a la seva regió C terminal. S'ha vist que la inserció de 9 residus en el llaç que separa les làmines β 3 i 4 del domini C_2A provoca un reajustament estructural –l'antiga làmina 4 es converteix en dues petites hèlixs i la inserció ocupa el seu lloc com a làmina- que comporta la baixa afinitat pel Ca^{2+} de la isoforma llarga (Garcia et al., 2004).

Una altra variació estructural, causada per l'*splicing* alternatiu, que pot comportar canvis funcionals molt importants és el canvi en la forma de la proteïna. Així, en el cas de la Glutatió-S-transferasa *d'Anopheles dirus* s'ha trobat que la causa de les diferències funcionals entre AdGST1-3 i AdGST1-4 (Jirajaroenrat et al., 2001) és una inserció de cinc residus a la clivella en forma de V que provoca que el centre actiu de la isoforma AdGST1-4 sigui menys accessible al substrat (veure Figura 6) (Oakley et al., 2001). Un altre exemple és l'agrina, que és un proteoglicà heparan sulfat important a les unions neuromusculars. És el responsable de l'agregació dels receptors d'acetilcolina en les membranes post-sinàptiques i també s'uneix a α-distroglicà. L'agrina té moltes variants d'*splicing* que generen isoformes amb diferent funció i expressió. Unes de les variants es donen per la inserció de 8, 11 o 19 residus al domini G3. El domini G3 té una

estructura de β-sandvitx. Les isoformes amb la inserció es pleguen correctament, però s'observa una expansió del llaç on s'inserten els residus i una làmina β es reorienta passant de còncava a convexa (Stetefeld et al., 2004).

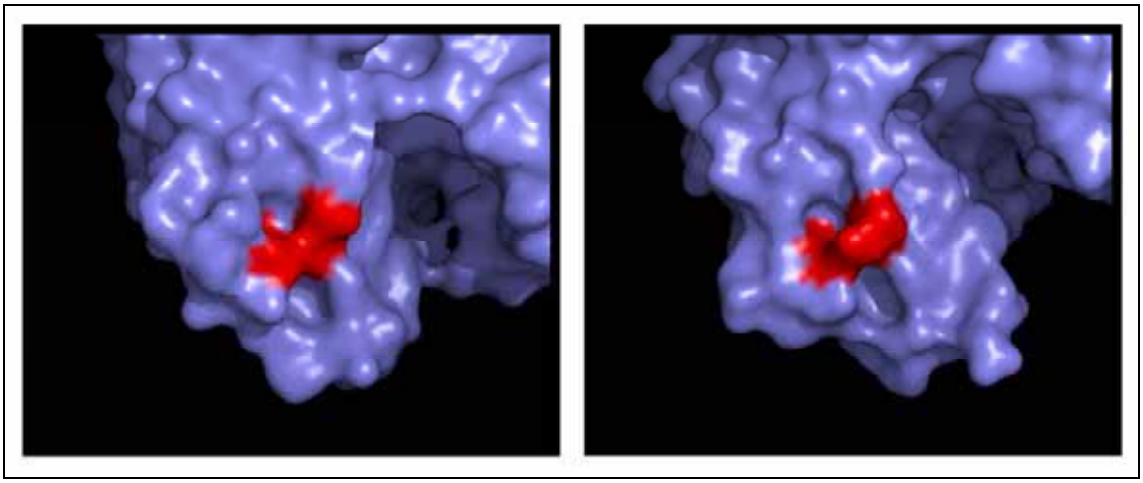

Figura 6. Impacte funcional de l'*splicing* alternatiu sobre el centre actiu de la Glutatió-S-transferasa. A l'esquerra, AdGST1-3 i a la dreta, AdGST1-4. En vermell, els residus responsables de la diferència d'accessibilitat (Asp110 i Asn126 a AdGST1-3; Glu116 i Arg134 a AdGST1-4).

Finalment, hi ha casos en que diverses de les anteriors característiques són presents en les variants d'*splicing*. EDA és un membre de la família de factors de necrosi tumoral involucrat en el desenvolupament ectodèrmic. Té dues isoformes –EDA-A1 i EDA-A2- que tan sols es diferencien per una inserció de dos residus (Glu308 i Val 309) en la regió d'unió al receptor d'EDA-A1. Aquesta inserció fa que les dues isoformes s'uneixin a diferents receptors perquè provoca canvis de conformació, de càrregues i de forma del lloc d'unió al receptor (Hymowitz et al., 2003).

Tots aquests estudis posen de manifest l'abundància d'efectes sobre l'estructura i funció de les proteïnes que pot tenir l'*splicing* alternatiu. No obstant això, el seu limitat nombre (Stetefeld and Ruegg, 2005) fa que no es puguin treure conclusions més generals.

1.4 Aproximació bioinformàtica a l'estudi de l'splicing alternatiu

La bioinformàtica pot permetre ampliar el coneixement que es té sobre les variants d'*splicing*, més enllà dels resultats experimentals. Així, la bioinformàtica s'ha aplicat a l'estudi de l'*splicing* alternatiu, tant en la predicció de llocs d'*splicing* com en l'anàlisi de les seves conseqüències. Tot seguit s'aborden aquests punts.

1.4.1 La bioinformàtica com a tècnica de suport

1.4.1.1 Estudi de l'*splicing* alternatiu a partir d'ESTs i mRNAs

Els projectes de seqüenciació genòmica s'han trobat amb unes necessitats imperioses a l'hora de la predicció de l'estructura gènica: la predicció de llocs d'*splicing*, en general, i dels esdeveniments d'*splicing* alternatiu, en particular. Així, doncs, l'anàlisi bioinformàtica de les dades d'aquests projectes ha intentat ajudar en la predicció de gens i, com a conseqüència, s'ha generat molta informació referent als patrons d'*splicing*, que ha comportat la construcció de diverses bases de dades per emmagatzemar-la.

Els projectes bioinformàtics de predicció d'estructura gènica i d'esdeveniments d'*splicing* alternatiu es basen, principalment, en l'alineament d'ESTs i/o mRNAs a fragments de DNA genòmic. Aquests projectes han anat apareixent a mesura que s'anaven obtenint les dades de la seqüènciació, primer amb els genomes animals (Kan et al., 2001; Modrek et al., 2001), però després s'han estès als de plantes (Haas et al., 2002; Iida et al., 2004). A més, també s'han fet projectes per trobar variants d'*splicing* amb especificitat tumoral (Wang et al., 2003b; Xie et al., 2002) o tissular (Xu et al., 2002). El seu principal problema és que, al treballar amb ESTs, són molt sensibles a la seva cobertura, tan pel que fa a la seva quantitat (Brett et al., 2002) com a la distribució de teixits estudiats (Gupta et al., 2004).

Pel que fa a les bases de dades, n'hi ha de diversos tipus: algunes es dediquen a extreure la informació d'altres bases de dades més generals, com ara SwissProt (Boeckmann et al., 2003), mentre que la majoria prediuen esdeveniments d'*splicing* alternatiu a partir d'alineaments d'mRNAs i ESTs. Dins del primer grup hi encabiríem ASDB (Dralyuk et al., 2000), ASMamDB (Ji et al., 2001) i PASDB (Zhou et al., 2003a). Pel que fa a les bases de dades generades automàticament, hi ha PALS db (Huang et al., 2002), ASAP (Lee et al., 2003) i la seva versió més desenvolupada (Kim et al., 2007), EASED (Pospisil et al., 2004). Finalment, un cas a part és el projecte ASD (Stamm et al., 2006; Thanaraj et al., 2004) que conté alhora bases de dades anotades manualment (Stamm et al., 2000) i altres generades computacionalment (Clark and Thanaraj, 2002).

1.4.1.2 Ús de microxips en l'estudi de l'*splicing* alternatiu

L'extensió de l'ús dels microxips i l'aparició de noves tècniques d'hibridació diferencial, juntament amb el gran desenvolupament de les eines bioinformàtiques, ha

permès que ens els darrers anys s'hagin fet diversos estudis a gran escala de l'*splicing* alternatiu (Blanchette et al., 2005; Castle et al., 2003; Johnson, 2003; Le et al., 2004; Pan et al., 2006; Pan et al., 2004; Shai et al., 2006; Stolc et al., 2004; Sugnet, 2006; Ule et al., 2005; Wang et al., 2003a; Watahiki et al., 2004; Yeakley et al., 2002).

En paral·lel, el més gran coneixement sobre l'*splicing* ha obligat a tenir-lo en compte a l'hora de dissenyar els microxips. Bàsicament, hi ha dues metodologies pel que fa al disseny dels microxips i les sondes: uns són els "anotar per dissenyar" (Davis et al., 2000; Spingola et al., 1999; Srinivasan et al., 2005), mentre els altres són els "dissenyar per anotar" (Johnson, 2003; Shoemaker et al., 2001) En el primer cas, l'*splicing* és anotat prèviament i el xip es dissenya per obtenir informació quantitativa del nivell d'expressió de les isoformes d'mRNA ja conegudes. Pel que fa a la segona estratègia, l'objectiu és trobar noves isoformes d'mRNA o nous exons alternatius, és a dir, anotar l'estructura gènica.

A més a més, hi ha diverses plataformes de microxips: *tiling*, exó i/o unió i disseny dirigit (Cuperlovic-Culf et al., 2006). Les dues primeres plataformes permeten el descobriment de noves isoformes i fer una certa mesura de l'expressió –no obstant, només la primera permet cobrir tot el genoma (Mockler and Ecker, 2005), ja que per les d'unió es requereix un coneixement previ de les unions d'*splicing* (Castle et al., 2003; Fehlbaum et al., 2005; Srinivasan et al., 2005). Per contra, les plataformes de disseny dirigit permeten quantificar el nivell d'expressió de les isoformes, però necessiten que es conegui l'estructura gènica; per tant, no serveixen per detectar noves variants d'*splicing*.

Paral·lelament al desenvolupament de la metodologia dels microxips també s'ha hagut de produir un gran avenç en les tècniques bioinformàtiques que en possibiliten l'anàlisi, car ha estat necessari adaptar-se al paradigma "un gen, múltiples productes" (Cuperlovic-Culf et al., 2006; Lee and Roy, 2004).

En principi, les eines d'anàlisi de dades utilitzades en els experiments estàndard de microxips també es poden utilitzar per l'anàlisi dels microxips d'isoformes. No obstant això, els microxips per estudiar l'*splicing* alternatiu tenen una problemàtica específica: cada sonda pot representar la suma d'expressions de diverses isoformes i diferents sondes representen l'expressió del mateix gen. A més, s'obté la informació sobre la coexpressió dels exons i les unions d'*splicing*, però no informació de la seqüència,

quantitat o coexpressió de les isoformes (Cuperlovic-Culf et al., 2006). Per això, s'han utilitzat diverses correccions i aproximacions metodològiques al problema (Clark et al., 2002; Cline et al., 2005; Hu et al., 2001; Johnson, 2003; Le et al., 2004; Pan et al., 2004; Shai et al., 2006; Srinivasan et al., 2005; Wang et al., 2003a). Tot i això, aquest és un camp d'estudi que continua obert (Cuperlovic-Culf et al., 2006).

1.4.2 La bioinformàtica com a eina per estudiar problemes biològics associats a l'*splicing* alternatiu

És evident que l'evolució ha comportat un augment de la complexitat del genoma dels organismes. Aquesta major complexitat –per raó de l'augment de mida del genoma o de complexitat de l'estructura gènica- porta a un augment de la complexitat del proteoma (Lynch and Conery, 2003; Valentine, 2000). En el seu interès per l'estudi de la diversitat lligada a l'*splicing* alternatiu, la bioinformàtica ha destacat clarament per les anàlisis des de dues vessants: la conservació dels esdeveniments d'*splicing* i les seves conseqüències.

1.4.2.1 Estudis de conservació de l'*splicing* alternatiu

L'anàlisi de la conservació de l'*splicing* alternatiu entre espècies, té un valor afegit al mer fet d'estudiar l'evolució dels gens, ens diu quins esdeveniments són funcionalment importants. Els canvis d'*splicing* poden portar a un canvi de la pauta de lectura, per això hi ha qui ha dit que els canvis que són múltiples de tres estarien afavorits evolutivament (Magen and Ast, 2005). D'aquesta manera prenen molta importància els casos d'*Alu* com a generador de canvis de pauta (Kreahling and Graveley, 2004) i l'NMD com a control dels canvis sense sentit (Lewis et al., 2003). Els elements transposables, com ara *Alu*, estan àmpliament repetits dins dels genoma humà. Així es creu que un 5% dels exons alternatius s'han originat a partir d'aquests elements (Sorek et al., 2002), molts cops a partir de l'exonització d'elements insertats en els introns que han patit mutacions originadores de llocs d'*splicing* (Lev-Maor et al., 2003). D'aquesta manera, la inclusió del nou exó pot portar una pèrdua de funció de la proteïna –pel trencament d'un domini funcional, el canvi de la pauta de lectura o l'aparició de codons d'STOP-, però si l'exó només és inclòs rares vegades dins dels transcrits, la funció es manté i l'exó alternatiu pot començar a evolucionar tal com un exó duplicat (Kreahling and Graveley, 2004; Sorek et al., 2002).

La bioinformàtica ha permès classificar els exons segons la seva presència en les

isoformes de les diferents espècies. Així, s'ha vist que els exons es poden dividir en tres tipus: constitutius –són presents en totes les isoformes-, conservats –exons alternatius que són en diverses espècies- i específics –exons alternatius propis d'una espècie (Blencowe, 2006). Una altra manera de classificar els exons és la que utilitzaren Modrek i Lee, basada en la seva abundància en els mRNAs (Modrek and Lee, 2003). Ells van classificar els exons alternatius en majoritaris i minoritaris i observaren que els graus d'inclusió eren semblants per als exons ortòlegs, quan compararen gens humans i murins. A més, veieren que la màxima variació es trobava en els exons minoritaris (Modrek and Lee, 2003), el què ha fet pensar que les formes minoritàries serien exons de nova aparició i amb una accelerada taxa evolutiva (Wang, 2005). A més, l'anàlisi d'humans i ratolins ha mostrat la conservació interespecifica en canvis de pauta de lectura i ús de codons STOP alternatius en determinats esdeveniments d'*splicing* (Thanaraj et al., 2003). Així mateix, s'ha vist que l'omissió d'exons és el mecanisme d'*splicing* alternatiu més freqüent en els casos conservats en mamífers (Sugnet et al., 2004). Finalment, sembla que la conservació de la seqüència gènica està estretament relacionada amb la conservació del seu patró d'*splicing* (Cusack and Wolfe, 2005).

Un altre punt d'interès en el camp de la conservació de l'*splicing* alternatiu ha estat la seva freqüència en les diverses espècies. Brett i col·laboradors van demostrar que el percentatge d'*splicing* alternatiu que es detecta depèn del repertori d'ESTs i que aquest percentatge en els diferents organismes animals no era diferent i que, per tant, les diferències fenotípiques no eren degudes a la capacitat d'empalmar els exons de diferents maneres (Brett et al., 2002). No obstant això, altres científics han trobat resultats oposats utilitzant aproximacions diferents (Kim et al., 2004), deixant, doncs, el camp obert per a estudis posteriors.

Pel que fa a l'efecte de l'*splicing* alternatiu sobre l'evolució proteica, de forma natural, alguns grups han mirat l'evolució dels exons excloents obtinguts per una duplicació exònica (Kondrashov and Koonin, 2001; Letunic et al., 2002). Després de la duplicació dels exons, aquests acostumen a no ser inclosos alhora al mateix transcrit per restriccions estructurals o funcionals (Letunic et al., 2002) –pèrdua de pauta, alteració del plegament estructural o regulació dels introns flanquejants. A partir d'aquí, sembla que els dos exons comencen a evolucionar i especialitzar-se de manera diferent, sense cap període de redundància (Kondrashov and Koonin, 2001). Més recentment, s'ha vist que les formes minoritàries d'*splicing* alternatiu estarien afectades per una pressió

selectiva força relaxada que els permetria evolucionar ràpidament (Xing and Lee, 2005) i que, a més a més, tindrien una tendència més gran a modificar els dominis funcionals (Pan et al., 2005). Això fa pensar que l'*splicing* alternatiu pot ser un mecanisme per a permetre una evolució neutral a escala genòmica (Masel, 2006).

1.4.2.2 Estudis d'impacte estructural i funcional

Tot i el seu interès, hi ha molt pocs estudis experimentals (Stetefeld and Ruegg, 2005). És per això que la bioinformàtica pot ser força útil per ampliar el coneixement de les conseqüències estructurals i funcionals de l'*splicing* alternatiu.

Els primers estudis van anar destinats a entendre els efectes dels canvis d'*splicing* en la variabilitat proteica, ja que les isoformes que es generen poden arribar a tenir menys d'un 50% d'identitat entre elles –encara que la majoria només varien en pocs residus- i poden tenir diferències de mida considerables (Kersey et al., 2000). Així, Kondrashov i Koonin van veure que la majoria dels canvis d'*splicing* alternatiu portaven associat un canvi en la mida de la proteïna (Kondrashov and Koonin, 2003). Van analitzar l'evolució de les proteïnes tenint en compte quina era l'isoforma ancestral i van concloure que les insercions anirien associades a canvis de funció, mentre les delecions generarien reguladors negatius (Kondrashov and Koonin, 2003). Pel que fa a les substitucions, una part s'originen en la duplicació d'exons (Kondrashov and Koonin, 2001). Els exons duplicats codifiquen, majoritàriament, per 30-50 aminoàcids (Letunic et al., 2002) i tenen una similitud entre ells que va del 30 al 90% (Kondrashov and Koonin, 2001). Aquests exons acostumen a ser excloents, fet que s'ha relacionat amb la conservació de la pauta de lectura, per evitar problemes en el plegament de la proteïna o bé perquè els introns flanquejants no són del mateix tipus (Letunic et al., 2002). Finalment, el seu patró de variació és mimètic en els gens ortòlegs, la qual cosa indica que hi ha unes restriccions funcionals i estructurals importants (Kondrashov and Koonin, 2001). Així, s'ha vist que encara que hi hagi una gran variació en la seqüència, la predicció d'estructura secundària no acostuma a mostrar grans canvis (Boue et al., 2002), encara que hi ha algunes excepcions (Wen et al., 2004).

Altres estudis s'han centrat en estudiar possibles biaixos funcionals o estructurals de l'*splicing* alternatiu. Pel que fa a l'estructura, s'ha vist que la distribució de l'*splicing* alternatiu no estaria esbiaixada cap al centre ni cap als extrems de la proteïna (Kriventseva et al., 2003), però algun anàlisi de genòmica estructural ha permès

concloure que la majoria de residus alternatius estarien situats a la superfície de les proteïnes i, principalment, en zones sense elements d'estructura secundària definits (Wang et al., 2005), però sense correlació amb els llocs d'interacció entre proteïnes (Offman et al., 2004). D'altra banda, s'ha vist que hi ha dominis funcionals que estan més afectats que altres per l'*splicing* alternatiu (Liu and Altman, 2003; Resch et al., 2004) i que l'*splicing* alternatiu té tendència a actuar sobre dominis o elements funcionals sencers, enlloc de fer-ho sobre fragments (Kriventseva et al., 2003), encara que s'ha constatat que hi ha casos en que el domini només és funcional en la isoforma curta, quan desapareix la disrupció (Hiller et al., 2005).

Finalment, s'han intentat utilitzar tècniques de modelatge per homologia (Furnham et al., 2004) i *threading* (Wang et al., 2005) per analitzar l'estructura tridimensional de les isoformes. Els resultats obtinguts són ambivalents, ja que encara que s'obtenen models de bona qualitat per a les delecions és molt més difícil en el cas de les insercions (Furnham et al., 2004), a causa de la manca de plantilla (Marti-Renom et al., 2000), i per un terç de les isoformes no es pot construir l'estructura (Wang et al., 2005). A més, s'ha vist que els canvis d'*splicing* molt curts, que haurien de ser més fàcils de modelar, tindrien una estructura secundària diferent a la de les zones flanquejants (Wen et al., 2004).

Tots aquests estudis han començat a aclarir el rol que juga l'*splicing* alternatiu com a regulador de la funció proteica, mitjançant canvis estructurals. Tanmateix, encara hi ha molts aspectes que no s'han clarificat i que ens han impulsat a realitzar el treball presentat en aquesta tesi.

OBJECTIUS

2 Objectius

El treball realitzat en aquesta tesi s'emmarca en l'estudi de la variabilitat originada per l'*splicing* alternatiu i intenta aprofundir en els seus efectes a nivell de les proteïnes. Per acomplir aquest objectiu general, ens hem marcat una sèrie de fites:

- **Comparar l'*splicing* alternatiu i la duplicació gènica com a fonts de variabilitat proteica.** El coneixement que es té sobre els efectes estructurals i funcionals de la duplicació gènica és superior al que es té de l'*splicing* alternatiu; per tant, pensem que la comparació pot permetre inferir conseqüències estructurals i/o funcionals dels canvis en les isoformes.
- **Analitzar la conservació interespecífica dels efectes de l'*splicing* alternatiu com a mecanisme de modulació funcional.** L'*splicing* alternatiu s'ha postulat com a causant de les diferències entre organismes, però ben poca cosa se sap sobre si les diverses espècies l'utilitzen de la mateixa manera a l'hora de generar diversitat proteica.
- **Estudiar els efectes de l'*splicing* alternatiu sobre una família funcional de proteïnes.** Hem elegit la família dels factors de transcripció pel seu gran interès biològic i perquè és el sistema en què els efectes de d'*splicing* alternatiu poden tenir un impacte fenotípic més ric.
- **Trobar un protocol que permeti la identificació d'esdeveniments d'*splicing* alternatiu equivalents en diverses espècies.** L'anotació automàtica dels genomes i l'avenç en la recerca biomèdica necessiten determinar quins esdeveniments d'*splicing* alternatiu són homòlegs, és a dir, juguen el mateix rol biològic en diferents espècies.

MATERIALS I MÈTODES

3 Materials i mètodes

Durant la realització dels treballs que han donat lloc a la present tesi s'han utilitzat les mateixes tècniques per obtenir els resultats presentats en diferents capítols. Així, s'ha cregut oportú tenir una sola secció de materials i mètodes, enlloc d'anar repetint les mateixes explicacions en diversos capítols.

Aquesta secció agrupa les tècniques utilitzades en les anàlisis genòmiques i proteòmiques dels capítols 4, 5 i 6.

3.1 Bases de dades utilitzades

La major part del treball presentat en aquesta tesi es va fer utilitzant dades de la base de dades SwissProt (Boeckmann et al., 2003). No obstant això, en algunes parts o com a controls ha estat necessari l'ús de bases de dades alternatives.

3.1.1 SwissProt

SwissProt (http://expasy.org/sprot/) (Boeckmann et al., 2003) és una base de dades de proteïnes supervisada manualment. El servidor conté una sola seqüència proteica per entrada; no obstant això, també té informació per regenerar les isoformes alternatives, quan n'hi ha.

En l'anàlisi genòmica (capítol 4), utilitzàrem les dades de SwissProt per extreure informació tant de gens amb variants d'*splicing* com de gens amb duplicats.

Les anàlisis proteòmiques (capítols 4, 5 i 6) es feren preferentment amb aquesta base de dades, si no és que s'especifica al text.

3.1.2 AltSplice

AltSplice (http://www.ebi.ac.uk/asd/altsplice/) (Thanaraj et al., 2004) és una base de dades de variants d'*splicing* alternatiu generada automàticament a partir de les anotacions d'Ensembl (Birney et al., 2006), basades en gens coneguts, cDNAs i ESTs (Curwen et al., 2004).

En l'anàlisi genòmica (capítol 4), utilitzàrem les dades d'AltSplice (versions 2.0 per humà i ratolí) com a font dels gens amb *splicing* alternatiu.

3.1.3 Ensembl

Ensembl (http://www.ensembl.org/index.html) (Birney et al., 2006) és una base de dades secundària per a l'anotació automàtica de genomes eucariotes.

En l'anàlisi genòmica (capítol 4), utilitzàrem les dades d'Ensembl (versió 37.35j per humà; versió 37.34e per ratolí) per extreure la informació sobre els duplicats. Així mateix, també s'utilitzaren les prediccions de variants d'*splicing* per comprovar els resultats obtinguts.

A l'hora de buscar les famílies gèniques vam eliminar la redundància dels genomes, quedant-nos només amb la seqüència del transcrit predit de més llargada i vam utilitzar un filtre de baixa complexitat per evitar agupar els gens d'acord amb zones altament repetitives (Wootton and Federhen, 1996).

3.1.4 InParanoid

InParanoid (http://inparanoid.sbc.su.se/) (O'Brien et al., 2005) és una base de dades que agrupa els gens eucariotes en clústers d'ortòlegs.

InParanoid s'utilitzà per detectar els gens humans amb ortòlegs a llevat, mosca i ratolí. Aquestes dades s'utilitzaren en l'anàlisi genòmica (capítol 4).

3.1.5 SEGE

SEGE (http://sege.ntu.edu.sg/wester/intronless/) (Sakharkar and Kangueane, 2004) és una base de dades generada automàticament a partir de les dades de GenBank (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide) (Benson et al., 2007) que conté gens eucariotes sense introns.

Aquesta base de dades fou utilitzada en l'ànalisi genòmica (capítol 4) per descartar la retrotransposició com a causa de l'anticorrelació entre *splicing* alternatiu i duplicació gènica.

3.1.6 ASAP

ASAP (http://bioinfo.mbi.ucla.edu/ASAP/) (Lee et al., 2003) és una base de dades d'isoformes d'*splicing* alternatiu generada automàticament a partir de les ESTs dipositades a UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) (Wheeler et al., 2003).

L'utilitzàrem com a control en algunes de les nostres anàlisis proteòmiques (capítol 4).

3.1.7 Pfam

Pfam (http://www.sanger.ac.uk/Software/Pfam/) és una base de dades d'alineaments múltiples de dominis proteics o regions conservades (Sonnhammer et al., 1997). Està basada en alineaments múltiples i perfils obtinguts amb *Hidden Markov Models* a partir de les seqüències de SwissProt i TrEMBL (http://expasy.org/sprot/) –traducció automàtica de les seqüències dipositades a l'EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) i als seus homòlegs americà i japonès.

Particularment, nosaltres utilitzàrem la fracció Pfam-A, que està anotada manualment i que en les darreres versions, ja té milers d'alineaments de famílies (Bateman et al., 2004).

Utilitzàrem Pfam en la predicció de dominis funcionals dels factors de transcripció (capítol 6) i com un exemple de famílies de duplicats altament divergents (capítol 4).

3.1.8 SMART

SMART (*Simple Modular Architecture Research Tool*) (http://smart.embl-heidelberg.de/) és una base de dades de dominis funcionals basada en perfils construïts amb *Hidden Markov Models* extrets de seqüències homòlogues (Schultz et al., 1998). SMART està especialitzat en dominis de senyalització, nuclears i extracel·lulars.

D'aquesta manera, la base de dades, que ja té diversos centenars de dominis funcionals en les darreres versions (Letunic et al., 2004), permet anotar la composició de dominis de les proteïnes. Nosaltres l'utilitzàrem en la predicció de dominis funcionals dels factors de transcripció (capítol 6)

3.2 Obtenció de les dades

La informació sobre el nombre d'isoformes, les posicions inicials i finals dels canvis d'*splicing* i el seu tipus –substitucions o insercions/delecions- s'extragué analitzant la base de dades SwissProt. En examinar els fitxers en format text de la base de dades, totes les línies comencen amb un identificador de dues lletres que indica la secció d'informació. En el nostre cas, buscàvem l'etiqueta "FT VARSPLIC", que indica que hi ha un canvi en la seqüència causat per l'*splicing* alternatiu. En aquest camp hi ha la informació de les posicions inicial i final del canvi i les seqüències substituïdes o

insertades. En el cas de les delecions, apareix la paraula "MISSING". S'ha de fer notar que en les darreres versions de la base de dades, aquesta informació s'indica amb l'etiqueta "FT VARSEQ"

3.2.1 Obtenció dels factors de transcripció

Els factors de transcripció analitzats al capítol 6 es van obtenir de la base de dades SwissProt mitjançant la següent petició SRS (Etzold et al., 1996) (http://expasy.org/srs5bin/cgi-bin/wgetz): *Keywords "DNA-binding AND transcription" OR Description "transcription factor" OR Gene Name "transcription factor" OR Comment "transcription factor" BUTNOT Description "polymerase OR kinase OR aldolase OR hydrolase OR putative"*.

3.2.2 Obtenció dels enzims

Per obtenir el conjunt d'enzims, que s'utilitzen com a control en l'anàlisi d'expressió gènica al capítol 6, es buscaren totes les entrades que tenien un número EC associat.

3.3 Obtenció de les famílies de paràlegs

Per obtenir les famílies de gens duplicats, es clusteritzàren les seqüències mitjançant el programa CD-HIT (Li et al., 2001) utilitzant diversos percentatges d'identitat i els corresponents paràmetres optimitzats (veure Taula 2). Aquest programa ordena les seqüències segons la seva mida i, començant per la més llarga, agupa les seqüències que tinguin una identitat que iguali o superi la que s'utilitza com a llindar. Per accelerar l'algoritme cerca pèptids idèntics de diferents mides (Li et al., 2001). Quan el llindar d'identitat de seqüència era baix, utilitzàvem la variant MCD-HIT.

Identitat	Programa	-c	-n
90%	cd-hit	0.9	5
80%	cd-hit	0.8	5
70%	cd-hit	0.7	5
60%	cd-hit	0.6	4
50%	mcd-hit	0.5	3
40%	mcd-hit	0.4	2

Taula 2. Paràmetres per a la clusterització de seqüències proteiques. Depenent del llindar identitat del clúster s'ha d'utilitzar un programa o un altre i uns paràmetres optimitzats: c és la identitat llindar, mentre que n és la mida de pèptid que s'utilitza en la cerca.

Com més alt és el percentatge d'identitat utilitzat com a llindar, més conservades són les famílies de paràlegs –les famílies al 40% se suposa que han de ser més grans i contenir membres més distants que les famílies al 80%.

3.4 Alineament de seqüències

Per fer alineaments de seqüències hi ha diverses estratègies. Nosaltres utilitzàrem l'algoritme d'alineament global proposat per Neddleman i Wunsch (Needleman and Wunsch, 1970). Aquest algoritme de programació dinàmica té l'avantatge que proporciona un alineament global òptim –a diferència dels algoritmes basats en alineaments locals, com BLAST (Altschul et al., 1990)-, però només pot ser utilitzat per alinear dues seqüències, a causa de l'elevat cost computacional.

Com a matriu de substitució d'aminoàcid utilitzàrem, generalment, la matriu BLOSUM 62 (Henikoff and Henikoff, 1992). Els paràmetres per penalitzar l'obertura i eixamplament de forats en l'alineament foren -11 i -1, respectivament.

3.5 Predicció de dominis

Els dominis funcionals es prediren mitjançant l'eina RPS-BLAST –una variant del PSI-BLAST (Altschul et al., 1997)- utilitzant un E-valor de 0.02 i amb el filtre de baixa

complexitat actiu. Aquest programa proporciona el nom i la localització dels extrems N i C terminal dels dominis.

Cercàrem els dominis funcionals a la base de dades CDD (per *Conserved Domains Database*) (Marchler-Bauer et al., 2005). CDD agrupa la informació de diverses bases de dades amb informació de dominis funcionals –predits a partir de perfils d'alineaments múltiples. Tot i que dóna més informació, nosaltres només ens quedàrem amb els dominis funcionals basats en les bases de dades SMART (Letunic et al., 2004) i Pfam (Bateman et al., 2004).

Després, eliminàrem redundància de la composició de dominis: quan dos dominis se solapaven, eligíem el més llarg i descartàvem l'altre si la regió solapada era almenys el 60% de la mida del domini curt. Això ho férem perquè havíem utilitzat dues bases de dades i perquè alguns patrons de dominis inclouen subdominis al seu interior.

Algunes cerques amb RPS-BLAST poden ser insatisfactòries i no trobar cap domini funcional. L'èxit de les cerques té a veure amb la mida de la seqüència, el llindar d'E-valor utilitzat i la informació continguda a la base de dades elegida.

3.6 Modelatge comparatiu

L'estructura de la proteïna quinasa activada per mitògens número 9 (MAPK 9) d'humà, que apareix a la discussió del capítol 4, es va modelar a partir de l'estructura de la MAPK 10 humana (1jnk) (Xie et al., 1998). Les dues proteïnes pertanyen a la mateixa subfamília de MAPKs (Kultz, 1998) i la identitat de seqüència entre les dues proteïnes és del 84%, fet que permet obtenir un model bastant fidedigne.

L'alineament global entre la seqüència de la MAPK9 i la de l'estructura 1jnk es va obtenir amb un algoritme de programació dinàmica (Needleman and Wunsch, 1970) i la matriu de substitució d'aminoàcids BLOSUM 62 (Henikoff and Henikoff, 1992). Aquest alineament va ser l'entrada que es va utilitzar per al programa de modelatge comparatiu MODELLER (Sali and Blundell, 1993), en el seu ús més simple: la construcció d'un model a partir d'una plantilla que està alineada amb la proteïna d'interès. El programa té una sèrie de restriccions espacials per millorar la qualitat dels models (Eswar et al., 2003).

3.7 Identificació d'esdeveniments homòlegs d'splicing alternatiu entre diferents espècies

Considerarem que dos esdeveniments d'*splicing* alternatiu en dues espècies diferents eren homòlegs quan tenien un rol biològic similar en les dues espècies, és a dir, quan la variabilitat en les dues espècies –tant de funció com de lloc d'expressió- era la mateixa (veure Figura 7). Això vol dir que cada isoforma implicada en un esdeveniment d'*splicing* alternatiu en una espècie té una isoforma equivalent en un esdeveniment d'*splicing* alternatiu d'una altra espècie.

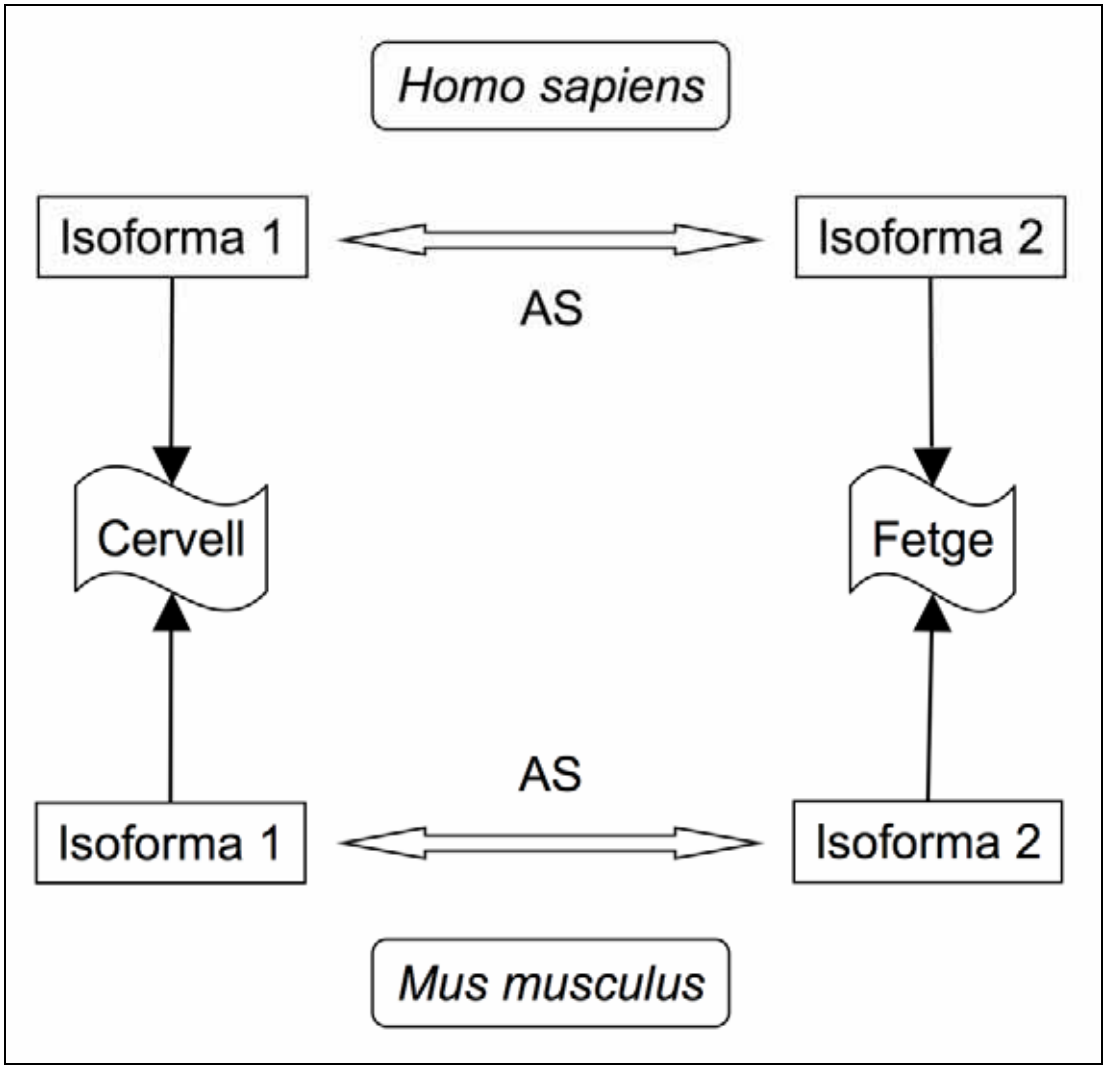

Figura 7. Exemple d'esdeveniment equivalent. En aquest cas, una proteïna humana i una de ratolí tenen la mateixa funció bioquímica i ambdues tenen diverses isoformes amb especificitat tissular: la isoforma 1 humana i la isoforma 1 de ratolí són equivalents perquè s'expressen al cervell; mentrestant, la isoforma 2 humana i la isoforma 2 murina s'expressen ambdues al fetge. L'equivalència particular de les isoformes comporta l'equivalència de l'esdeveniment d'*splicing* alternatiu definit com els canvis entre les isoformes 1 i 2 en l'humà i les isoformes 1 i 2 en el ratolí.

Desafortunadament, no vam trobar bases de dades amb anotacions sobre l'equivalència biològica de les isoformes; per tant, primer de tot, vam haver d'examinar manualment la bibliografia per trobar gens ortòlegs –humans i d'altres espècies- amb esdeveniments d'*splicing* alternatiu que podien tenir un rol similar en ambdues espècies. Ens fixàrem en l'expressió tissular i en l'especificitat per estadis de desenvolupament.

Després, alineàrem les isoformes humanes amb les isoformes equivalents en altres espècies. També férem alineaments creuats (veure Figura 8). Finalment, calculàrem uns índexs d'identitat global i local per a cada alineament (veure Figura 8).

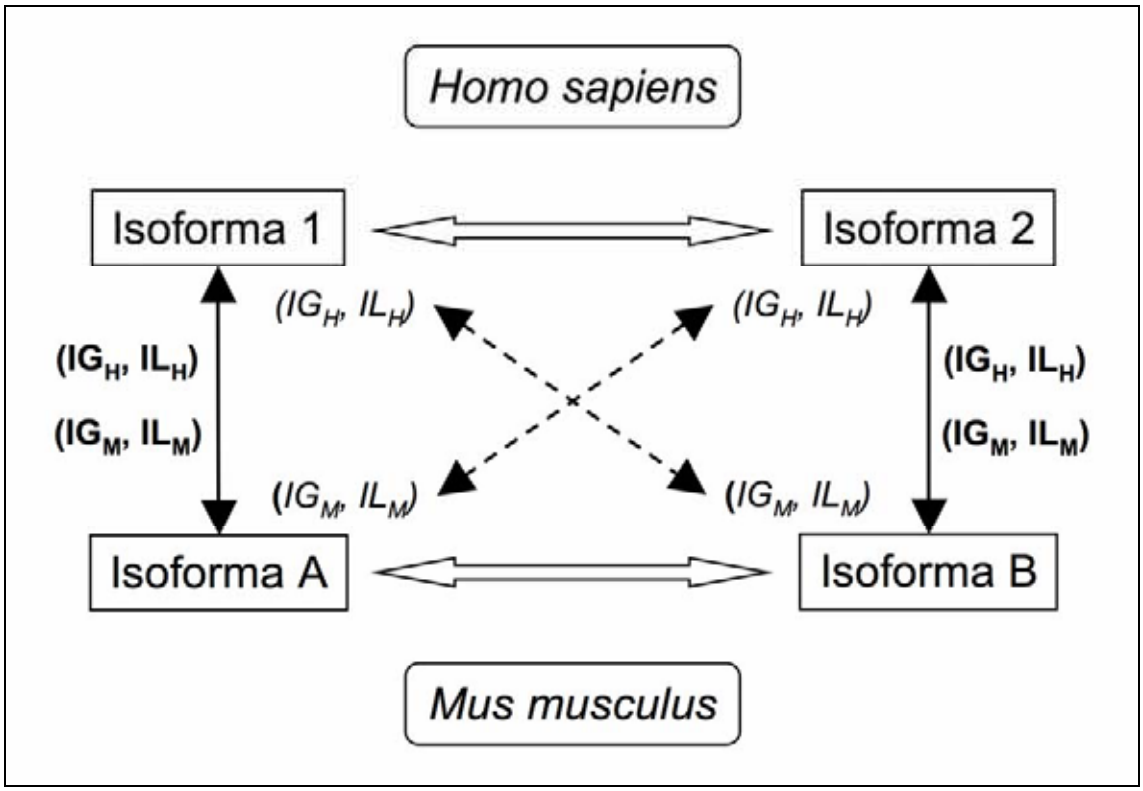

Figura 8. Comparacions i índexs per trobar esdeveniments equivalents. IG_H i IL_H són els índexs d'identitat global i local prenent la isoforma humana com a referència. IG_M i IL_M són els índexs d'identitat global i local prenent la isoforma de ratolí com a referència. Les fletxes contínues senyalen les comparacions d'isoformes equivalents. Les fletxes discontínues corresponen a les comparacions creuades.

Analitzant amb cura el nostre petit conjunt de dades, ens adonàrem que les isoformes equivalents tenien uns percentatges d'identitat –tant global com local- força alts i similars, mentre que les isoformes no equivalents tenien una identitat local molt més baixa. Per això, d'acord amb Kondrashov i Koonin (Kondrashov and Koonin, 2003) decidírem utilitzar un criteri basat en la seqüència per generar la llista d'esdeveniments homòlegs. Aquest criteri es basa en el fet que, en les isoformes equivalents, la identitat

de la zona variable és similar a la de la resta de la proteïna (Kondrashov and Koonin, 2003).

Així, desenvolupàrem un protocol per a l'assignació automàtica d'isoformes equivalents basat en l'existència de diverses isoformes en els gens ortòlegs i unes identitats global i local superiors al 50% –aquest llindar permet confiar en la conservació de la funció bioquímica (Tian and Skolnick, 2003; Wilson et al., 2000). En el cas de les insercions/delecions, alguns índexs locals no es poden calcular; per aquesta raó, miràrem que els canvis es localitzessin més o menys al mateix lloc. Posteriorment, tots els assignaments foren revisats manualment.

Amb aquest protocol semiautomàtic, trobàrem uns conjunts d'esdeveniments d'*splicing* alternatiu homòlegs: 99, entre humà i ratolí i 58, entre humà i rata. No se'n trobà entre humà i mosca del vinagre.

3.8 Prediccions d'accessibilitat i estructura secundària

Les prediccions d'accessibilitat i estructura secundària s'obtingueren amb els programes PREDACC (Mucchielli-Giorgi et al., 1999) i PREDATOR (Frishman and Argos, 1997), respectivament. Cal assenyalar que PREDACC prediu l'accessibilitat per a cada residu de la seqüència en dos estats: accessible i amagat o enterrat. D'altra banda, PREDATOR proporciona una predicció en tres estats de la regió d'estructura secundària a què pertany cada residu: hèlix α, làmina β i sense estructura definida (*coil*).

3.9 Anàlisis estadístiques

3.9.1 Solapament de les distribucions

El solapament de les distribucions de freqüència permet comparar fàcilment dos histogrames, donant una mesura intuïtiva de la seva semblança.

3.9.2 Tests estadístics

La prova de la t d'Student (http://home.clara.net/sisa/t-test.htm) s'utilitza per calcular la probabilitat de que dos promitjos o dues freqüències siguin iguals.

Si la quantitat de dades és petita, pot ser problemàtic fer un histograma de freqüències –encara que calculem intervals de confiança. En aquest cas, és millor aplicar el test de Kolmogorov-Smirnov (http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html),

que a més no necessita cap assumpció prèvia sobre la forma de les distribucions comparades. En el nostre cas, utilitzàrem el test KS per a dues mostres.

La prova de la χ^2 (http://home.clara.net/sisa/twoby2.htm) s'utilitza per provar la independència entre dos criteris de classificació de les dades.

3.9.3 Intervals de confiança

Pels histogrames de freqüències, calculàrem un interval de confiança per a cadascuna de les freqüències de la distribució assumint la hipòtesi d'independència (Durbin et al., 1998) seguint el treball de Goodman (Goodman, 1965) i tal com es veu a l'Equació 1.

$$p_i \pm \sqrt{\chi^2_{(\alpha/k,1)} * \frac{p_i * (1 - p_i)}{n}} \quad \text{(Equació 1)}$$

on p_i és la freqüència observada per la i-èsima variable, α és el nivell de significació utilitzat (95%, en el nostre cas), k és el nombre d'intervals i n és la mida total de la mostra. El valor de $\chi^2_{(\alpha/k,1)}$ s'obté de la distribució χ^2 per un grau de llibertat i una significació de α/k.

En alguns casos, hem calculat la quantitat de barres de l'histograma de freqüències que tenen els intervals de confiança solapats.

3.10 Anàlisi a nivell genòmic

3.10.1 Correlació/anticorrelació

Per mirar si hi havia relació entre la duplicació gènica i l'*splicing* alternatiu, s'agruparen totes les famílies gèniques segons la seva mida i, posteriorment, es calculà la fracció de gens amb variants d'*splicing* per a cada grup.

La fracció esperada per a una distribució a l'atzar és el resultat de dividir el nombre de gens amb *splicing* alternatiu pel nombre total de gens.

3.10.2 Anàlisi de la funció

Per estudiar si l'anticorrelació entre l'*splicing* alternatiu i la duplicació gènica té relació amb les funcions dels gens, analitzàrem les funcions de les proteïnes humanes dels quatre subconjunt de dades que teníem –amb o sense variants d'*splicing* i duplicats-, és a dir, gens únics sense variants d'*splicing*, gens únics amb *splicing* alternatiu, gens

duplicats sense isoformes i duplicats amb variants d'*splicing*.

Analitzàrem les funcions dels gens de les proteïnes de les bases dades SwissProt i AltSplice, mitjançant el servidor DAVID (http://david.abcc.ncifcrf.gov/home.jsp) (Dennis et al., 2003).

Aquest servidor analitza les proteïnes humanes per veure si hi ha biaixos en una gran diversitat de categories -termes de Gene Ontology (Ashburner et al., 2000), dominis proteics, vies metabòliques, categories funcionals, interaccions entre proteïnes, malalties, bibliografia i propietats de la seqüència.

3.10.3 Expressió de gens ortòlegs

Vam utilitzar les dades d'expressió per a diferents teixits de gens humans i murins del servidor SymAtlas (http://symatlas.gnf.org) (humà: GNF1H; ratolí: GNF1M) (Su et al., 2004). Els nivells d'expressió s'havien obtingut utilitzant algoritmes RMA (per *robust multiarray averaging*) (Bolstad et al., 2003; Irizarry et al., 2003)

S'analitzaren 30 teixits diferents comuns a les dues espècies –les dades de medul·la espinal de ratolí són el promig de les dades per les parts superior i inferior del teixit. Es compararen els nivells d'expressió de 553 factors de transcripció –109 amb variants d'*splicing* i la resta, sense. Com a control s'utilitzaren els patrons d'expressió de 1923 enzims.

Per aquells gens que tenien diverses rèpliques, les dades d'expressió es promitjaren.

Seguint els comentaris de Liao i Zhang (Liao and Zhang, 2006), utilitzàrem l'abundància relativa (RA) enlloc de la intensitat de senyal (S) (veure Equació 2).

$$\boldsymbol{RA}(\boldsymbol{i},\boldsymbol{j}) = \frac{\boldsymbol{S}(\boldsymbol{i},\boldsymbol{j})}{\sum_{\boldsymbol{j}=1}^{\boldsymbol{n}} \boldsymbol{S}(\boldsymbol{i},\boldsymbol{j})} \quad \text{(Equació 2)}$$

on *i* representa un gen i *j* és un teixit concret.

Seguint les recomanacions de la bibliografia (Huminiecki and Wolfe, 2004; Jordan et al., 2005; Liao and Zhang, 2006; Makova and Li, 2003; Yanai et al., 2004), per analitzar l'expressió dels factors de transcripció utilitzàrem la correlació de Pearson (r, veure Equació 3) i la distància euclidea (d, veure Equació 4).

$$r = \frac{\sum_{j=1}^{n}\left[RA_H(i,j) * RA_R(i,j)\right] - \left[\sum_{j=1}^{n} RA_H(i,j)\right] * \left[\sum_{j=1}^{n} RA_R(i,j)\right]}{\frac{n}{\sqrt{\sum_{j=1}^{n}\left[RA_H(i,j)\right]^2 - \left[\sum_{j=1}^{n} RA_H(i,j)\right]^2} * \sqrt{\sum_{j=1}^{n}\left[RA_R(i,j)\right]^2 - \left[\sum_{j=1}^{n} RA_R(i,j)\right]^2}}{n}}$$

(Equació 3)

on *i* representa un gen, *j* és un teixit concret, *H* i *R* indiquen si el patró d'expressió és d'humà o de ratolí, respectivament.

$$d = \sqrt{\sum_{j=1}^{n}\left[RA_H(i,j) - RA_R(i,j)\right]^2} \qquad \text{(Equació 4)}$$

on *i* representa un gen, *j* és un teixit concret, *H* i *R* indiquen si el patró d'expressió és d'humà o de ratolí, respectivament.

3.11 Anàlisi a nivell de proteïnes

3.11.1 Estudi de la conservació de les propietats físico-químiques

A l'hora d'estudiar la conservació de l'accessibilitat i l'estructura secundària dels residus, es feia un alineament global de les dues seqüències i, posteriorment, se substituïen els residus per les seves prediccions.

3.11.2 Caracterització de les substitucions

En el nostre treball hem hagut de fer diversos anàlisis per caracteritzar les substitucions d'aminoàcids causades per l'*splicing* alternatiu o la duplicació gènica.

La informació de la localització i residus implicats en les substitucions d'*splicing* alternatiu s'extragué de la base de dades SwissProt (Boeckmann et al., 2003).

Pel que fa a les duplicacions gèniques, la informació dels residus substituïts s'obtingué a partir dels alineaments de seqüència.

3.11.3 Identitat global

La identitat global correspon al percentatge de residus idèntics entre tots els residus alineats (veure Figura 9). Es calcula a partir d'un alineament d'un parell de seqüències

completes.

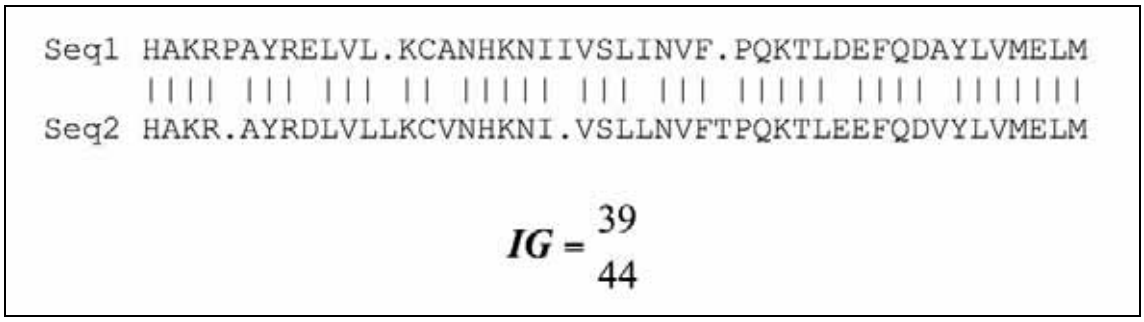

Figura 9. Càlcul de la identitat global. Després de l'alineament global es compten tots els residus idèntics i es divideixen pel nombre de residus alineats.

3.11.4 Identitat local

La identitat local es refereix al percentatge d'identitat entre parts de la seqüència. Normalment, la identitat local es refereix a parts substituïdes, però també s'aplica als dominis funcionals dels factors de transcripció.

3.11.4.1 Identitat local entre isoformes

En la major part dels casos, sempre que s'havia de calcular una identitat local, es feia un alineament global, però només es calculava el percentatge d'identitat localment (veure Figura 10). No obstant això, la identitat local entre variants d'*splicing* que apareix al capítol 4 es va calcular de manera diferent: s'alineaven només els fragments implicats en l'esdeveniment d'*splicing* i a partir d'aquí es comptaven els aminoàcids idèntics de tots els residus substituïts (veure Figura 10).

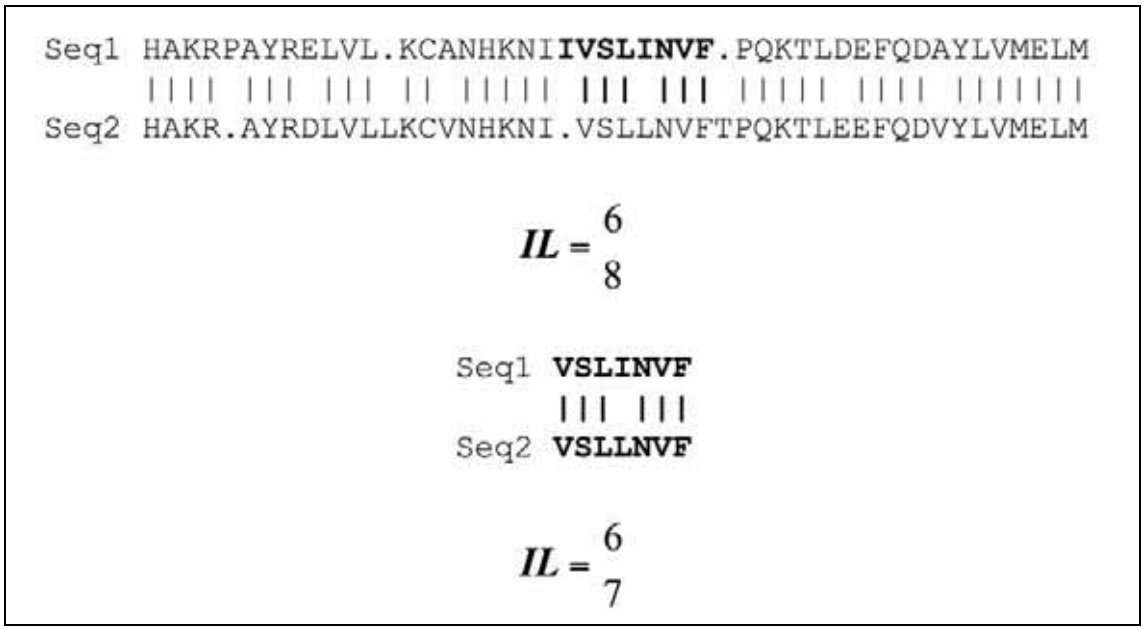

Figura 10. Càlcul de la identitat local. A la part superior, després de l'alineament global es compten tots els residus idèntics de la part en negreta i es divideixen per la seva mida. A la part inferior, s'extreuen els fragments i es procedeix com en el càlcul de la identitat global. Els resultats no han de ser necessàriament idèntics, perquè en el segon cas es té en compte el nombre de residus alineats, enlloc de la mida del canvi.

En aquest cas, per evitar comparacions sense sentit vam imposar que ambdues seqüències substituïdes havien de ser més llargues de 10 aminoàcids i la més curta havia de tenir una longitud d'almenys el 60% de la llargada de la seqüència més llarga. Aquests filtres s'usaren en el càlcul de la identitat local, però en cap dels altres anàlisis.

3.11.4.2 Identitat local entre duplicats

En el càlcul de la identitat local entre duplicats, vam diferenciar entre les famílies que tenien algun membre amb *splicing* alternatiu i les que no en tenien. El punt d'inici és un alineament global entre els duplicats i, a partir d'aquí, vam seguir estratègies diferents per als dos casos, depenent de si un dels duplicats tenia *splicing* alternatiu o no.

En el cas de les famílies de duplicats en que algun membre té *splicing* alternatiu, seguírem dos protocols diferents. En el primer, després de fer l'alineament entre la proteïna que té variants i els seus duplicats –sempre alineaments per parelles entre la isoforma de referència i els duplicats-, utilitzàrem una finestra lliscant de la mateixa mida que el canvi d'*splicing* i calculàrem la identitat local per a totes les possibles localitzacions de la finestra (veure Figura 11). En el segon protocol, analitzàrem la capacitat d'introduir la mateixa variabilitat; per tant, després d'alinear les seqüències, calculàrem el percentatge d'identitat local per la zona d'*splicing* i la seva corresponent

en el duplicat (veure Figura 12).

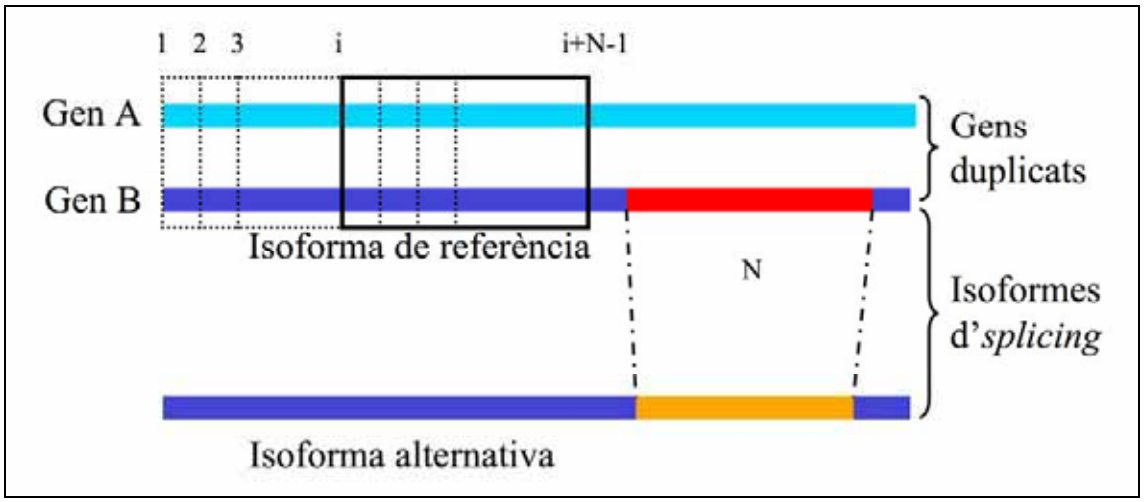

Figura 11. Identitat local en duplicats amb presència d'*splicing* alternatiu. Mètode totes posicions. Es genera una finestra lliscant de mida N (mida del canvi d'*splicing*) i es calcula la identitat local entre els duplicats en totes les posicions de la finestra.

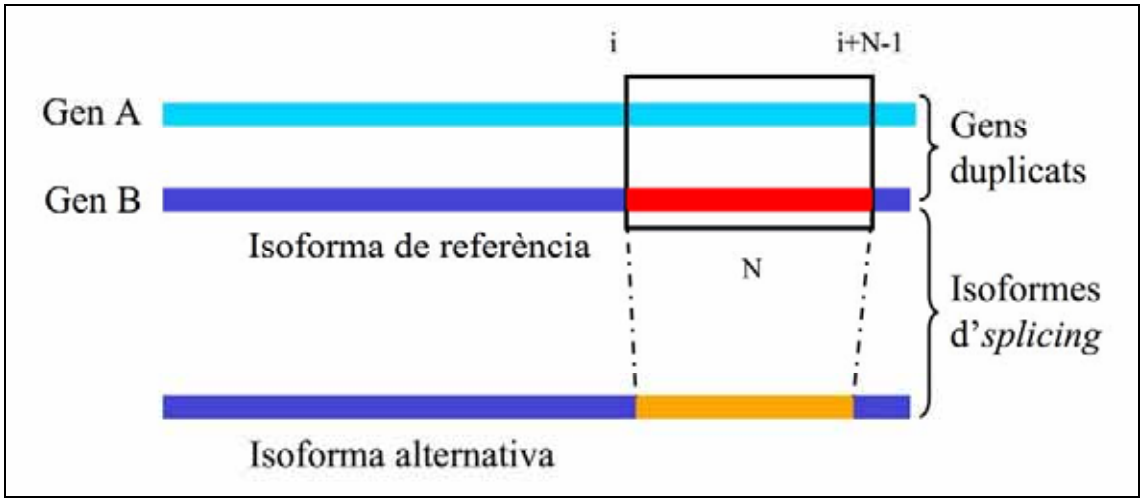

Figura 12. Identitat local en duplicats amb presència d'*splicing* alternatiu. Mètode mateixa posició. Es calcula la identitat local entre els duplicats en la mateixa regió on hi ha el canvi d'*splicing* alternatiu.

En el cas de les famílies sense membres amb variants d'*splicing* la comparació directa no és possible; per tant, utilitzàrem una finestra lliscant de mida fixa –100 residus de llargada (veure Figura 13).

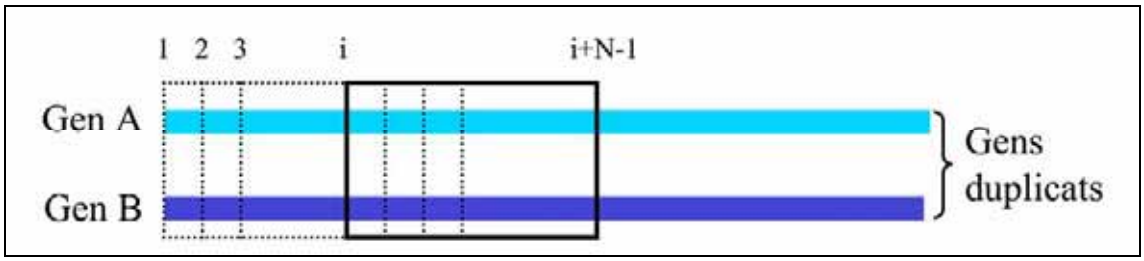

Figura 13. Identitat local en duplicats sense *splicing* alternatiu. Es genera una finestra lliscant de mida igual a 100 residus i es calcula la identitat local entre els duplicats en totes les posicions de la finestra.

3.11.4.3 Identitat local entre dominis funcionals

En aquest cas, s'alinearen només les seqüències que formaven part dels dominis funcionals, definides a partir de les prediccions fetes amb l'RPS-BLAST (veure secció 3.5).

3.11.5 Similitud local

La similitud entre regions substituïdes s'obtingué puntuant l'alineament entre els dos fragments, mitjançant una matriu de substitució PAM250 (Dayhoff et al., 1978), que està considerada una matriu prou general per permetre alinear correctament seqüències divergents.

Es van utilitzar les mateixes restriccions de mida que en el cas de la identitat local i la puntuació de l'alineament es va normalitzar utilitzant la seva longitud.

3.11.6 Canvis no conservatius

Definim com a canvis no conservatius aquelles substitucions d'aminoàcids que tenen una puntuació negativa en la matriu de substitucions BLOSUM 62 (Henikoff and Henikoff, 1992). Aquest mateix criteri fou utilitzat anteriorment en l'anotació d'SNPs (*Single Nucleotide Polymorphisms*) (Cargill et al., 1999).

La fracció de canvis no conservatius s'obtingué dividint el nombre de canvis amb puntuació negativa pel nombre total de canvis en l'alineament.

Els percentatges de canvis no conservatius apareguts a causa de l'*splicing* alternatiu o la duplicació gènica foren comparats amb el test de t-Student (http://home.clara.net/sisa/t-test.htm).

3.11.7 Distribució de la distància màxima entre canvis no conservatius

Per calcular les distàncies màximes seguírem diferents estratègies segons el nombre de substitucions involucrades en l'esdeveniment d'*splicing* alternatiu. Així, analitzàrem els casos amb una sola substitució i els casos amb dues substitucions.

En el cas dels esdeveniments d'*splicing* alternatiu amb una sola substitució –independentment del nombre d'insercions o delecions-, la distància màxima correspon a la distància que hi ha entre els canvis de residu més separats (veure Figura 14). La distància es normalitzà dividint-la per la mida de la proteïna.

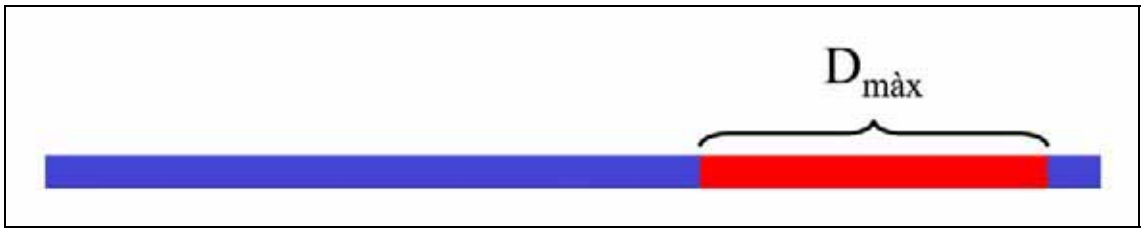

Figura 14. Distància màxima quan només hi ha una substitució d'*splicing* alternatiu.

El mateix criteri s'utilitzà pels corresponents duplicats.

En els cas de les isoformes amb dues substitucions, la distància màxima es mesurà com la quantitat d'aminoàcids separant els dos canvis d'*splicing*, normalitzada dividint per la distància separant els dos canvis de residus més allunyats (D'_{max}) (veure Figura 15).

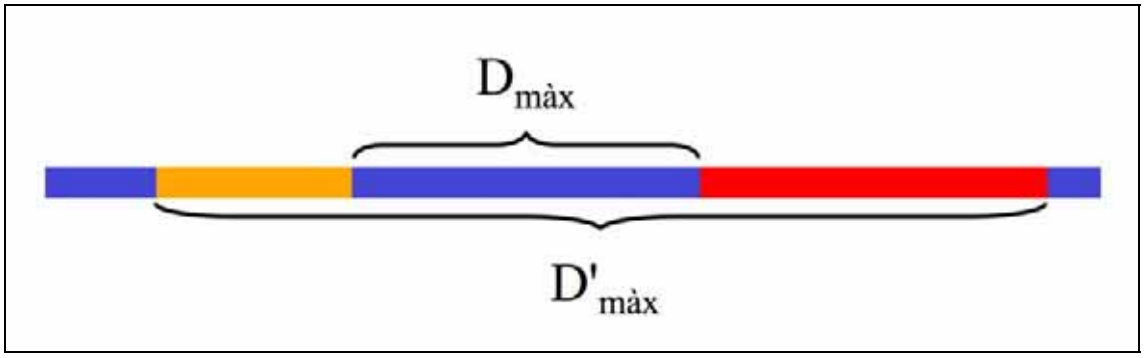

Figura 15. Distància màxima quan hi ha dues substitucions d'*splicing* alternatiu.

En els corresponents duplicats mesuràrem la distància entre els canvis de residus contigus més separats. També la normalitzàrem amb la distància entre els canvis més separats.

Les distàncies màximes en el cas de dos canvis foren comparades amb el test KS per a dues mostres.

3.11.8 Caracterització de les insercions/delecions

La informació de les delecions s'extragué de la base de dades SwissProt (Boeckmann et al., 2003), a partir de l'anotació "MISSING" en el camp VARSPLIC.

Les insercions estaven anotades com una substitució, amb la particularitat que el canvi era d'un sol residu per una seqüència que començava per aquell mateix aminoàcid.

Les insercions/delecions entre duplicats es trobaren alineant les seqüències per parelles. Allà on l'algoritme generava un forat en l'alineament, es comptava la quantitat de residus implicats.

3.11.8.1 Mida de les insercions/delecions

Les insercions/delecions s'agruparen en intervals, dels quals es calcularen les freqüències per generar un histograma.

En el cas de les duplicacions, es va aplicar una correcció de redundància consistent en ponderar la contribució de cada duplicat a la corresponent barra de l'histograma tenint en compte la mida de la família gènica, és a dir, dividíem cada compte sumat a l'histograma pel nombre de duplicats dins de la família.

Posteriorment, els dos grups de dades foren dividits, tenint en compte si els canvis afectaven els extrems N- o C-terminals. Els canvis que incloïen els extrems s'anomenaren externs i la resta, interns.

3.11.8.2 Solapament de les insercions/delecions

Per estimar el solapament entre les insercions/delecions de les variants d'*splicing* i els duplicats seguirem el següent protocol: primer, mapejàrem els canvis per raó de l'*splicing* alternatiu en l'isoforma més llarga; en segon lloc, alineàrem aquesta isoforma amb els corresponents duplicats de la família; després, mapejàrem les insercions i delecions trobades en els alineaments anteriors –a causa de la duplicació gènica- en l'isoforma més llarga; seguidament, per cada possible comparació entre els canvis d'*splicing* i els de duplicats, calculàrem el nombre de residus comuns i ho dividírem per la mida del canvi causat per l'*splicing* alternatiu (veure Figura 16); finalment, agrupàrem els resultats després de fer una correcció per redundància, a fi de no

sobreestimar els resultats de les famílies gèniques més nombroses.

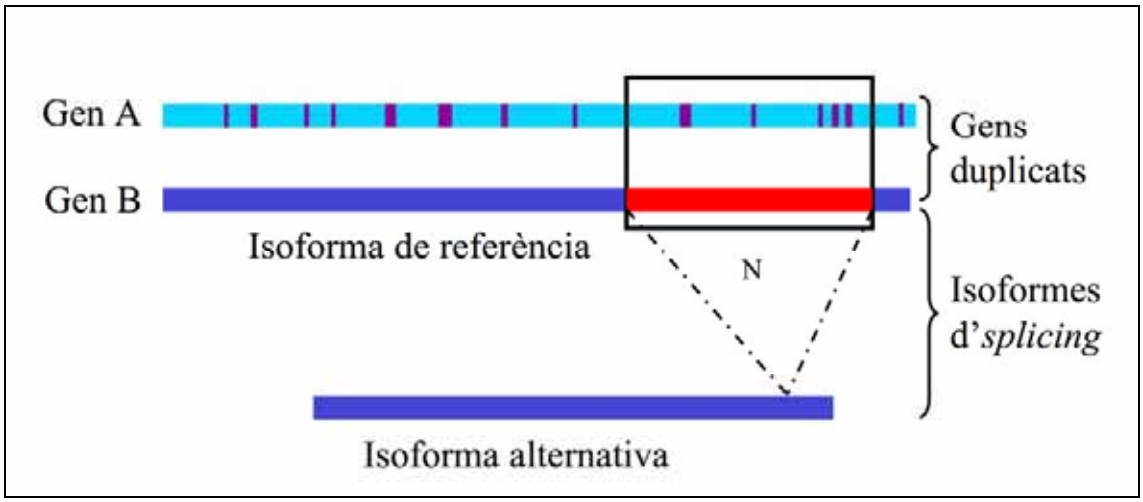

Figura 16 Solapament entre canvis. En la regió on hi ha la deleció d'*splicing* es compta quantes delecions apareixen entre els duplicats i es divideix per N.

La correcció de redundància utilitzada consistí en afegir un sol compte a l'histograma de freqüència quan els solapaments entre un canvi entre isoformes i una sèrie de canvis entre duplicats eren sempre el mateix.

3.11.9 Anàlisi de l'especificitat dels efectes de l'*splicing* alternatiu sobre l'estructura modular dels factors de transcripció

Analitzàrem l'especificitat de l'acció de l'*splicing* alternatiu sobre els dominis funcionals, és a dir, la relació entre els límits de les zones variables i les fronteres dels dominis funcionals. Per fer el càlcul, ordenàrem les quatre posicions –N i C terminal del domini, N i C terminal del canvi d'*splicing*- de manera ascendent i calculàrem quina fracció de tots els residus era compartida (veure Equació 5).

$$\textit{Especificitat} = \frac{D_i}{D_e} \quad \text{(Equació 5)}$$

on D_i és la distància entre les dues posicions internes del grup $-S_1$ i S_2- i D_e és la distància entre les dues posicions extremes $-M_1$ i M_2- (veure Figura 17).

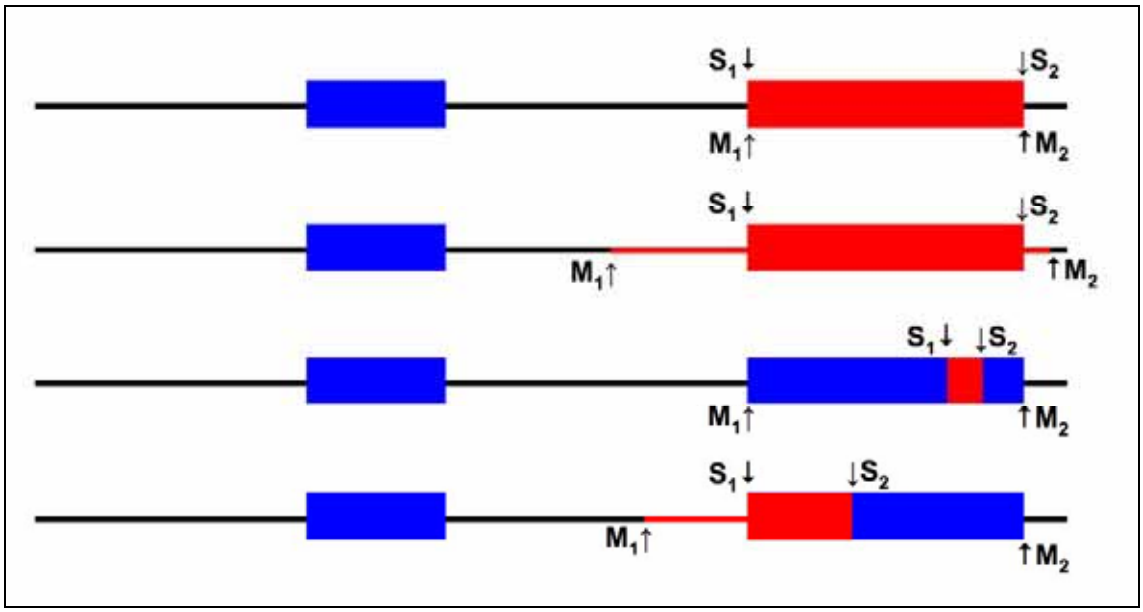

Figura 17. Càlcul de l'especificitat dels efectes de l'*splicing* alternatiu. Les capses representen dominis funcionals, mentre que el color vermell assenyala on hi ha la regió afectada per l'*splicing* alternatiu.

Una especificitat propera a 1 implica que l'efecte de l'*splicing* alternatiu se centra, principalment, en el domini funcional. Per la seva banda, si l'especificitat s'apropa a 0, això indica que l'*splicing* alternatiu modifica molts residus de les regions entre els dominis o que l'efecte sobre el domini funcional és molt petit.

RESULTATS I DISCUSSIÓ

4 *Splicing* alternatiu i duplicació gènica

4.1 Introducció

Hi ha moltes incògnites que envolten els efectes de l'*splicing* alternatiu, des de la viabilitat de les diverses isoformes (Lewis et al., 2003; Magen and Ast, 2005; Neu-Yilik et al., 2004) fins a la seva funció (Blencowe, 2006; Neu-Yilik et al., 2004). Una de les poques coses que es consideren segures és la gran importància que té en l'increment de la diversitat proteica (Brett et al., 2002; Graveley, 2001), però no és possible racionalitzar les conseqüències estructurals i funcionals en les proteïnes que provoquen els canvis en la seqüència (Lee and Wang, 2005).

Òbviament, l'*splicing* alternatiu no és l'únic mecanisme que incrementa la variabilitat del proteoma. Així, la duplicació gènica –seguida de la divergència entre els gens duplicats- és un mecanisme alternatiu que té molta importància (Chothia et al., 2003; Koonin et al., 2000; Lynch and Conery, 2000). L'estudi de la duplicació gènica està més avançat que el de l'*splicing* alternatiu i així, per exemple, se sap que les modificacions –les substitucions d'aminoàcids- que introdueix l'evolució en la seqüència dels diversos homòlegs no poden ser a l'atzar, sinó que estan sotmeses a una sèrie de restriccions estructurals, per permetre un plegament correcte i una funcionalitat plena (Chothia and Lesk, 1986). De la mateixa manera, s'ha vist que les insercions o delecions que s'observen entre els homòlegs tendeixen a ser a zones poc estructurades i superficials (Pascarella and Argos, 1992), on els canvis són menys lesius.

4.2 Intercanviabilitat com a font de diversitat proteica

Essent els dos fenòmens –*splicing* alternatiu i duplicació gènica- fonts de variabilitat del proteoma, és normal preguntar-se fins a quin punt les funcions de l'*splicing* alternatiu i la duplicació gènica se solapen. Els darrers anys s'han trobat diversos exemples en que les variants d'*splicing* o els gens duplicats semblen aportar la mateixa variabilitat (Altschmied, 2002; Dominguez et al., 2004; Lister et al., 2001; Pacheco et al., 2004) i que recolzarien la hipòtesi de què els dos fenòmens poden ser intercanviables. Un d'aquests és el cas del gen *U2AF1* d'humà i els gens *u2af1-α* i *u2af1-β* de *Fugu rubipres* (Pacheco et al., 2004). El gen humà dóna dues isoformes –U2AF^{35}a i U2AF^{35}b, que difereixen per només 7 residus del motiu de reconeixement de l'RNA-, mentre els

duplicats de Fugu no tenen *splicing* alternatiu. Quan es miren les seqüències proteiques, es veu com la diversitat proteica introduïda per l'*splicing* alternatiu en humans i la duplicació gènica a Fugu és la mateixa. Un altre exemple és el del gen *mitf*, duplicat al llinatge dels peixos i amb variants d'*splicing* en mamífers i aus (Lister et al., 2001). Sembla que després de la duplicació, els gens de peixos s'haurien especialitzat –cobrint les diverses funcions del gen ancestral- i haurien perdut l'*splicing* alternatiu (Altschmied, 2002).

A més a més, quan mirem el proteoma globalment, tant en les nostres dades (veure Figura 18) com en estudis recents (Kopelman et al., 2005; Su et al., 2006) s'observa una clara anticorrelació entre l'*splicing* alternatiu i la duplicació gènica. Així, els gens que no s'han duplicat tenen més *splicing* alternatiu que l'esperat, mentre les famílies gèniques més nombroses tenen un volum d'*splicing* alternatiu més baix del que haurien de tenir si la distribució fos a l'atzar.

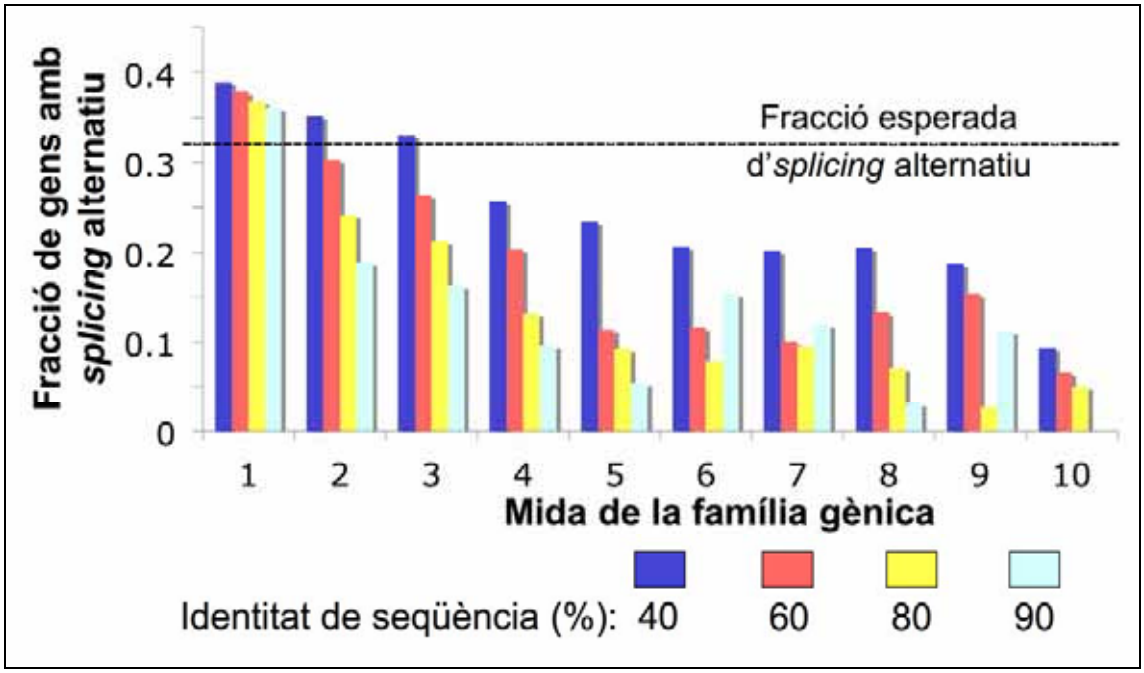

Figura 18. Anticorrelació entre *splicing* alternatiu i duplicació gènica. La gràfica mostra la fracció de gens que tenen *splicing* alternatiu segons la mida de la família gènica. Les famílies estan construïdes amb el programa CD-HIT (Li et al., 2001), utilitzant diversos llindars d'identitat de seqüència. La línia discontínua mostra la fracció de gens amb *splicing* alternatiu que s'esperaria si la distribució no estès esbiaixada. Resultats semblants s'obtenen utilitzant diferents bases de dades o els gens ortòlegs.

Tot això (anticorrelació a nivell genòmic i exemples específics a nivell proteòmic) suggereix que l'*splicing* alternatiu i la duplicació gènica podrien arribar a ser

intercanviables a l'hora d'introduir diversitat en el proteoma, és a dir, que els dos fenòmens podrien ser redundants com a font de variabilitat (Kopelman et al., 2005).

4.3 Anàlisi genòmica

Primer de tot, estudiàrem si la causa de la relació inversa entre els dos fenòmens podia tenir un origen genòmic –tant estructural com funcional. Per això analitzàrem l'estructura i la localització dels gens i les classes funcionals a les que pertanyen.

4.3.1 Estructura i localització dels gens

Pel que fa a l'estructura gènica, calculàrem el nombre d'exons per a cada gen. Analitzàrem tres grups de dades: els gens únics amb variants d'*splicing*, les famílies de gens duplicats sense *splicing* alternatiu i les famílies gèniques que tenen algun membre amb variants d'*splicing* (veure Figura 19).

Els nostres resultats mostren que l'estructura dels gens és diferent depenent de la via utilitzada per diversificar-se. Mentre que els gens sense *splicing* alternatiu tenen tendència a tenir un nombre d'exons menor al que tenen els gens amb variants d'*splicing*, els gens únics tenen una quantitat d'exons superior a la dels gens duplicats.

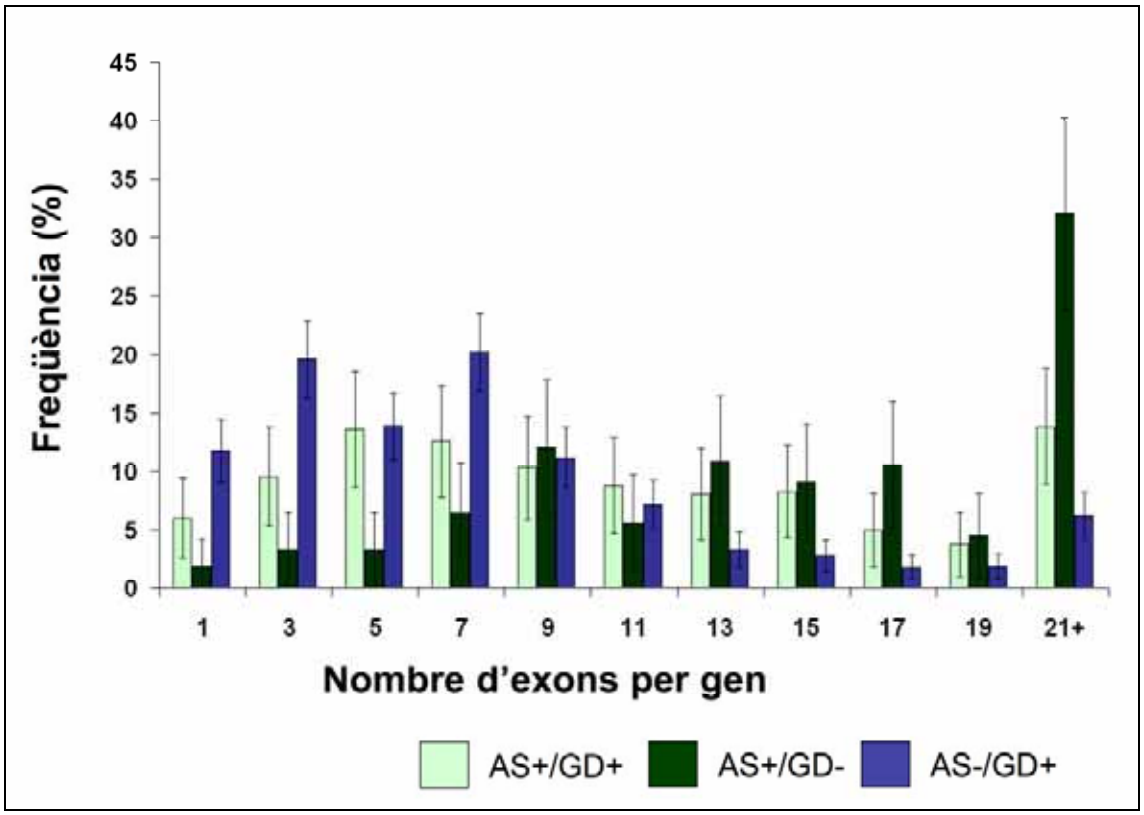

Figura 19. Estructura gènica de les famílies gèniques. La gràfica mostra la freqüència del nombre d'exons per gen. En verd clar, les famílies gèniques que tenen membres amb *splicing* alternatiu (AS+/GD+); en verd fosc, els gens únics amb variants d'*splicing* (AS+/GD-); en blau, les famílies gèniques amb una sola isoforma per gen (AS-/GD+).

La freqüència més gran de gens sense *splicing* alternatiu (AS-/GD+) en la primera barra de l'histograma –un o dos exons per gen- es podria atribuir parcialment al fenomen de la retrotransposició, és a dir, als gens duplicats per raó de la inserció al genoma d'un transcrit processat. Òbviament, aquests gens no tenen introns i, per tant, no poden fer *splicing*. No obstant això, aquest mecanisme ja havia estat descartat prèviament com a causa de l'anticorrelació (Kopelman et al., 2005) i, a més, només serviria per explicar un petit subconjunt de casos. Una altra explicació, més general, pel major nombre d'exons en els gens amb *splicing* alternatiu (AS+/GD+) és que molts d'aquests exons apareixen per duplicació exònica (Kondrashov and Koonin, 2001; Letunic et al., 2002).

Pel que fa a la localització dels gens duplicats, s'observa que en la majoria de famílies gèniques, independentment de la seva mida, la majoria dels seus membres tenen diferent localització cromosòmica (veure Figura 20), la qual cosa suggereix diferent regulació de l'expressió dels duplicats (Semon and Duret, 2006) i en dificulta la intercanviabilitat amb l'*splicing* alternatiu, perquè els dos fenòmens van molts cops lligats (Maniatis and

Reed, 2002; Maniatis and Tasic, 2002).

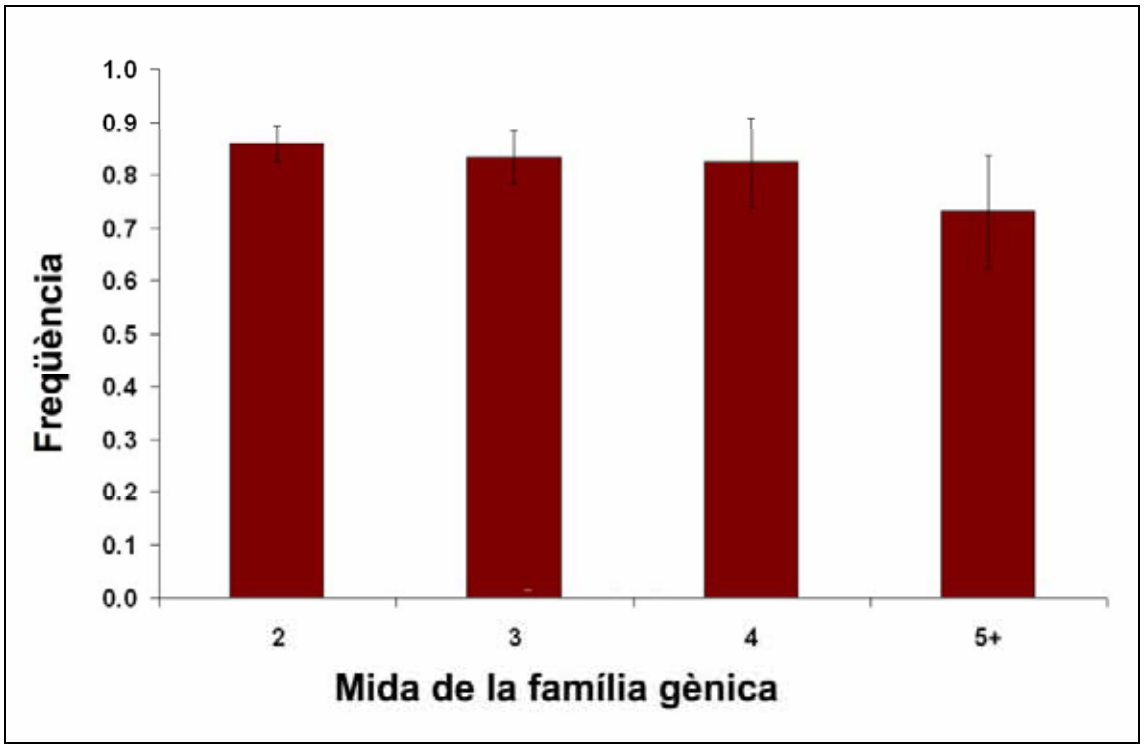

Figura 20. Freqüència de gens amb diferent localització cromosòmica. La gràfica mostra quina freqüència de gens tenen diferent localització dins de cada família. Es feren totes les comparacions dins de la família i les contribucions de cada comparació es corregiren tenint en compte el nombre de comparacions no redundants. Les dades estan agrupades per la mida de la família.

Així doncs, mentre sembla que hi ha una relació entre l'estructura gènica i la manera d'augmentar la variabilitat de la família gènica, la localització cromosòmica semblaria descartar-nos una regulació que permetés intercanviar els dos fenòmens com a fonts de diversitat.

4.3.2 *Splicing* alternatiu, duplicació gènica i funció

Podria ser que determinades famílies gèniques –lligades a una funció biològica concreta- mostressin una preferència per una via d'increment de la seva variabilitat, és a dir, podria ser que una de les raons de l'anticorrelació fos l'ús de l'*splicing* alternatiu o la duplicació gènica depenent de la funció dels gens. El que s'observà, però, és que els gens amb *splicing* alternatiu i els gens amb duplicats tenen distribucions similars entre la majoria de les categories funcionals.

Les proteïnes ribosomals i els receptors pertanyen a les famílies proteiques més nombroses en humans (Madera et al., 2004) i és conegut que els gens amb *splicing*

alternatiu no acostumen a tenir aquestes funcions (Neverov et al., 2005), cosa que suggereix que aquests grups de gens podrien ser els causants de l'anticorrelació. No obstant això, quan s'exclogueren aquestes dues famílies de l'anàlisi (Madera et al., 2004), es continuà trobant una relació inversa significativa (p-valor < 0.001) entre l'*splicing* alternatiu i la duplicació gènica.

Aquests resultats indiquen que l'anticorrelació entre aquests fenòmens –*splicing* alternatiu i duplicació gènica- no està relacionada amb una distribució diferencial de les funcions.

4.4 Anàlisi proteòmic

Per investigar si l'*splicing* alternatiu i la duplicació gènica són intercanviables des d'una perspectiva proteòmica –nivell estructural i funcional- vam comparar els canvis en la seqüència que introdueixen els dos fenòmens.

Encara que siguin tan diferents, des de punts de vista molecular i evolutiu, les modificacions introduïdes a nivell proteic són del mateix tipus: insercions/delecions i/o substitucions. Nosaltres caracteritzàrem aquests canvis mitjançant la seva mida, el tipus de residus afectats i la localització dins de la seqüència, ja que a partir d'aquestes variables es poden inferir característiques estructurals i/o funcionals (Shortle and Sondek, 1995).

Les dades de duplicació gènica s'obtingueren a partir de dos models: un basat en identitat de seqüència i l'altre, en dominis funcionals. Elegírem dues identitats de seqüència força diferents –40% i 80%-, perquè la força de l'anticorrelació sembla estar relacionada amb la identitat dins de les famílies (veure Figura 18). Les famílies amb una identitat mínima del 80% tenen un alt grau d'homologia estructural i funcional, mentre que les famílies amb identitats de fins al 40% permeten recuperar duplicats més vells o altament divergents però que encara poden tenir la mateixa funció (Tian and Skolnick, 2003; Wilson et al., 2000). Finalment, assenyalar que el model basat en famílies de dominis funcionals és un model més divergent que inclou casos més extrems de canvis de seqüència.

4.4.1 Insercions/delecions

Les insercions/delecions vam caracteritzar-les mirant la seva mida i la posició que tenen dins de la seqüència proteica. La primera propietat ens pot donar una idea de la

magnitud de l'efecte –més o menys dràstic-, mentre que la segona ens pot permetre inferir lleument algunes conseqüències estructurals. No obstant això, per tenir una idea clara dels efectes de les insercions/delecions cal tenir en compte les dues variables.

4.4.1.1 Mida

La mida de les insercions/delecions és un factor determinant a l'hora de valorar els canvis entre isoformes o duplicats. Així, s'ha d'esperar que els canvis de més llargària –ja siguin l'eliminació o l'addició d'un fragment- tinguin uns efectes estructurals i funcionals més grans que els de mida petita. Tanmateix, canvis petits poden tenir efectes importants tant a nivell d'estructura –per exemple, impedint la formació d'elements d'estructura secundària- com de funció –com ara, eliminant el residu catalític del centre actiu.

La Figura 21 mostra com la distribució de mides de les insercions/delecions en duplicats i en isoformes d'*splicing* alternatiu és molt diferent. Mentre la majoria de canvis introduïts per l'*splicing* tenen una mida que permetria extreure o afegir un domini funcional sencer, els canvis dels duplicats –qualsevol que sigui el model que mirem- són menors que 5 residus.

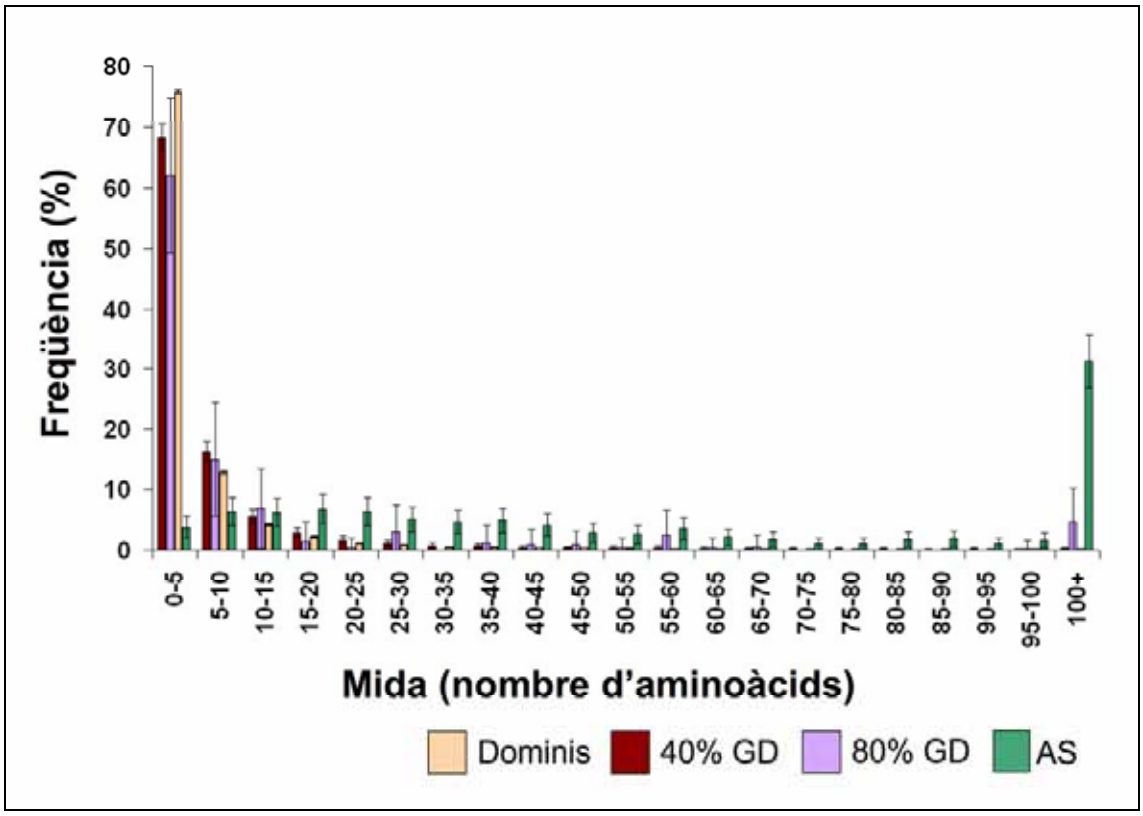

Figura 21. Mida de les insercions/delecions de les isoformes i els duplicats. S'utilitzaren diversos

models de famílies gèniques –identitats del 40% i 80% i famílies de dominis.

Posteriorment, vam refinar el nostre anàlisi separant les insercions/delecions que ocorren als extrems de la seqüència de les interiors. Això es féu perquè se sap que els extrems N- i C-terminal tendeixen a ser a la superfície de la proteïna (Hovmoller and Zhou, 2004); per tant, grans canvis afectant els extrems podrien ser més fàcilment acomodats i no estarien tan restringits estructuralment (Chothia and Lesk, 1986). Les Figures 20 i 21 mostren el resultat d'aquesta anàlisi. En ambdós casos les diferències entre les isoformes són molt més grosses que les que hi ha entre els duplicats gènics. Tal com esperàvem, tant en el cas dels duplicats com en el de les isoformes els canvis són més petits a l'interior de la seqüència (veure Figura 22) indicant l'efecte de les restriccions estructurals; d'aquesta manera, els canvis de mida més grossa entre les isoformes (>100 aminoàcids) es troben majoritàriament als extrems N- i C-terminal (veure Figura 23).

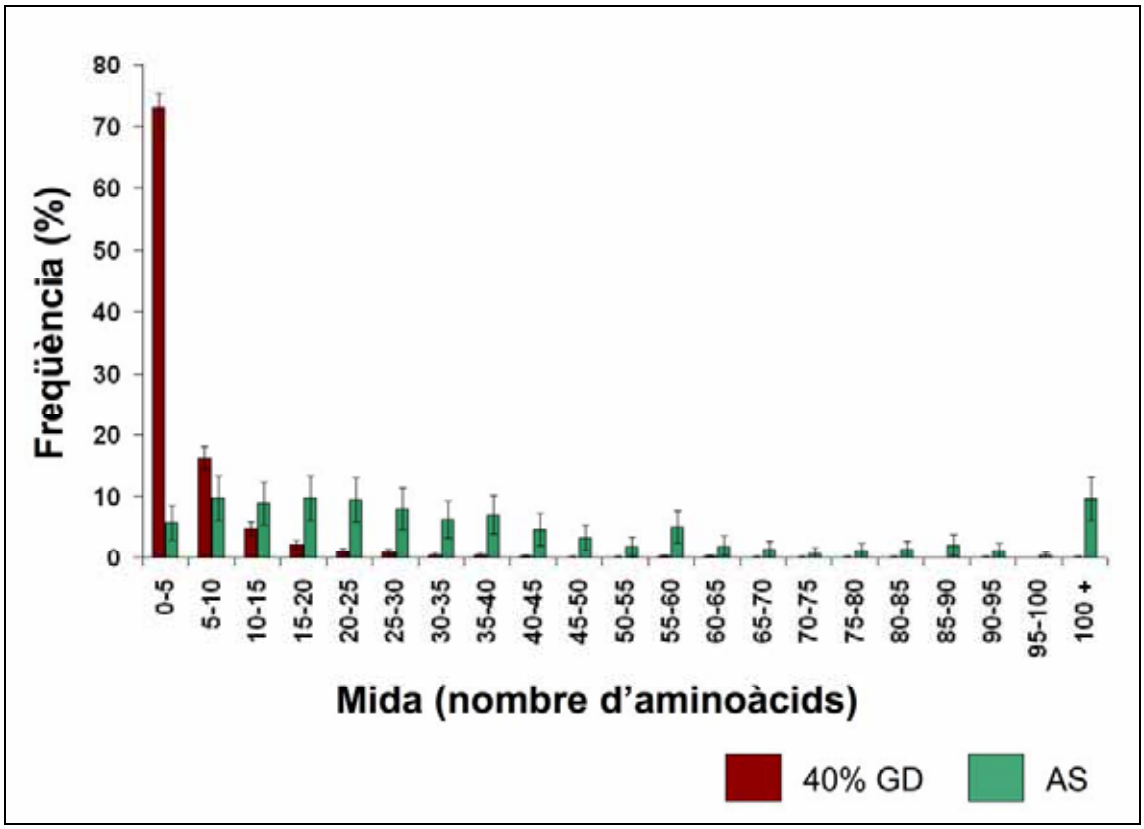

Figura 22. Mida de les insercions/delecions internes de les isoformes i els duplicats. S'utilitzaren famílies amb 40% d'identitat.

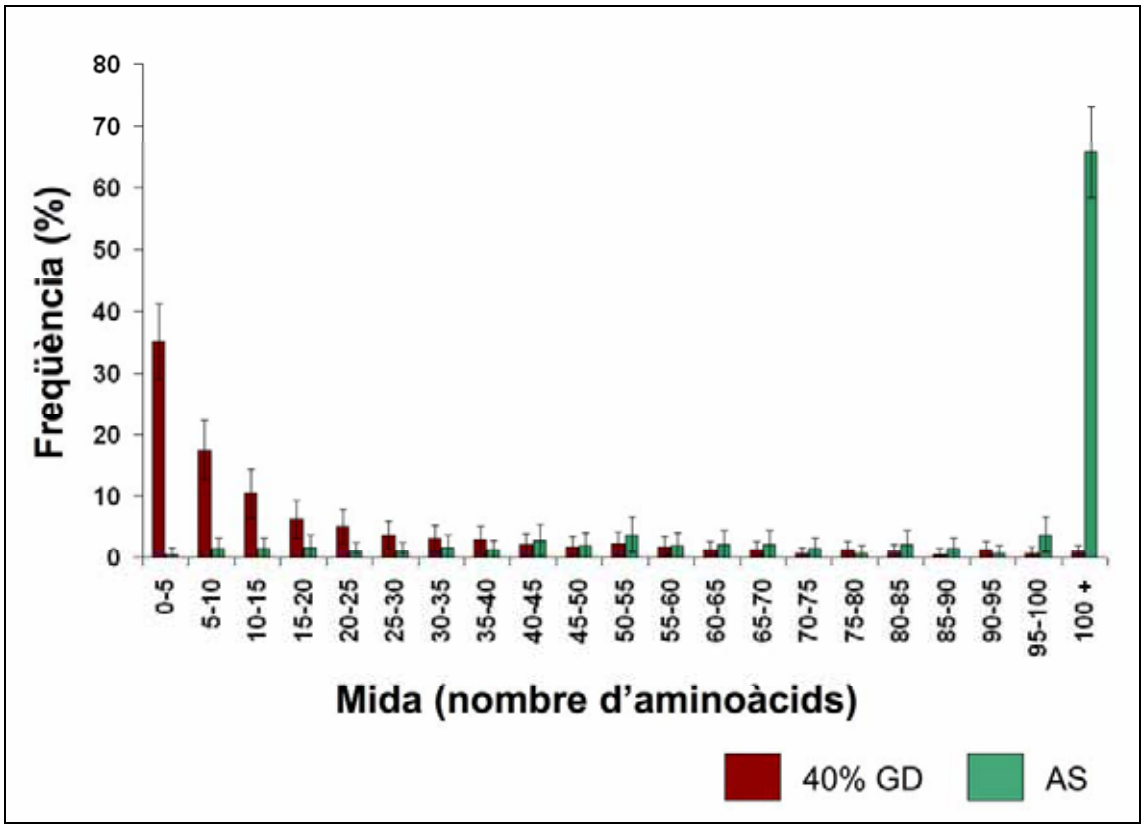

Figura 23. Mida de les insercions/delecions externes de les isoformes i els duplicats. S'utilitzaren famílies amb 40% d'identitat.

4.4.1.2 Posició relativa de les insercions/delecions

Amb l'objectiu de veure si els efectes de les insercions/delecions aparegudes en variants d'*splicing* i duplicats són similars, ens fixàrem en la seva posició dins de la seqüència. Com s'ha mencionat anteriorment, no només la mida del canvi és important, sinó que molts dels seus efectes depenen del lloc –l'entorn- on hi ha el canvi. Ho férem calculant la posició relativa dels canvis causats per la duplicació respecte a la posició dels canvis d'*splicing*, és a dir, calculàrem quin percentatge dels residus insertats o delecionats a les variants d'*splicing* també ho eren als duplicats del mateix gen (veure Figura 24). Aquesta mesura ens permet saber si les regions afectades pels canvis són les mateixes, fet que seria molt important ja que ens indicaria si els efectes estructurals i funcionals són comparables.

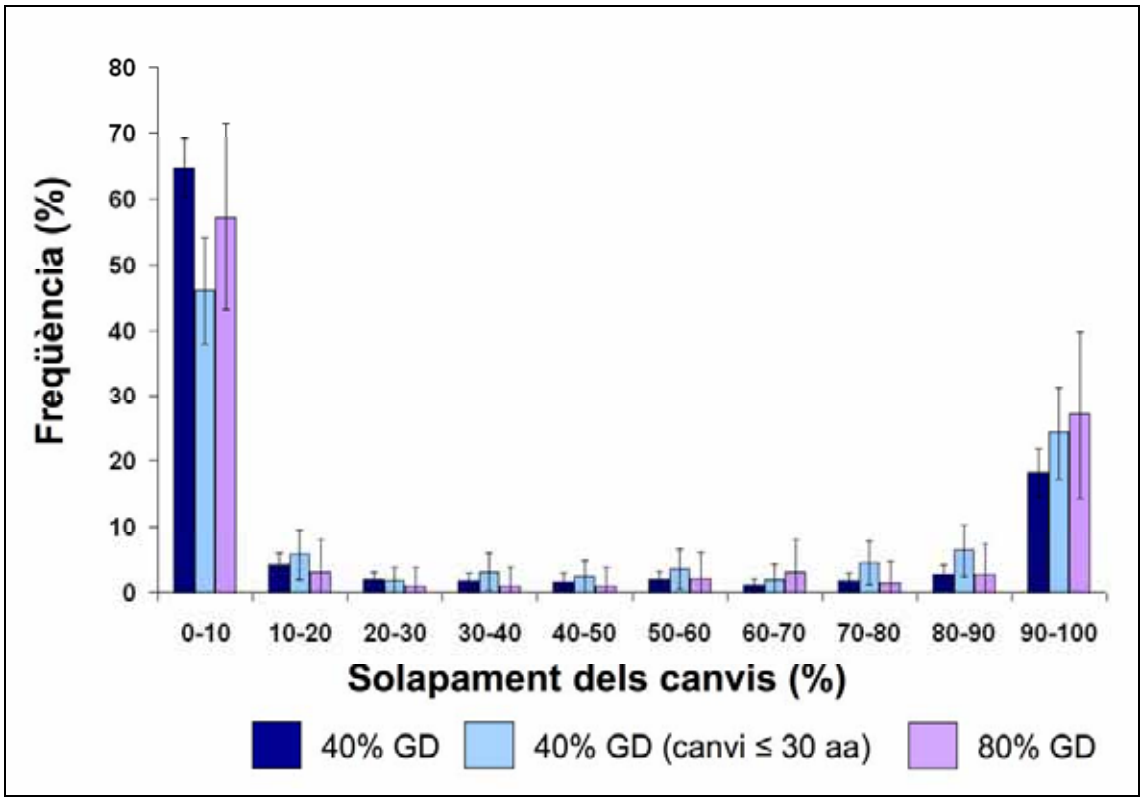

Figura 24. Solapament de les insercions/delecions de variants d'*splicing* i duplicats. Percentatge de residus d'una inserció/deleció causada per *splicing* alternatiu que pertanyen a un canvi del mateix tipus entre duplicats. S'utilitzaren famílies amb 80% i 40% d'identitat. De les darreres també s'analitzà el subconjunt de canvis de mida petita.

La majoria de les insercions/delecions causades per *splicing* alternatiu (~80%, en famílies duplicades al 40%) tenen un solapament exigu amb les que venen dels duplicats. Trobem un resultat semblant quan ens centrem en els canvis de mida petita (≤30 aminoàcids) o en les famílies al 80%.

Per tant, les insercions o delecions causades per l'*splicing* alternatiu o la divergència gènica no afecten, en general, les mateixes regions de la seqüència, cosa que ens fa pensar que els dos fenòmens tindran diferents efectes estructurals i funcionals. Amb tot, no podem oblidar que hi ha un petit percentatge de casos (~15%, en famílies duplicades al 40%) amb un solapament considerable, que després de mirar-los amb més cura, concloguérem que alguns d'aquests corresponen a canvis equivalents que semblen recolzar la hipòtesi de la intercanviabilitat com un mecanisme aplicable a un nombre petit però significatiu de proteïnes.

4.4.2 Substitucions

Les substitucions foren caracteritzades mitjançant la seva identitat de seqüència, tan globalment com localment, la natura físico-química dels canvis de residus i la distribució d'aquests canvis al llarg de la seqüència.

4.4.2.1 Identitats de seqüència global i local

La identitat global entre dues proteïnes ens dóna una idea del grau de conservació funcional (Tian and Skolnick, 2003). No obstant això, la identitat global no ens diu res del que passa localment, és a dir, proteïnes amb les mateixes identitats globals poden tenir una distribució de canvis completament diferent. Per això, també vam estudiar el que passava en determinades regions de la proteïna –fragments substituïts en l'*splicing* alternatiu i finestres de mides variables en gens duplicats.

La Figura 25 ens mostra les distribucions d'identitat global per les isoformes d'*splicing* i pels duplicats. En el cas de les isoformes, només es diferencien per la part substituïda; per tant, tal com esperàvem, el seu percentatge d'identitat és molt alt –majoritàriament, per sobre del 90%. En els cas dels duplicats, la seva distribució cobreix un interval més gran i depèn del llindar utilitzat a l'hora d'establir les famílies. Les famílies que tenen un 80% de residus idèntics són les que tenen una distribució més semblant a la de les isoformes. Finalment, si ens fixem en el model de dominis, podem veure que les seqüències han divergit molt.

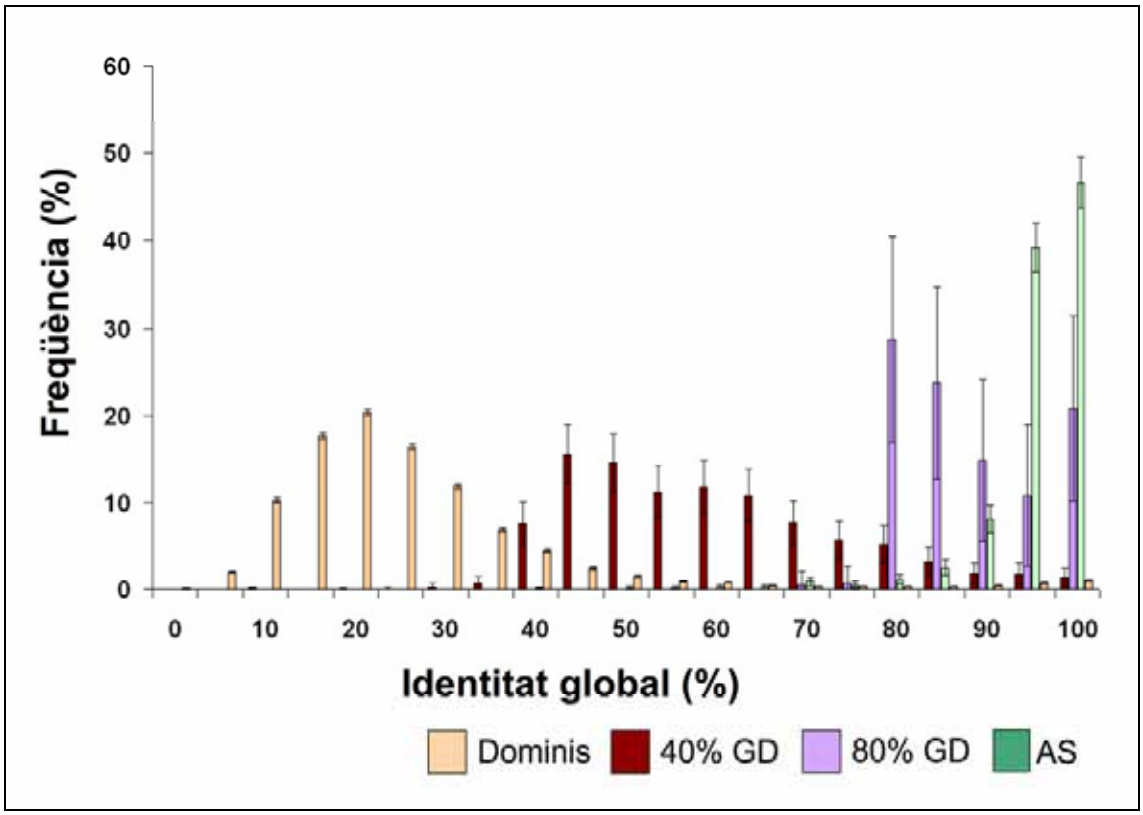

Figura 25. Identitat global dels alineaments entre isoformes o entre duplicats. S'utilitzaren diversos models de famílies gèniques –identitats del 40% i 80% i famílies de dominis.

Si ens fixem en la identitat local (veure Figura 26), els resultats que observem són força diferents. En el cas de les variants d'*splicing*, comparàrem les regions substituïdes. En canvi, pels duplicats calculàrem totes les possibles identitats locals dins de la família gènica –amb la mida dels canvis d'*splicing*- utilitzant la metodologia de la finestra lliscant (veure Materials i mètodes, secció 3.11.4.2). Les seqüències substituïdes en l'*splicing* alternatiu són molt diferents (20-30%), mentre que pels duplicats al 80% –que tenen una identitat global semblant a la de les isoformes (veure Figura 25)- la identitat local és molt superior. Això ens indica que tot i tenir un nombre de canvis comparables, l'*splicing* alternatiu els concentra més i als duplicats estan més esparsos. Els duplicats al 40% també tendeixen a ser localment més similars que les isoformes, però en aquest cas la diferència no és tan clara.

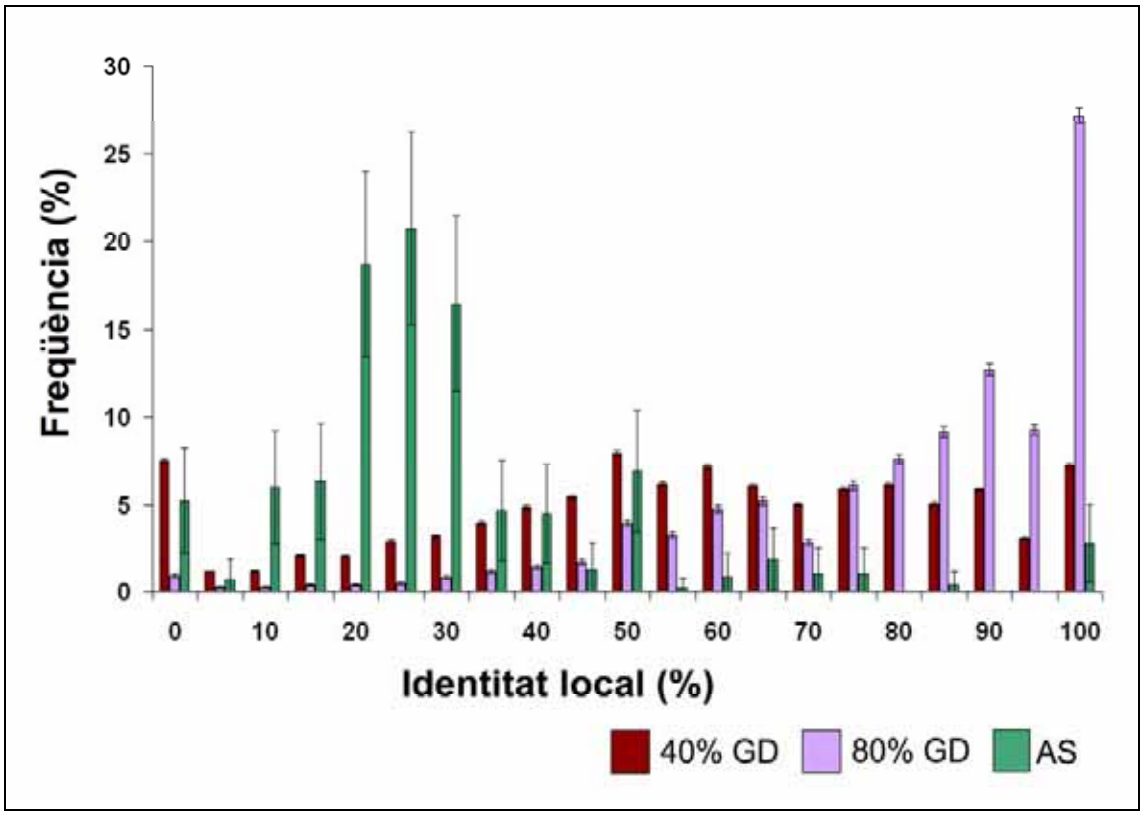

Figura 26. Identitat local dels alineaments entre fragments substituïts de les isoformes o entre regions dels duplicats. S'utilitzaren famílies amb 40% i 80% d'identitat.

Per completar aquesta anàlisi vam comparar les regions substituïdes en l'*splicing* alternatiu amb les regions equivalents als duplicats. En la majoria dels casos (78% pels duplicats al 80%; 64% pels duplicats al 40%), la identitat local dels duplicats és més alta que la de les isoformes (veure Figura 27).

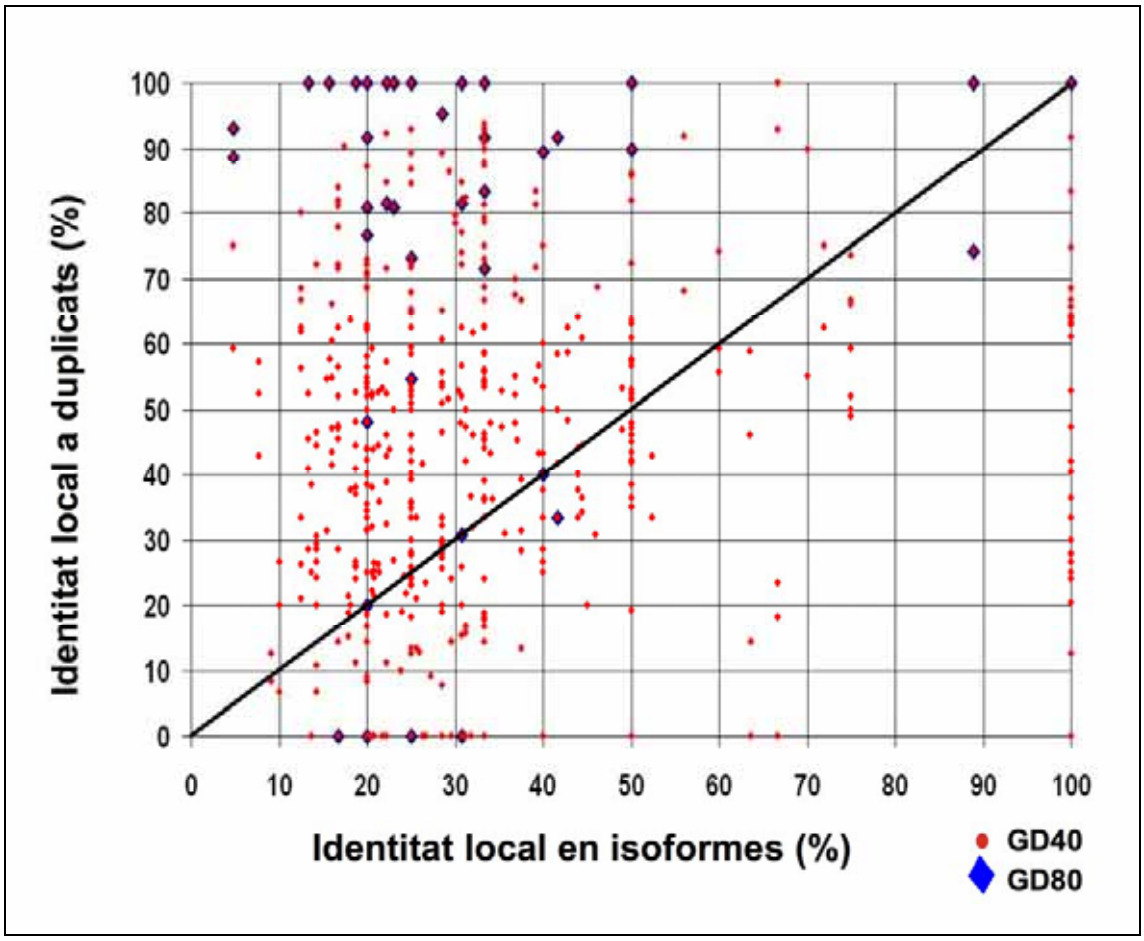

Figura 27. Identitat local de la mateixa regió en variants d'*splicing* i duplicació gènica. Els rombes blaus corresponen a les famílies al 80%; els punts rojos, a les famílies al 40%.

4.4.2.2 Natura i distribució dels canvis

Havent vist que l'*splicing* alternatiu tendeix a generar zones amb més substitucions d'aminoàcids que la duplicació gènica, decidírem estudiar la natura d'aquests canvis. Per fer-ho, ens centrarem en els canvis d'aminoàcids no conservatius. Aquestes substitucions confereixen canvis en les propietats físico-químiques de la proteïna i tenen més probabilitat d'estar implicades en alteracions estructurals i/o funcionals (Ferrer-Costa et al., 2002; Topham et al., 1997).

Veiérem que la quantitat relativa de canvis no conservatius és més alta en l'*splicing* alternatiu que en la duplicació gènica, tant si ho mirem en les famílies al 80% –58% i 44%-, com en les famílies al 40% –60% i 43%. Aquests canvis són estadísticament significatius (test t-Student; p-valor < 0.05 en ambdós casos).

Finalment, mesurarem la concentració d'aquests canvis més dràstics al llarg de la seqüència de la proteïna. Amb aquest objectiu calcularem la distància entre les

substitucions no conservatives més allunyades en la seqüència. Per descartar biaxos causats per la quantitat de canvis d'*splicing*, dividírem les dades en diversos subconjunts atenent al nombre de regions substituïdes per raó de l'*splicing* alternatiu. És a dir, vam creure que no tenia sentit analitzar les dades conjuntament perquè en aquest cas la mida i nombre dels exons podia afectar molt els nostres resultats.

En el subconjunt de dades amb una sola substitució per *splicing*, mesuràrem la distància màxima entre canvis –la distància entre els canvis d'aminoàcids més allunyats. La Figura 28, d'acord amb els resultats previs, mostra que les distribucions per *splicing* alternatiu i per duplicació gènica són molt diferents. Així, mentre els canvis no conservatius en les isoformes amb una sola substitució estan concentrats en una porció molt petita de la proteïna, les substitucions entre els corresponents duplicats estan espargides al llarg de tota la seqüència.

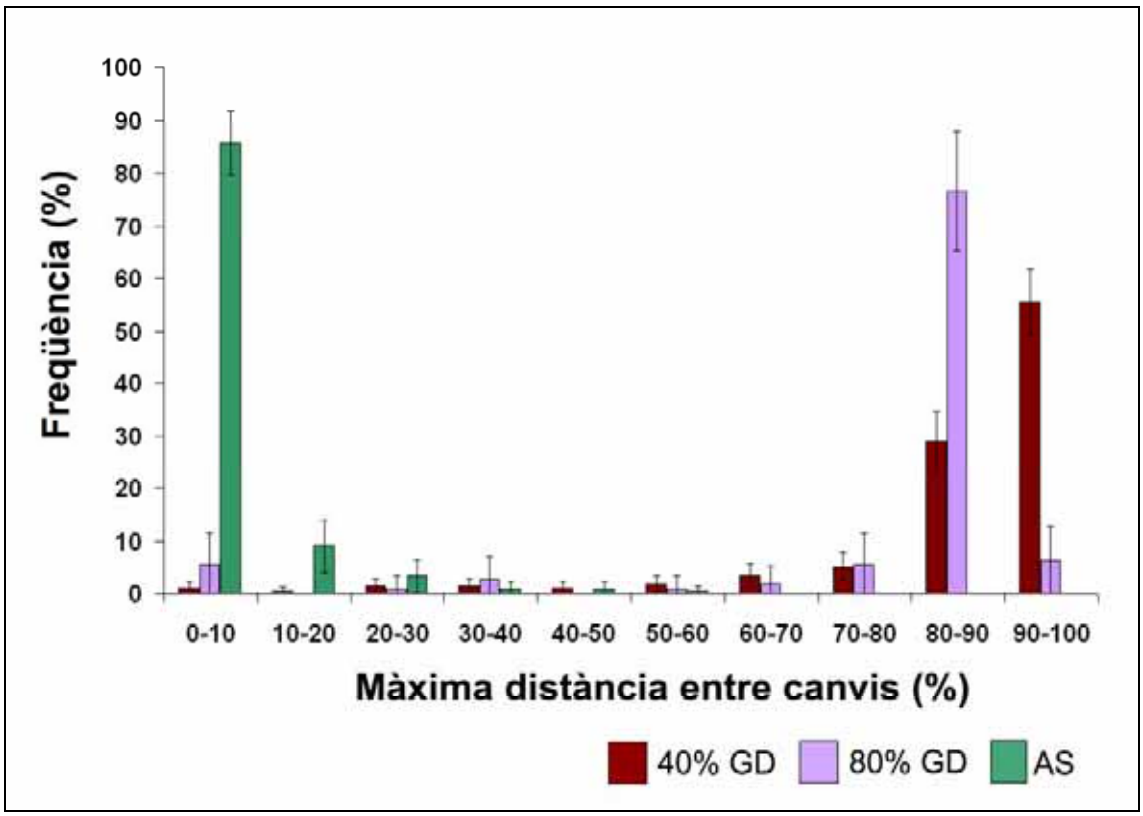

Figura 28. Distància màxima entre canvis no conservatius. La distància fou normalitzada dividint per la mida de la proteïna. S'utilitzaren famílies amb 40% i 80% d'identitat.

Pel que fa als casos amb dues substitucions en les variants d'*splicing*, ens interessàrem per la distància que hi ha entre les dues regions variants. Per això, calculàrem el quocient entre la distància màxima entre substitucions contigües de residus i la distància

entre els canvis més allunyats (veure Materials i mètodes, secció 3.11.7, per una explicació més completa). A causa de la menor quantitat de dades, trobàrem preferible fer un test estadístic per comparar els dos conjunts de dades, enlloc d'un histograma de freqüències amb intervals de confiança. Com no sabíem a priori quina distribució tenien les dades, elegírem el test KS. Tant quan analitzàrem totes les substitucions com quan ens centràrem en les no conservatives, trobàrem diferències entre les mostres de duplicats i variants d'*splicing* –$D = 0.80$ ($p < 0.05$) i $D = 0.65$ ($p < 0.05$) per a tots el canvis i per als canvis no conservatius, respectivament.

4.5 Discussió

Com hem vist, l'*splicing* alternatiu i la duplicació gènica estan inversament relacionats a nivell genòmic. Aquesta anticorrelació entre els dos fenòmens (veure Figura 18) (Kopelman et al., 2005; Su et al., 2006) es manté per diferents graus de divergència de les seqüències proteiques i pot tenir relació amb l'estructura dels gens. En canvi, sembla ser independent de les funcions de les proteïnes. Així doncs, la hipòtesi que suggereixen aquests resultats és l'existència d'intercanviabilitat funcional entre els dos fenòmens (Kopelman et al., 2005) –*splicing* alternatiu i duplicació gènica-, que a nivell de proteïnes recolzarien diversos exemples (Altschmied, 2002; Dominguez et al., 2004; Lister et al., 2001; Pacheco et al., 2004).

D'aquesta manera, l'*splicing* alternatiu i la duplicació gènica, tot i ser tan diferents, jugarien el mateix rol com a promotors de la variabilitat proteica i l'explicació més plausible per aquesta redundància seria la de l'existència de restriccions estructurals per mantenir la viabilitat funcional dels productes proteics (Chothia and Lesk, 1986).

Per contra, les nostres anàlisis de les seqüències proteiques ens indiquen clarament que la variabilitat introduïda per l'*splicing* alternatiu i la duplicació gènica és tan diferent que, en general, no es poden considerar els dos fenòmens com a redundants. La Taula 3 resumeix els resultats obtinguts.

		Variants d'*splicing*	Duplicats gènics
Substitucions	Identitat de seqüència	Global: alta Local: baixa	Tant la global com la local depèn del llindar utilitzat
	Natura dels canvis d'aminoàcids	Majoritàriament, no conservatius	Majoritàriament, conservatius
	Distribució dels canvis	Agrupats en regions petites que acostumen a estar força separades	Esparsos per tota la seqüència
Insercions/delecions	Mida quan els canvis passen als extrems N o C terminal	Molt llargs (més de 100 residus)	Curts (menys de 15 residus)
	Mida quan els canvis passen entre els extrems	Mitjans o llargs (més de 15 residus)	Molt curts (menys de 5 residus)
	Localització	Normalment no se solapen massa amb els canvis dels duplicats	Preferencialment, a zones exposades (Pascarella and Argos, 1992)

Taula 3. Efectes de l'*splicing* alternatiu i la duplicació gènica.

Finalment, hem de tenir en compte que les diferències a nivell de la seqüència proteica tenen la seva traducció en diferències a nivell de l'estructura terciària. La Figura 29 exemplifica aquesta conseqüència en un cas especialment rellevant (família de les MAP quinases) –els canvis entre les variants d'*splicing* són pocs, dràstics i concentrats, mentre que els duplicats tenen molts més canvis, esparsos i més conservatius. Així, llevat d'alguns casos particulars, podem afirmar que la majoria d'efectes sobre

l'estructura i la funció no poden ser intercanviables.

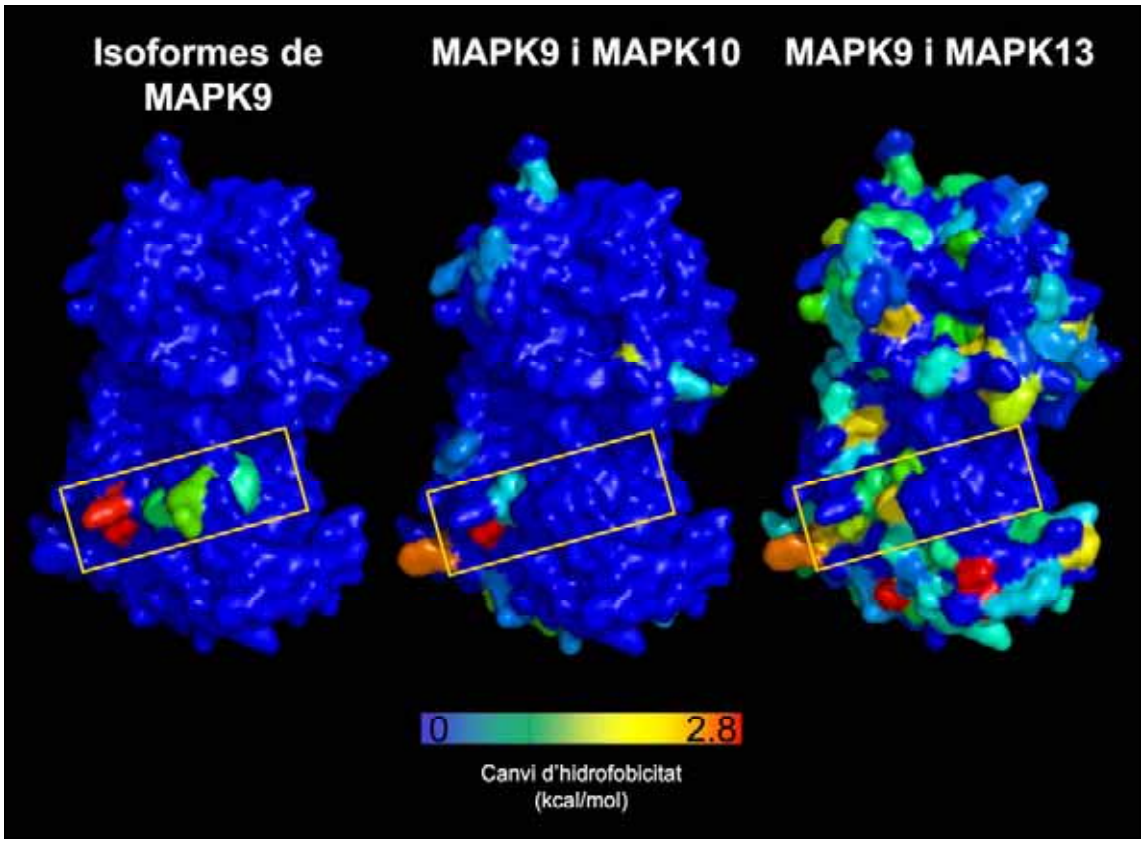

Figura 29. Efectes de l'*splicing* alternatiu i la duplicació gènica. Canvi d'hidrofobicitat entre les variants d'*splicing* de MAPK9 i entre MAPK9 i els seus duplicats MAPK10 (84% d'identitat) i MAPK13 (46% d'identitat). Com a mesura d'hidrofobicitat utilitzem l'energia lliure de la transferència dels aminoàcids d'aigua a octanol (Fauchere et al., 1988), assignant a cada canvi d'aminoàcid el valor absolut de la diferència L'escala va del blau (sense canvi) al vermell (màxim canvi). Per facilitar la comparació, la regió substituïda en l'*splicing* alternatiu apareix marcada amb un marc daurat en totes les estructures.

Així doncs, ens trobem davant d'una paradoxa aparent: la relació existent entre l'*splicing* alternatiu i la duplicació gènica és contradictòria depenent del nivell al qual la mirem. Per intentar explicar-la, nosaltres proposem fixar-nos en la història evolutiva de les famíles gèniques i en el concepte de l'equilibri de la dosi gènica.

Tal com em vist, la majoria de gens duplicats tenen una localització diferent, a partir de la qual s'infereix una regulació gènica diferent (Semon and Duret, 2006); per tant, els diversos productes proteics podrien interferir els uns amb els altres, afectant la funció biològica original. Així, per mantenir estable la dosi gènica (Papp et al., 2003), pot ser que la duplicació dels gens estigui desafavorida o es pot donar el cas que una de les còpies sigui silenciada després de la duplicació (Lynch and Conery, 2000).

Evidentment, aquest punt va en contra de qualsevol duplicació gènica, però no ens pot passar per alt que és probable que la pressió selectiva sigui més forta en el cas dels gens amb *splicing* alternatiu que en el dels gens que tan sols donen un producte proteic.

Un cas en que la duplicació i posterior retenció dels gens pot resultar molt beneficiosa per a l'organisme és quan aquesta afecta als gens essencials, ja que pot ser un sistema de seguretat (Shakhnovich and Koonin, 2006). En canvi, la introducció de canvis en el patró d'*splicing* dels gens essencials no es pot utilitzar per aportar aquest plus de seguretat a l'organisme que li confereix la duplicació.

Finalment, si un gen amb *splicing* alternatiu es duplica, la pèrdua de variants d'*splicing* en un dels duplicats, o en ambdós, pot ser assumida perquè hi ha altres versions idèntiques de la isoforma. Aquest fet, que és una extensió del sistema de seguretat dels paràlegs (Kafri et al., 2005), queda reforçat pel fet que l'anticorrelació és més forta com més semblants són les proteïnes (veure Figura 18) (Kopelman et al., 2005; Su et al., 2006) i el descobriment que immediatament després de la duplicació es perden moltes variants d'*splicing* (Su et al., 2006).

En conclusió, una combinació de tots aquests efectes ha de donar, com a resultat general, una fracció menor de gens amb *splicing* alternatiu en les famílies gèniques més nombroses. Aquest efecte seria més dramàtic en els duplicats recents, mentre que la divergència posterior hauria de permetre un ressorgiment de noves variants d'*splicing*.

5 *Splicing* alternatiu en un context evolutiu

5.1 Introducció

Havent establert que la contribució de l'*splicing* alternatiu a la diversitat del proteoma és, en general, diferent a la que aporta la duplicació gènica, ens centrarem a analitzar la conservació d'aquesta contribució en diverses espècies animals.

La presència d'*splicing* alternatiu en els diferents organismes i la seva conservació interespecífica van començar a ser temes de gran interès arrel de la troballa de menys gens dels esperats en la seqüenciació del genoma humà (IHGSC, 2001; Venter, 2001). Llavors, se suggerí que la fracció de gens amb *splicing* alternatiu o la quantitat d'isoformes podrien explicar les diferències fenotípiques entre els organismes (IHGSC, 2001). Malgrat tot, a dia d'avui no hi ha resultats concloents en aquesta línia d'estudi, car s'han trobat resultats oposats –descartant (Brett et al., 2002) o recolzant (Kim et al., 2004) la hipòtesi sobre el paper de l'*splicing* alternatiu en l'especiació.

D'altra banda, el que sí s'ha vist és que hi ha diversos gens en humans i ratolins que tenen *splicings* alternatius conservats a nivell genòmic (Sugnet et al., 2004; Thanaraj et al., 2003) i, fins i tot, exemples específics de conservació en altres espècies (Bomze and Lopez, 1994). No obstant això, hi ha qui dubta de que la conservació sigui global i mostra que hi ha moltes isoformes que són específiques de les espècies (Nurtdinov et al., 2003).

Molt menys clara és encara la possible conservació del fenomen de l'*splicing* alternatiu a nivell proteic, és a dir, fins a quin punt les diferents espècies pateixen els mateixos efectes proteics per raó de l'*splicing* alternatiu. Així, saber si diferents organismes utilitzen els mateixos mecanismes per modular la funció fou el següent tema que centrà el nostre interès.

5.2 Anàlisi dels mecanismes de modulació funcional en proteïnes

Les diverses isoformes d'una proteïna poden tenir diferents substractes o afinitats (Zarich et al., 2000), diferents constants d'activitat o, fins i tot, efectes antagònics (Taylor et al., 1999); per tant, es pot veure l'*splicing* alternatiu com un mecanisme per modular la funció proteica. Les diferències funcionals provenen dels canvis que hi ha a

nivell de la seqüència que porten a canvis estructurals (Garcia et al., 2004) i de les propietats físico-químiques.

En primer lloc, caracteritzàrem a nivell de proteïna els efectes de l'*splicing* alternatiu en quatre espècies diferents –humà, ratolí, rata i mosca del vinagre-, mitjançant diverses propietats, tant de seqüència com estructurals. Posteriorment, comparàrem els histogrames de freqüència de les diverses propietats obtinguts amb les proteïnes humanes amb els de les altres espècies determinant el seu grau de semblança o diferència.

En el cas dels esdeveniments d'*splicing* alternatiu consistents en insercions/delecions, ens fixàrem en la seva mida –normalitzada utilitzant les longituds d'exons més freqüents per a cada espècie segons Deutsch i Long (Deutsch and Long, 1999)-, la composició aminoacídica de les insercions de longitud menor o igual a 30 residus, l'accessibilitat al solvent del fragment insertat/delecionat i la seva estructura secundària. Totes aquestes propietats ens permeten relacionar els canvis de seqüència causats per l'*splicing* alternatiu amb el seu possible efecte estructural i/o funcional (Kim et al., 2001; Shortle and Sondek, 1995; Sondek and Shortle, 1990).

La Taula 4 mostra els resultats pel solapament entre les distribucions per les variables utilitzades en el cas de les insercions/delecions. Per la majoria de variables, el solapament entre la superfície de les distribucions és superior al 90%. Només la mida de les insercions/delecions sembla ser una mica més divergent, quedant per sota del 80% en dues espècies. Addicionalment, es calculà l'interval de confiança al 95% per a cada barra de l'histograma i en la majoria dels casos es veu solapament entre els intervals de les diverses espècies (veure Figura 30). Una altra conclusió important és que no sembla que les diferències estiguin relacionades amb la distància evolutiva entre els organismes: els resultats per a la comparació entre humà i rata són, alguns cops, més semblants als del ratolí, mentre que altres són similars als de la mosca del vinagre.

	Ratolí	Rata	Drosophila
Mida	90%	78%	79%
Composició	97%	95%	89%
Accessibilitat	98%	98%	96%
Estructura secundària	98%	93%	93%

Taula 4. Solapament de les distribucions per insercions/delecions. Percentatge de solapament de les distribucions humanes amb les distribucions de ratolí, rata i Drosophila.

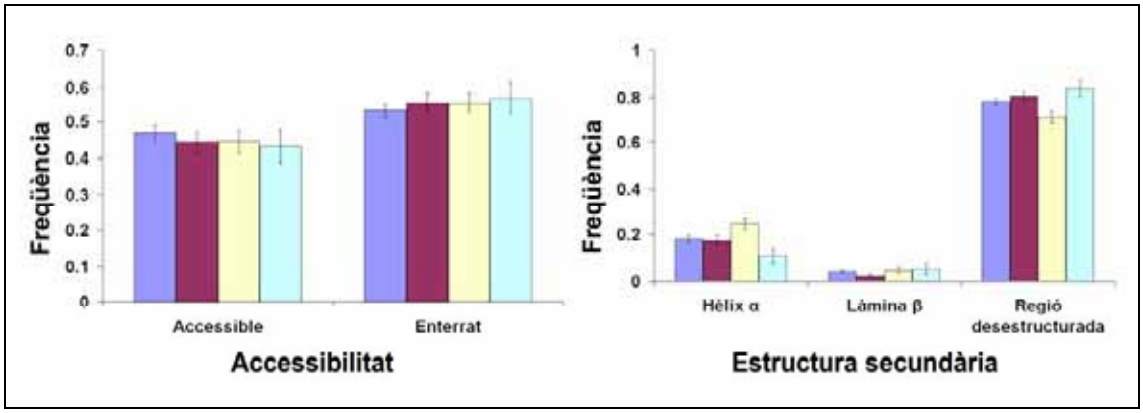

Figura 30. Accessibilitat i estructura secundària de les insercions/delecions. En blau fosc, humans; en morat, ratolins; en groc, rates; en blau cel, mosca del vinagre.

Pel que fa a les substitucions, analitzàrem la similitud de seqüència entre els fragments substituïts i els canvis de mida que provoquen a les proteïnes. La similitud local es calculà com la puntuació de l'alineament entre els dos fragments normalitzada per la seva longitud. Les dues variables mesuren diferents aspectes del mecanisme de modulació funcional: la similitud de seqüència ens dóna una idea de la natura físico-química del canvi, mentre la diferència de mida ens pot donar una idea de si hi pot haver canvis en el plegament de la proteïna; així, si els dos fragments tenen una mida similar, els canvis estructurals i funcionals entre les isoformes dependran, probablement, de la similitud entre els fragments substituïts; en canvi, si un dels fragments és molt més llarg que l'altre, l'efecte pot ser semblant al de les insercions/delecions.

La Taula 5 conté els resultats en el cas de les substitucions. De manera semblant a

l'exposat més amunt, tant el solapament de la superfície de les distribucions de similitud de seqüència com el de les de canvi de mida se situa al voltant del 90% per totes les espècies, independentment de la distància evolutiva.

	Ratolí	Rata	Drosophila
Similitud	93%	92%	90%
Canvi de mida	91%	97%	88%

Taula 5. Solapament de les distribucions per substitucions. Percentatge de solapament de les distribucions humanes amb les distribucions de ratolí, rata i Drosophila.

En aquest cas, també es trobà solapament en la majoria d'intervals de confiança de les barres de l'histograma.

5.3 Anàlisi de parells equivalents

Després d'aquest anàlisi global, seleccionàrem un subconjunt de dades que contenia parells d'esdeveniments equivalents –amb el mateix rol biològic (veure Figura 31)-, entenent com a esdeveniment d'*splicing* alternatiu dues isoformes de la mateixa proteïna. Segons el nostre criteri, per a què l'esdeveniment sigui equivalent, les isoformes homòlogues han de tenir el mateix rol, ja sigui perquè s'expressen en el mateix teixit o estadi de desenvolupament o perquè tenen, potencialment, la mateixa funció bioquímica. Un exemple del darrer cas és el de la proteïna homeobox Hox-A1: tant en humans com en ratolins hi ha isoformes amb i sense el domini homeobox (Hong et al., 1995; LaRosa and Gudas, 1988), la qual cosa fa pensar que les isoformes poden tenir un paper semblant.

Les dades –parells d'esdeveniments equivalents- s'obtingueren utilitzant un protocol inspirat en observacions de Kondrashov i Koonin (Kondrashov and Koonin, 2003).

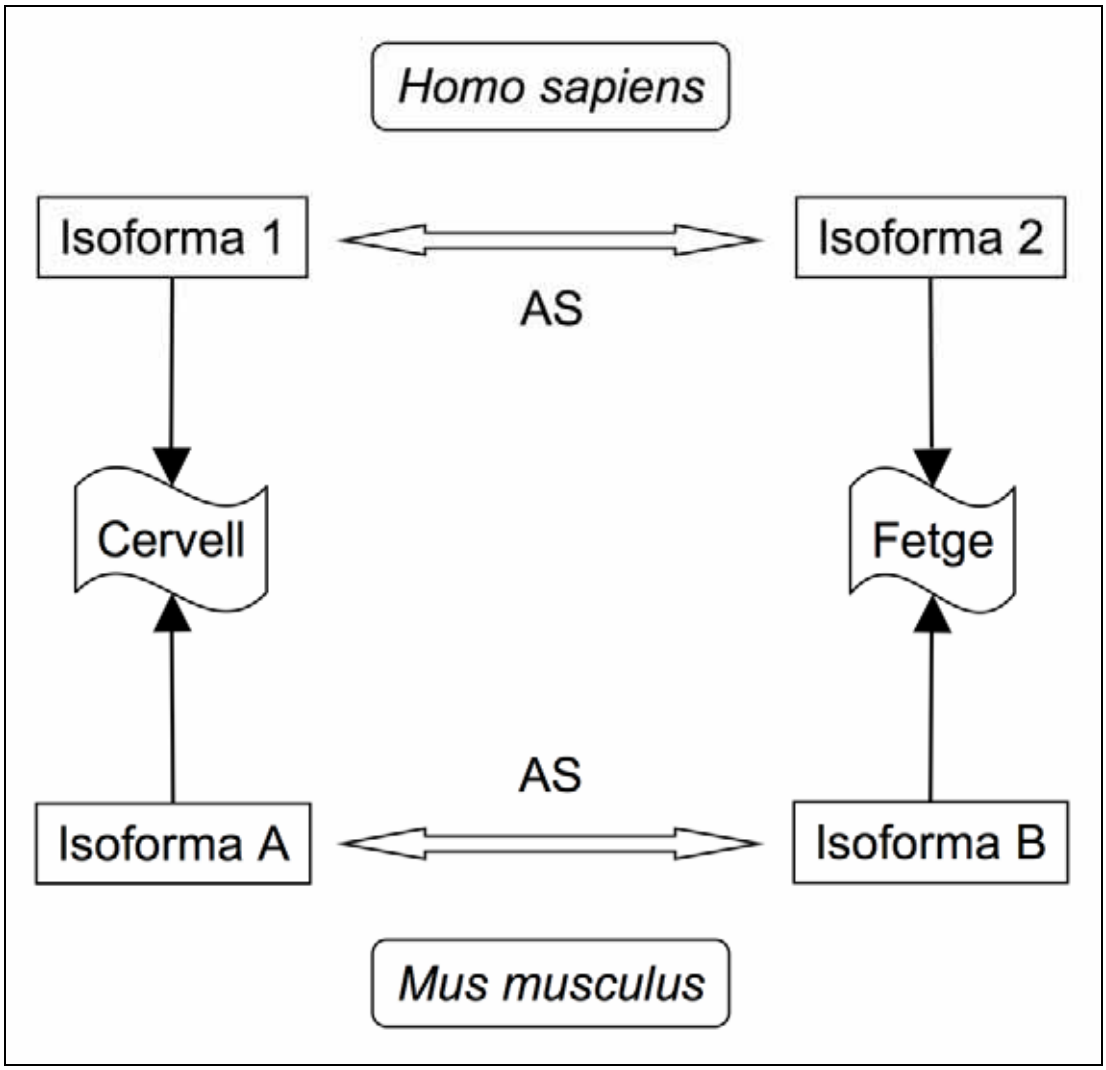

Figura 31. Exemple d'esdeveniments equivalents. En aquest cas, una proteïna humana i una de ratolí tenen la mateixa funció bioquímica (provenen de gens ortòlegs) i ambdues tenen diverses isoformes amb especificitat tissular: la isoforma 1 humana i la isoforma A de ratolí són equivalents perquè s'expressen al cervell; mentrestant, la isoforma 2 humana i la isoforma B murina s'expressen ambdues al fetge. L'equivalència particular de les isoformes comporta l'equivalència de l'esdeveniment d'*splicing* alternatiu definit com els canvis entre les isoformes 1 i 2 en l'humà i les isoformes A i B en el ratolí.

Aquesta anàlisi es féu comparant proteïnes humanes i murines, car no hi havia dades per utilitzar la mosca del vinagre. La Taula 6 mostra alguns exemples d'esdeveniments d'*splicing* alternatiu equivalents.

Gen	**Espècie**	**Esdeveniments d'*splicing* alternatiu (nom de les isoformes) i lloc d'expressió**	
AP2A1	Humà	A	B
		Teixit neuronal i múscul esquelètic (Scorilas et al., 2002)	Ubiqua (Scorilas et al., 2002)
	Ratolí	A	B
		Teixit neuronal i múscul esquelètic (Ball et al., 1995)	Ubiqua (Ball et al., 1995)
CLTA	Humà	Brain	Non brain
		Cervell (Jackson and Parham, 1988)	No trobada al cervell (Jackson and Parham, 1988)
	Rata	Brain	Non brain
		Cervell (Jackson and Parham, 1988; Kirchhausen et al., 1987)	No trobada al cervell (Jackson and Parham, 1988; Kirchhausen et al., 1987)
CLTB	Humà	Brain	Non brain
		Cervell (Jackson and Parham, 1988)	No trobada al cervell (Jackson and Parham, 1988)
	Rata	Brain	Non brain
		Cervell (Jackson and Parham, 1988; Kirchhausen et al., 1987)	No trobada al cervell (Jackson and Parham, 1988; Kirchhausen et al., 1987)

ECE1	Humà	B	A
		Fetge (Schweizer et al., 1997; Valdenaire et al., 1999; Valdenaire et al., 1995)	Múscul llis (Schweizer et al., 1997; Valdenaire et al., 1999; Valdenaire et al., 1995)
	Rata	B	A
		Fetge (Valdenaire et al., 1999)	Múscul llis (Valdenaire et al., 1999)
HMBS	Humà	Non erythropoietic	Erythropoietic
		Teixits no eritropoiètics (Grandchamp et al., 1987)	Teixits eritropoiètics (Grandchamp et al., 1987)
	Ratolí	Non erythropoietic	Erythropoietic
		Teixits no eritropoiètics (Beaumont et al., 1989)	Teixits eritropoiètics (Beaumont et al., 1989)
GCK	Humà	1	2
		Pàncreas (Tanizawa et al., 1992)	Fetge (Tanizawa et al., 1992)
	Ratolí	1	2
		Pàncreas (Ishimura-Oka et al., 1995)	Fetge (Ishimura-Oka et al., 1995)
	Rata	1	3
		Pàncreas (Magnuson and Shelton, 1989)	Fetge (Magnuson and Shelton, 1989)

RGS9	Humà	1	3
		Cervell (Granneman et al., 1998; Zhang et al., 1999)	Fotoreceptors retinals (Zhang et al., 1999)
	Ratolí	2	1
		Cervell (Garzon et al., 2001; Rahman et al., 1999)	Fotoreceptors retinals (He et al., 1998)
TPM3	Humà	1	2
		Múscul esquelètic (Reinach and MacLeod, 1986)	Fibroblast (Reinach and MacLeod, 1986)
	Ratolí	1	2
		Múscul esquelètic (Pieples and Wieczorek, 2000)	Fibroblast (Takenaga et al., 1990)

Taula 6. Esdeveniments equivalents.

Pel que fa a les insercions/delecions, analitzàrem només aquelles que tenien una mida menor a un domini funcional –30 residus fou el llindar escollit. Això és perquè, en principi, la interpretació funcional de les grans delecions és més senzilla, ja que van associades a la pèrdua d'activitat a causa de l'eliminació de dominis funcionals, mentre que els canvis petits poden estar relacionats tant amb canvis lleugers com amb altres més dràstics, depenent dels ajustaments estructurals necessaris pel plegament de la isoforma (Garcia et al., 2004).

Les variables utilitzades foren la diferència de mides entre les regions delecionades equivalents –que es comparà amb la diferència de mida entre les isoformes curtes homòlogues- i la conservació d'accessibilitat i estructura secundària entre les isoformes equivalents. En el cas de l'accessibilitat, diferenciàrem dos estats pels residus –accessible i enterrat. Pel que fa a l'estructura secundària, agrupàrem els residus d'hèlix α i làmines β i els analitzàrem conjuntament i els diferenciàrem de les zones sense estructura definida (*coil*).

La conservació de l'accessibilitat i l'estructura secundària es calculà com el percentatge de residus alineats amb la mateixa propietat (veure Equació 6).

$$C_j = 100 \cdot \frac{\sum_{i=1}^{n} e_{i,j}}{n} \qquad \text{(Equació 6)}$$

on C és el percentatge de conservació de la propietat j; n és el nombre de residus humans alineats; e és qualsevol residu que té el mateix estat.

La Taula 7 ens mostra els resultats de la comparació entre isoformes equivalents d'esdeveniments d'insercions/delecions. Pel que fa a les mides, es veu com les regions delecionades són molt similars pel que fa a la seva mida (0.56 residus de diferència entre humans i ratolins, 0.12 entre humans i rates), molt més del que ho són les isoformes curtes entre elles (10.32 i 7.12 per les comparacions amb ratolins i rates, respectivament). L'accessibilitat i l'estructura secundària estan força conservades –en tots els casos-, sent els elements estructurats allò menys constant. No obstant això, més de tres quartes parts de residus tenen les propietats físico-químiques i estructurals conservades.

		Ratolí	Rata
Diferència de mida	Regió delecionada	0.56±1.98	0.12± 0.47
	Isoforma curta	10.32±42.53	7.12± 21.59
Accessibilitat	Residus enterrats	91%	87%
	Residus accessibles	94%	87%
Estructura secundària	Residus estructurats	77%	81%
	Residus desestructurats	94%	94%

Taula 7. Comparació dels esdeveniments equivalents que tenen insercions/delecions. La diferència

de mida es refereix a la que hi ha entre les regions delecionades equivalents i a la que hi ha entre les isoformes curtes homòlogues. Els percentatges es refereixen a la conservació per a cada estat de les propietats estudiades.

En el cas de les substitucions, comparàrem la identitat de seqüència entre els fragments substituïts en les isoformes equivalents –aquelles que tenen el mateix rol biològic- i les isoformes alternatives –les dues isoformes d'un esdeveniment d'*splicing* alternatiu. També calculàrem la conservació tant de l'accessibilitat a solvent com de l'estructura secundària, de la manera abans esmentada (veure Equació 6).

A la Taula 8 podem veure els resultats per les substitucions equivalents. Clarament es veu com les isoformes equivalents tenen un percentatge d'identitat de seqüència molt superior (per sobre de 0.80 en ambdues espècies) al que tenen els fragments alternatius (al voltant de 0.20). Això indica que pels organismes i gens considerats la variabilitat introduïda per l'*splicing* alternatiu és molt superior a la que aporta el procés d'especiació. Pel que fa a les característiques de les regions d'*splicing*, tant per l'accessibilitat com per l'estructura secundària dels residus, la conservació és superior al 80%.

		Ratolí	Rata
Identitat de seqüència	Isoformes equivalents	0.89±0.08	0.82±0.12
	Isoformes alternatives	0.24±0.15	0.23±0.08
Accessibilitat	Residus enterrats	94%	93%
	Residus accessibles	93%	85%
Estructura secundària	Residus estructurats	83%	80%
	Residus desestructurats	94%	84%

Taula 8. Comparació dels esdeveniments equivalents que tenen substitucions. La identitat de seqüència es refereix a l'alineament de les regions variables, implicades en l'*splicing* alternatiu, tant en les isoformes equivalents com en les alternatives. Els percentatges es refereixen a la conservació per a cada estat de les propietats estudiades.

5.4 Discussió

L'*splicing* alternatiu juga un paper molt important en la modulació de la funció proteica (Black, 2000; Graveley, 2001; Lopez, 1998). La seva acció es dóna mitjançant diversos mecanismes: deleció de dominis d'activació (Foulkes et al., 1992), creació de llocs d'unió addicionals (Zarich et al., 2000), substitució de fragments de seqüència per altres amb diferents propietats (Schmucker et al., 2000)...

Nosaltres hem estudiat el grau de conservació d'aquests mecanismes entre diverses espècies des de dos punts de vista –una comparació global i un estudi d'isoformes homòlogues.

Primer de tot, hem caracteritzat les zones variables per raó de l'*splicing* alternatiu, agrupant-les segons el mecanisme de modulació de la funció proteica. Posteriorment, les hem comparat –les d'humà amb les de ratolí, rata i mosca del vinagre.

En ambdós casos –insercions/delecions i substitucions- hi ha una gran similitud entre les diverses funcions de distribució (veure Taules 4 i 5). Aquesta semblança suggereix que hi ha conservació en l'efecte funcional de l'*splicing* alternatiu en les espècies considerades.

L'estructura gènica compartida podria explicar la gran similitud entre humans, ratolins i rates (Thanaraj et al., 2003), però l'estructura gènica dels invertebrats acostuma a ser molt diferent a la dels vertebrats (Deutsch and Long, 1999). Creiem que l'explicació més raonable dels nostres resultats és l'existència de restriccions físico-químiques (Benner et al., 1993) per a l'obtenció d'isoformes funcionals i, en conseqüència, organismes viables.

En segon lloc, hem analitzat el grau de conservació entre parelles d'esdeveniments d'*splicing* alternatiu homòlegs entre diferents espècies –humà comparat amb ratolí i rata-, és a dir, esdeveniments d'*splicing* alternatiu pels quals cada isoforma en una espècie té una isoforma equivalent en l'altra espècie.

Tant en les insercions/delecions com en el cas de les substitucions hi ha una gran

conservació de les propietats estructurals analitzades, quan comparem les isoformes equivalents. Per contra, quan ens fixem en les dues variants d'*splicing* que formen un esdeveniment, trobem que la conservació és molt menor (veure Taules 7 i 8). Això suggereix que, en el cas dels esdeveniments homòlegs, els efectes de l'*splicing* alternatiu sobre la funció en les diferents espècies –humà, ratolí i rata- deuen ser molt semblants, confirmant els resultats de l'estudi general.

En resum, els nostres resultats mostren que en espècies diferents hi ha una gran conservació pel que fa a la manera que l'*splicing* alternatiu modula la funció proteica. Probablement, les petites diferències que s'observen no tenen significat biològic i estan relacionades amb les diferències a nivell de l'estructura gènica que hi ha entre els diversos organismes (Russell and Barton, 1994).

Per tant, els nostres resultats i altres prèviament publicats semblen descartar que les diferències de complexitat entre els organismes siguin causades per diferències globals en l'ús dels mecanismes de l'*splicing* alternatiu per modificar la funció de les proteïnes, almenys en el cas dels esdeveniments majoritaris.

Si estem interessats en entendre les diferències entre les diferents espècies, haurem de parar atenció tant a l'*splicing* en famílies específiques de proteïnes (com ara els factors de transcripció (Latchman, 1996c)), com als esdeveniments minoritaris (Modrek and Lee, 2003; Wang, 2005) i a una altra gran varietat de mecanismes –modificacions post-traduccionals (Kondrashov and Koonin, 2001), evolució dels gens reguladors (Barrier et al., 2001), etcétera- però no a la magnitud del seu efecte sobre l'estructura de les proteïnes en els esdeveniments homòlegs.

6 *Splicing* alternatiu de factors de transcripció

6.1 Introducció

Als capítols anteriors hem vist com els efectes a nivell de proteïna produïts per l'*splicing* alternatiu són molt més dràstics que els que es produeixen després de la divergència dels duplicats (Talavera et al., 2007). Així mateix, també hem vist que els mecanismes de modulació proteica que utilitza l'*splicing* alternatiu estan molt conservats evolutivament (Valenzuela et al., 2004). Finalment, decidírem estudiar com es tradueixen aquests canvis en una família concreta de gran interès biològic: els factors de transcripció (Lopez, 1995; Lopez, 1998).

L'expressió dels gens es regula a diversos nivells –seqüència, cromatina i nucli (Arney and Fisher, 2004; Hurst et al., 2004; van Driel et al., 2003). Fins i tot, es pensa que l'ordre dels gens dins del genoma té a veure amb el seu control d'expressió (Hurst et al., 2004). Els factors de transcripció tenen un rol molt important a l'hora de controlar l'expressió gènica (Lopez, 1995), ja que tenen la facultat d'augmentar o reprimir la transcripció dels gens mitjançant llur unió a seqüències específiques del DNA; per tant, són responsables dels tipus i quantitats de proteïnes que existeixen a la cèl·lula. Per ser actius, els factors de transcripció han de ser capaços de llegir diferents senyals: des de seqüències específiques de nucleòtids fins al codi de modificació de les histones.

Els factors de transcripció acostumen a estar construïts d'una manera modular (Laoide et al., 1993; Lopez, 1995), és a dir, amb un o més dominis, que poden tenir diferents funcions: dominis d'unió al DNA, dominis d'interaccció i dominis reguladors (veure Figura 32). Això fa que, tal com es veu a la Figura 33, els factors de transcripció siguin sovint proteïnes de mida gran.

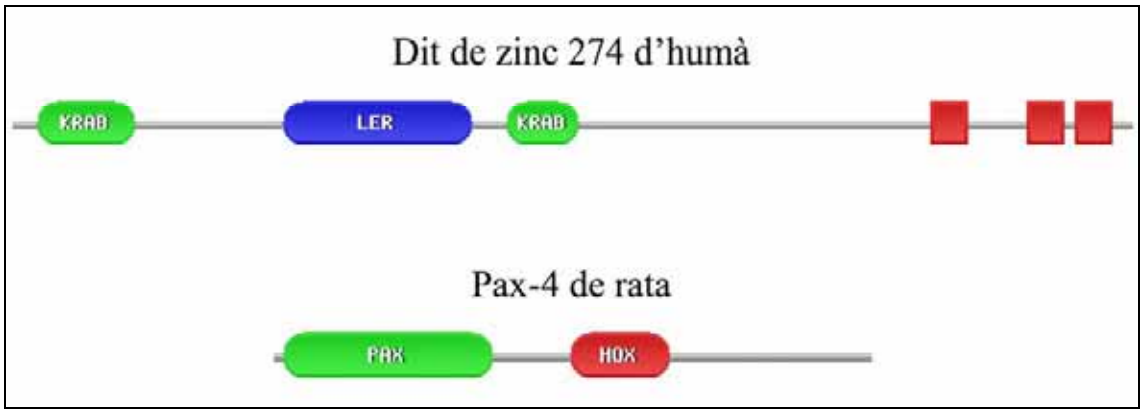

Figura 32. Composició modular dels factors de transcripció. Es mostren els dits de zinc 274 d'humà i Pax-4 de rata. Les diverses capses són dominis funcionals. En verd, els dominis reguladors; en blau, els d'oligomerització; en vermell els d'unió al DNA.

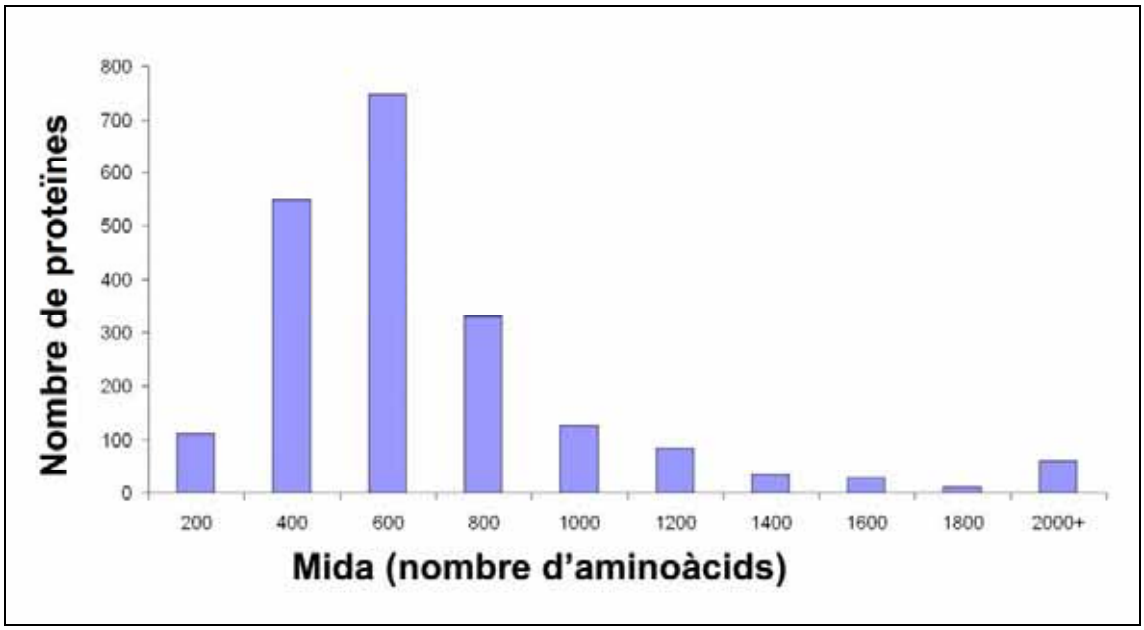

Figura 33. Distribució de mides dels factors de transcripció. Les mides són en nombre de residus. Els factors de transcripció són proteïnes multi-dominis que necessiten tenir una mida relativament gran

Els gens que codifiquen per factors de transcripció també poden tenir *splicing* alternatiu, generant diferents isoformes reguladores (Hsu et al., 1992; Laoide et al., 1993; Latchman, 1996c). L'*splicing* alternatiu pot influir en els factors de transcripció per mitjà de tres vies diferents: canviant les seves propietats d'unió al DNA, alterant les seves característiques d'activació/repressió i afectant els dominis de dimerització (Lopez, 1995). Aquests canvis poden afectar la funció dels factors de transcripció de diverses maneres: manca d'activitat, augment o disminució de l'activitat o efecte agonista (Latchman, 1996c). Addicionalment, tal com s'ha comentat anteriorment,

l'*splicing* pot ser específic del teixit o de l'estadi de desenvolupament (Lopez, 1995; Taneri et al., 2004). Per exemple, el gen *Nr1i3* de ratolí, que codifica pel receptor nuclear orfe NR1I3, dóna una isoforma curta (CAR2) que no té part del domini de dimerització/unió a lligand (veure Figura 34) i, per tant, no és capaç d'activar la transcripció (Choi et al., 1997). En el cas del factor de transcripció humà p65, codificat pel gen *RELA*, l'*splicing* alternatiu remou una làmina β inclosa en el domini RHD (veure Figura 34). La modificació del domini RHD en la isoforma delta desafavoreix la formació de l'heterodímer amb p50 –l'heterodímer activa la transcripció-, pensant-se que actua com un regulador negatiu (Ruben et al., 1992).

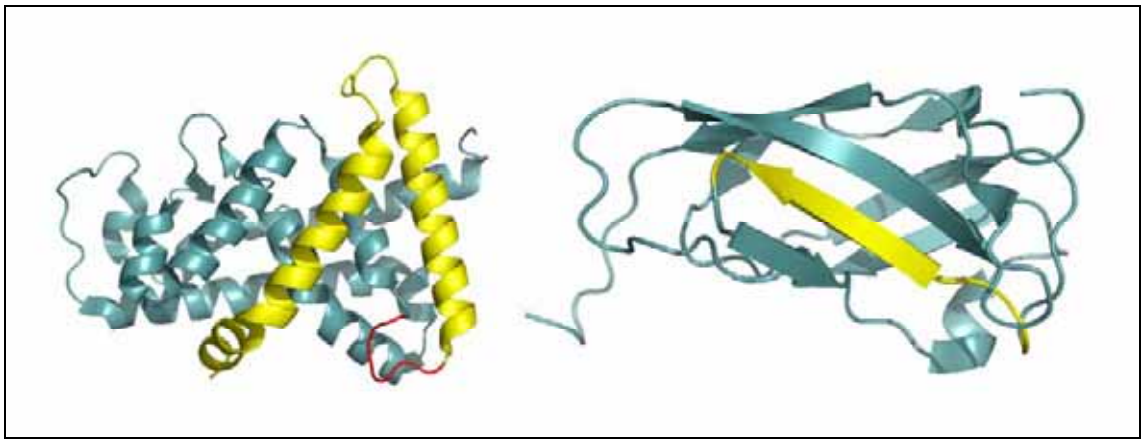

Figura 34. Efectes de l'*splicing* alternatiu sobre els factors de transcripció. En groc, les regions delecionades en les variants d'*splicing*; en vermell, els fragments substituïts. A l'esquerra, NR1I3 de ratolí (codi PDB: 1xnx (Shan et al., 2004)). A la dreta, p65 de ratolí (codi PDB: 1my7 (Huxford et al., 2002)).

En el nostre treball, hem intentat clarificar una mica més la importància de les variants d'*splicing* en la diversitat funcional dels factors de transcripció, fixant-nos en els tipus de dominis modificats i en la manera que té l'*splicing* alternatiu de modificar la composició global de dominis funcionals. Així mateix, en una segona part hem analitzat la conservació interespecífica de la variabilitat.

6.2 Diversitat dels factors de transcripció

Els factors de transcripció han de ser molt versàtils (Latchman, 1996a), car han de respondre a múltiples i diversos estímuls –tan interns com externs- unint-se a seqüències específiques del DNA. La seva manera més freqüent d'actuació consisteix a activar la transcripció dels gens, però alguns actuen com a inhibidors (Latchman,

1996b).

El nostre estudi començà per una anàlisi global sobre la variabilitat dels factors de transcripció. Treballàrem amb els factors de transcripció de la base de dades SwissProt (Boeckmann et al., 2003) que havíem extret pel sistema SRS (Etzold et al., 1996). Ens fixàrem en el seu nombre, la seva taxa d'*splicing* alternatiu i la semblança entre els diversos factors de transcripció

La Taula 9 fa una descripció de la quantitat i variabilitat dels factors de transcripció d'organismes eucariotes en general i d'humans i ratolins en particular. Els factors de transcripció són força nombrosos: un 5% dels gens presents a la base de dades corresponen a factors de transcripció -7% i 8% en ratolí i humà, respectivament. Malgrat tot, la gran proporció de factors de transcripció pot ser el resultat de biaixos en la composició de la base de dades pel gran interès que hi ha hagut en l'estudi d'aquestes proteïnes. Sembla que els gens de factors de transcripció tenen *splicing* alternatiu de manera més freqüent que la majoria de gens, així a humans es passa del 26.9% en tots els gens al 29.4% dels factors de transcripció, i quan mirem tots els organismes eucariotes es dobla el percentatge (de 8.2% a 17%) (test t-Student: p-valor = 0.08, p-valor = 0, p-valor = 0, per humans, ratolins i eucariotes, respectivament) (Taneri et al., 2004); per contra, la quantitat de variants que generen no és significativament diferent de la resta de proteïnes (test t-Student: p-valor > 0.1).

	Gens de factors de transcripció			**Tots els gens**		
	Eucariotes	**Ratolí**	**Humà**	**Eucariotes**	**Ratolí**	**Humà**
Nombre de gens	4377	737	1077	85952	10031	12946
% gens amb *splicing* alternatiu	17.0	26.1	29.4	8.2	17.6	26.9
Isoformes/gen	2.8	2.9	3.0	2.8	2.7	2.8

Taula 9. Breu descripció de les dades: nombre de gens totals i de factors de transcripció, percentatge de gens amb *splicing* alternatiu i nombre d'isoformes per gen.

Com hem vist anteriorment, una de les maneres d'incrementar la variabilitat és la duplicació gènica, seguida de la divergència evolutiva. Al capítol 4 hem vist com en

humans no s'observaven preferències a nivell funcional a l'hora de diversificar-se mitjançant l'*splicing* alternatiu o la duplicació; això no obstant, decidírem analitzar els factors de transcripció de diversos organismes per veure quin paper podia jugar-hi la duplicació gènica.

Per realitzar aquesta anàlisi vam mirar el percentatge d'homòlegs, és a dir, miràrem quina proporció de proteïnes compartien un determinat percentatge d'identitat (veure Taula 10). Això ens permetia tenir una idea aproximada de la importància de la duplicació gènica o, almenys, saber quina diversitat hi ha entre les proteïnes. Les restriccions estructurals (Chothia and Lesk, 1986) i un nombre limitat de dominis funcionals ens feien suposar que hi podria haver una alta taxa d'homologia dins de les diverses espècies. Com a control, utilitzàrem les proteïnes humanes que no eren factors de transcripció.

La primera cosa que crida l'atenció dels resultats mostrats a la Taula 10 és la baixa homologia dins de les diverses espècies, almenys per les identitats molt altes. Això ens indica que o bé hi ha molt poca duplicació gènica en aquest tipus de proteïnes, o bé hi ha una gran divergència entre les famílies i dins d'elles. Si els factors de transcripció es comporten igual que els enzims, això fa que difícilment puguin tenir la mateixa funcionalitat (Tian and Skolnick, 2003). Totes les espècies tenen almenys dos terços de proteïnes que no són homòlegs, és a dir, tenen menys del 40% de residus idèntics – essent el llevat el cas més dramàtic, amb un 3% de proteïnes similars en aquest llindar. En promig, menys d'un 1% de les seqüències dels factors de transcripció tenen un 90% d'identitat i hem de baixar el llindar fins al 50% per trobar un mínim d'un 10% de proteïnes homòlogues.

D'altra banda, el resultat més interessant des del punt de vista d'aquest treball és l'aparent manca d'influència de la capacitat de fer *splicing* alternatiu a l'hora de tenir proteïnes més o menys homòlogues. D'aquesta manera, mentre que *E. coli*, que no fa *splicing* alternatiu, té unes taxes d'homologia molt semblants a les de ratolí i rata, el llevat, que té un petit nombre de variants d'*splicing*, gairebé no té factors de transcripció similars

.

Identitat entre les proteïnes	Control	Humà	Ratolí	Rata	Mosca del vinagre	Llevat	*E. coli*
90%	0.04	0.01	0.01	0.01	0.00	0.00	0.00
80%	0.07	0.02	0.02	0.01	0.02	0.00	0.02
70%	0.10	0.07	0.04	0.04	0.03	0.01	0.06
60%	0.16	0.13	0.08	0.07	0.04	0.01	0.06
50%	0.22	0.22	0.15	0.14	0.08	0.01	0.10
40%	0.31	0.34	0.23	0.20	0.11	0.03	0.17
Nombre de gens	**10937**	**1077**	**737**	**255**	**199**	**183**	**239**

Taula 10. Divergència intraespecífica dels factors de transcripció. La taula mostra, per a cada llindar d'identitat, el percentatge d'homòlegs, és a dir, la proporció de proteïnes que són similars.

6.3 Selectivitat

Com que el nostre interès se centrava en la modulació de la funció per raó de l'*splicing* alternatiu, ens fixàrem només en el subconjunt de factors de transcripció que tenien variants d'*splicing* –2087 isoformes, a partir de 742 gens.

166 proteïnes (22.4%) tenen almenys una isoforma amb canvis a la composició de dominis, adquirint o perdent elements funcionals. La Taula 11 resumeix els principals canvis causats per l'*splicing* alternatiu. La taula mostra els deu dominis funcionals que, en nombres absoluts, són modificats més freqüentment. Així mateix, mostra el nombre de proteïnes que tenen aquests dominis i en quina proporció de casos almenys una isoforma té canvis de composició.

Domini	**Nombre de proteïnes amb aquest domini**	**Nombre (fracció) de proteïnes amb isoformes que perden o guanyen aquest domini**	**Funció**
KRAB	25	15 (60%)	Regulació
ZF-C2H2	24	13 (54%)	Unió al DNA
HOLI	92	12 (13%)	Unió a lligand
HOX	80	9 (11%)	Unió al DNA
HLH	46	8 (17%)	Unió al DNA
PHD	23	6 (26%)	No coneguda
ZNF_C4	92	6 (6%)	Unió al DNA
SAM	5	5 (100%)	Dimerització
PAS	16	5 (31%)	Unió a lligand
ZNF_GATA	7	4 (57%)	Unió al DNA

Taula 11. Dominis funcionals més afectats per l'*splicing* alternatiu.

El domini més afectat, en quantitat, és KRAB (*Krueppel-associated box*), que està normalment associat amb un tipus de dits de zinc: C2H2. És un domini regulador que reprimeix la transcripció (Margolin et al., 1994); per tant, les isoformes que no tenen el domini KRAB permeten l'expressió dels gens reprimits (Witzgall et al., 1994). Altres estudis han trobat que aquest domini està molt afectat per l'*splicing* alternatiu (Resch et al., 2004). Per la seva banda el domini de dits de zinc ZF-C2H2 també està molt afectat pels canvis d'*splicing*.

Un altre domini que està significantment afectat és SAM (*Sterile Alpha Motif*), implicat en l'homodimerització dels factors de transcripció. És un domini que està present en diverses proteïnes nuclears i de senyalització (Schultz et al., 1997), però nosaltres sempre el trobem associat al domini P53.

Tot i afegir o eliminar més freqüentment els dominis d'unió al DNA (Taneri et al., 2004), sembla que l'*splicing* alternatiu utilitza totes les variants per alterar els factors de transcripció.

Una altra cosa remarcable és que l'*splicing* alternatiu sembla afectar preferentment uns dominis respecte a uns altres. Així, per exemple, mentre ZNF_C4 és el dit de zinc més abundant en les nostres dades, altres com ZF-C2H2 i ZNF_GATA estan afectats relativament més sovint per l'*splicing* alternatiu. Aquest punt és molt important a l'hora de la regulació de l'expressió gènica, perquè quan un factor de transcripció perd el seu domini d'unió al DNA –per exemple, un dit de zinc- no pot activar la transcripció, però pot ser un regulador negatiu, ja que, si manté el seu domini de dimerització, pot segrestar isoformes actives i evitar-ne la unió al DNA (Lopez, 1995).

*6.4 Especificitat dels efectes de l'*splicing *alternatiu*

Alguns treballs anteriors deien que l'*splicing* alternatiu tendeix a modificar dominis sencers en lloc de fragments (Kriventseva et al., 2003). Nosaltres ens interessàrem en el grau de solapament entre els dominis funcionals i les regions d'*splicing* alternatiu; per això, mesuràrem la relació entre els límits de la zona variable i les fronteres dels dominis funcionals.

Primer de tot mesuràrem l'especificitat de l'efecte, és a dir, la quantitat de residus que formen part alhora de la regió d'*splicing* i el domini funcional, enfront de tots els residus que formen part d'una cosa o l'altra. Ho comparàrem amb un model d'atzar, consistent en anar movent una finestra de la mateixa mida de la zona d'*splicing* per tota la seqüència de la proteïna i mesurar l'especificitat en cada posició.

La Figura 35 mostra una especificitat baixa, gairebé idèntica a la que donaria la distribució a l'atzar de l'*splicing* alternatiu. Per tant, no hi ha gairebé mai coincidència entre els límits dels dominis i els dels fragments afectats per l'*splicing* alternatiu.

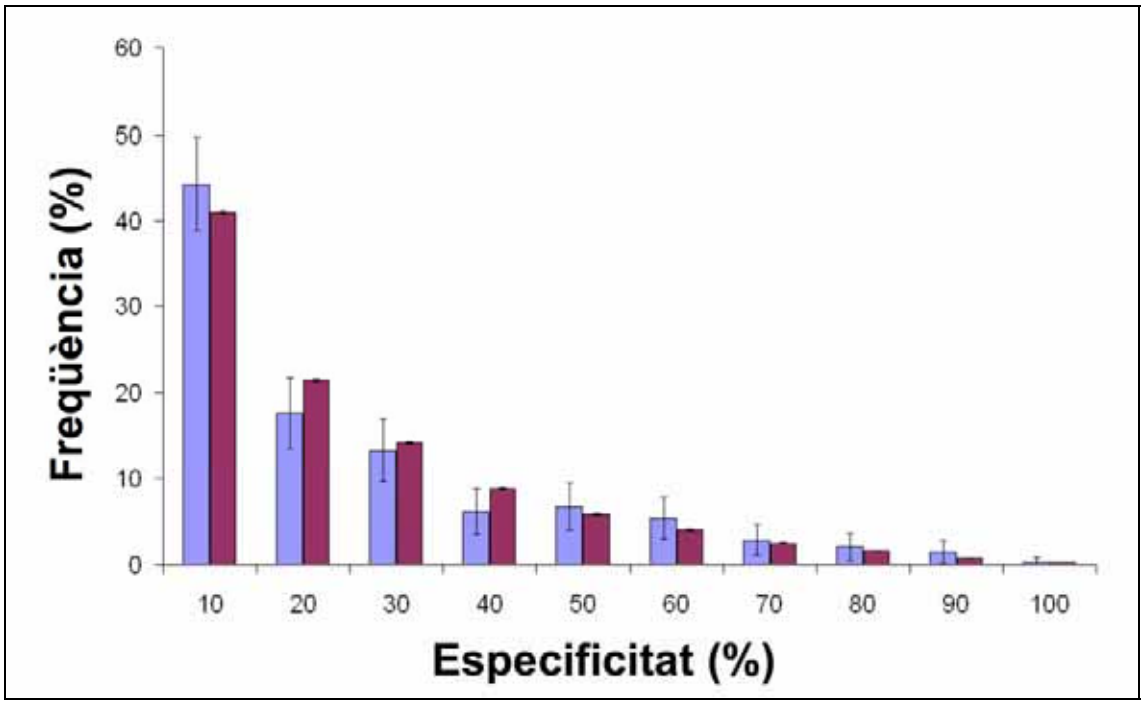

Figura 35. Especificitat dels efectes de l'*splicing* alternatiu sobre els dominis funcionals. En blau, la distribució observada. En fúcsia, l'esperada.

Posteriorment, classificàrem els efectes de l'*splicing* alternatiu en quatre tipus diferents (veure Figura 36): coincidència entre les fronteres funcionals i d'*splicing* (tipus 1); domini funcional inclòs totalment dins de la regió d'*splicing* alternatiu (tipus 2); regió d'*splicing* alternatiu completa a l'interior del domini funcional (tipus 3); *splicing* alternatiu afectant el domini funcional de manera parcial i incloent, a més, residus propers (tipus 4).

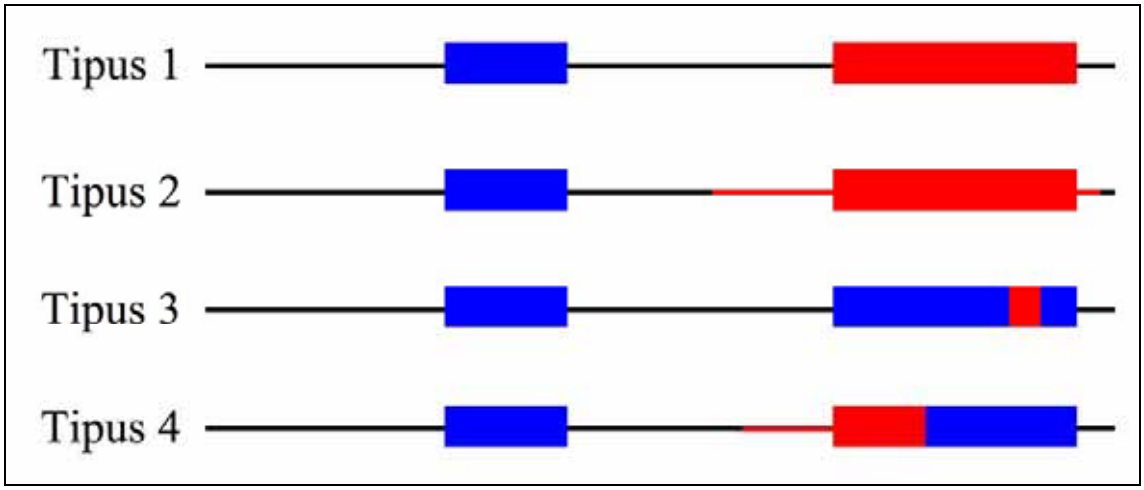

Figura 36. Tipus d'efecte de l'*splicing* alternatiu sobre els dominis funcionals. Les capses representen els dominis. En vermell, la regió d'*splicing* alternatiu.

La Taula 12 mostra que no hi ha casos de tipus 1 (tal com vèiem a la Figura 35). Els efectes més freqüents són els de tipus 4 –modificació parcial del domini i la seqüència propera. En canvi, el tipus 2, tot i estar per sobre de l'atzar, no és tan comú com era d'esperar d'acord amb la bibliografia existent (Kriventseva et al., 2003). Per la seva banda, els canvis d'*splicing* a l'interior dels dominis estan disminuïts respecte a l'atzar.

	Observat	Esperat
Tipus 1	0.00	0.00
Tipus 2	0.29	0.17
Tipus 3	0.32	0.44
Tipus 4	0.39	0.39

Taula 12. Efectes de l'*splicing* alternatiu sobre els dominis funcionals. Per a cada tipus d'efecte (tal com els hem definit abans) es dóna la freqüència observada i l'esperada seguint el model d'atzar.

6.5 Conservació interespecífica

Finalment, ens interessàrem per la conservació dels factors de transcripció –des d'uns punt de vista de variabilitat i de funcionalitat. Per analitzar la divergència després de l'especiació, estudiàrem 148 ortòlegs amb *splicing* alternatiu, presents a humà i ratolí. D'aquests 148 ortòlegs, 79 (53.4%) tenen diferent nombre d'isoformes.

6.5.1 Conservació estructural

La Taula 13 mostra una breu descriptiva dels factors de transcripció analitzats, centrant-nos en llur nombre d'isoformes, el nombre de dominis funcionals que té cadascuna i la quantitat que se'n perd o guanya a causa de l'*splicing* alternatiu.

	Humà	Ratolí
Nombre d'isoformes/gen	3.20 ± 1.85	2.95 ± 1.74
Nombre de dominis/isoforma	1.57 ± 1.40	1.58 ± 1.38
Dominis guanyats o perduts per raó de l'*splicing* alternatiu	0.16 ± 0.36	0.18 ± 0.45

Taula 13. Descripció dels factors de transcripció ortòlegs presents en humà i ratolí.

Els resultats mostren que no hi ha diferències significatives per les variables estudiades (test t-Student, p-valor > 0.05 en tots els casos), és a dir, tant la composició de dominis com la manera d'augmentar la variabilitat són semblants, de manera global, entre les espècies.

Tanmateix, vam intentar aprofundir més el nostre estudi, analitzant la conservació dels parells. Per això, miràrem els percentatges d'identitat dels ortòlegs i dels dominis homòlegs –constitutius i alternatius.

La Figura 37 mostra els resultats per la distribució d'identitats per a tota la proteïna i per a tots els dominis funcionals equivalents. S'observa clarament una alta conservació tant pel que fa als ortòlegs com als dominis concrets. Malgrat tot, els dominis funcionals estan més conservats que la seqüència sencera (test KS, p-valor < 0.05) –la majoria són quasi idèntics-, resultat que és consistent amb el fet que la major part de la divergència interespecífica se situï en les zones no definides com a funcionals a les bases de dades.

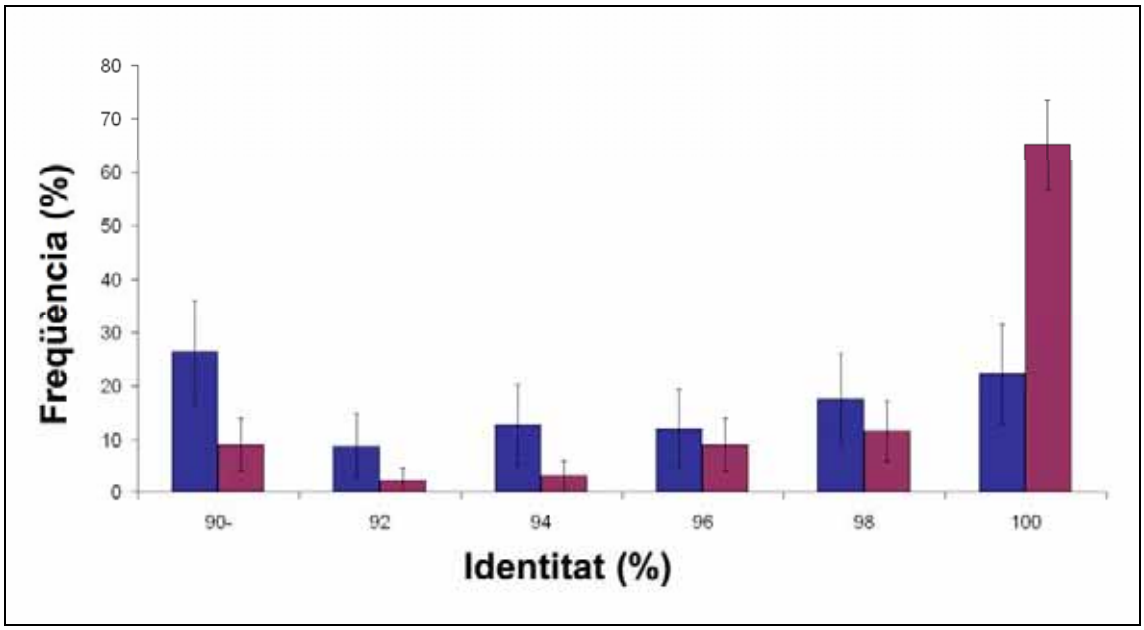

Figura 37. Identitat dels dominis funcionals. Distribucions d'identitat per tota la proteïna (blau) i els dominis funcionals equivalents (morat). Les identitats menors o iguals que 90% s'acumules en una sola barra de l'histograma.

Posteriorment, miràrem si hi havia diferències entre els dominis equivalents constitutius –presents a totes les isoformes- i alternatius –només en algunes. La Figura 38 mostra els resultats per les dues distribucions d'identitat, que mostren una gran conservació pels dos tipus de dominis equivalents. Aquest nivell d'identitat ens fa pensar que no hi haurà massa diferències interespecífiques pel que fa a la funcionalitat dels dominis. Tot i que estadísticament les dues distribucions siguin diferents (test KS, $p < 0.05$), probablement a causa de la baixa conservació d'uns pocs dominis constitutius.

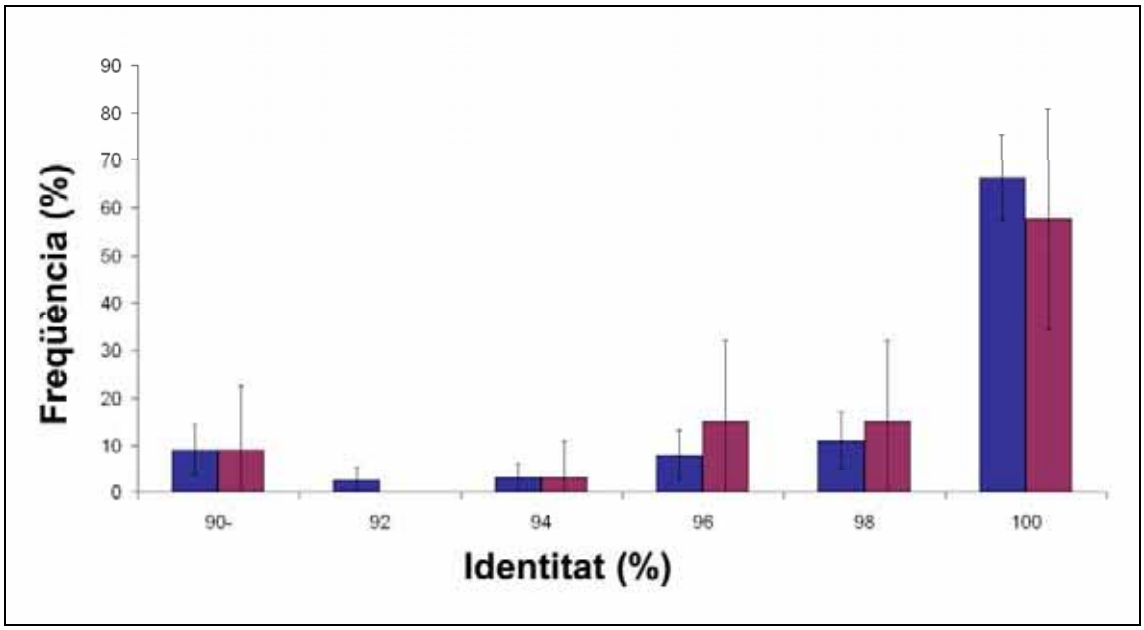

Figura 38. Identitat dels dominis constitutius i alternatius. Distribucions d'identitat per dominis equivalents constitutius (blau) i els dominis equivalents alternatius (morat). Les identitats menors o iguals que 90% s'acumules en una sola barra de l'histograma.

6.5.2 Conservació funcional

Per fer-nos una idea de la conservació funcional dels factors de transcripció i analitzar si la presència de variants d'*splicing* determinava d'alguna manera el seu patró d'expressió, analitzàrem els patrons d'expressió dels factors de transcripció ortòlegs. Vam extreure les dades d'expressió del servidor SymAtlas (http://symatlas.gnf.org) (Su et al., 2004) i mesuràrem la correlació de Pearson i la distància euclidea entre els patrons d'expressió dels gens humans i de ratolí, seguint la metodologia proposada per Liao i Zhang (Liao and Zhang, 2006). Les dues mesures han estat àmpliament utilitzades per comparar patrons d'expressió entre gens (Huminiecki and Wolfe, 2004; Jordan et al., 2005; Makova and Li, 2003; Yanai et al., 2004).

Dividírem els factors de transcripció ortòlegs en tres grups: presència d'*splicing* alternatiu en les dues espècies, variants d'*splicing* en una de les espècies i gens sense *splicing* alternatiu en les dues espècies. Comparàrem els resultats dels factors de transcripció amb un altre grup de proteïnes –els enzims- que fou utilitzat com a control.

La Figura 39 mostra les distribucions pels quatre grups esmentats –tres de factors de transcripció i un d'enzims- per les dos mesures utilitzades. En tots els casos la distribució és similar: una lleugera correlació positiva i poca distància dels patrons

d'expressió. Això significa que la presència o absència de variants d'*splicing* no altera de manera significativa els patrons d'expressió tissular dels factors de transcripció.

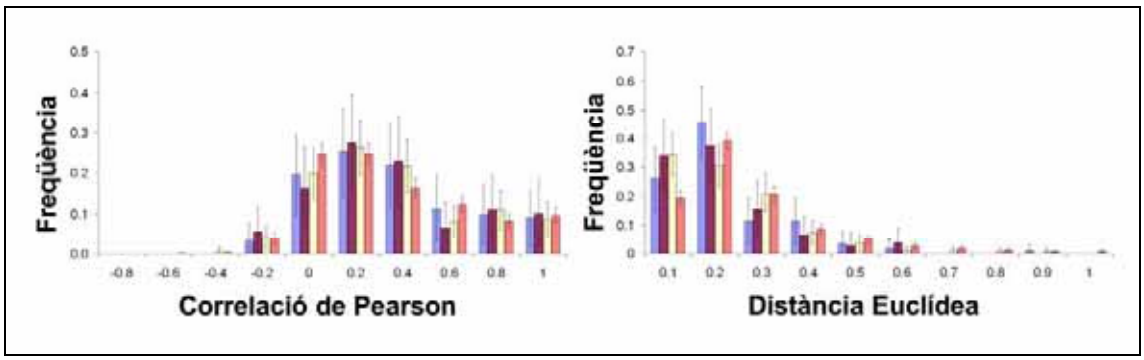

Figura 39. Expressió dels factors de transcripció. Mesures dels patrons d'expressió per factors de transcripció ortòlegs amb més d'una isoforma en les dues espècies (blau), factors de transcripció amb variants d'*splicing* en una de les espècies (grana), factors de transcripció sense *splicing* (groc) i enzims (taronja).

Finalment, analitzàrem més detingudament els patrons d'expressió dels ortòlegs amb *splicing* alternatiu en ambdues espècies. La idea d'aquesta anàlisi era veure si diferències en la composició de dominis duien associats canvis en els patrons d'expressió, ja que se sap que algunes isoformes tenen especificitat tissular (Lopez, 1995; Taneri et al., 2004) Comparàrem 13 gens que tenien diferències en el nombre de dominis guanyats o perduts en les dues espècies –senyal d'aparició d'isoformes no equivalents funcionalment i, per tant, més susceptibles de fer variar el patró d'expressió- amb 109 casos que no tenien diferències pel que fa al nombre de dominis funcionals afectats.

La Taula 14 mostra el promig i la mediana de cadascuna de les mesures utilitzades. Es veu com no hi ha diferències significatives per raó de l'aparició d'isoformes amb diferent funcionalitat (test t-Student, p-valor > 0.05 en tots els casos).

	Correlació de Pearson		Distància euclidea	
	Promig	Mediana	Promig	Mediana
Ortòlegs canviants	0.30±0.31	0.17	0.18±0.10	0.18
Ortòlegs constants	0.27±0.31	0.21	0.19±0.13	0.16

Taula 14. Comparació dels patrons d'expressió dels factors de transcripció amb diferències funcionals interespecífiques i els que no en tenen.

6.6 Discussió

Els factors de transcripció juguen un paper molt important en el control de les funcions biològiques i el cicle cel·lular (Lopez, 1995); per això, són molt nombrosos i divergents, fet que els permet una gran versatilitat en la resposta a estímuls (Latchman, 1996a; Latchman, 1996b).

Els factors de transcripció són proteïnes d'una mida considerable i tenen una composició modular: estan formats per diversos dominis funcionals –d'unió al DNA, d'interacció i de regulació. Per tant, els canvis en la composició de dominis per raó de l'*splicing* alternatiu poden donar isoformes amb funcions diferents (Lopez, 1995).

Els nostres resultats mostren que l'*splicing* alternatiu no modifica els dominis funcionals a l'atzar, sinó que n'hi ha uns que semblen ser-hi més propensos que altres. Això no obstant, contràriament a observacions prèvies (Taneri et al., 2004), no trobem una preferència clara per cap de les grans funcions (unió al DNA, dimerització...), sinó que sembla que la preferència és una característica dels dominis concrets.

Contràriament als resultats d'estudis previs que concloïen que l'*splicing* alternatiu eliminava dominis sencers en lloc de trencar-los (Kriventseva et al., 2003), nosaltres hem vist com l'efecte de l'*splicing* alternatiu sobre els dominis funcionals és similar al de l'atzar, si ens fixem en la seva precisió. No obstant això, si dirigim la nostra atenció vers la manera de modificar els dominis, veiem com l'eliminació de dominis sencers està enriquida respecte al que esperaríem si els efectes fossin a l'atzar, tot i ser clarament minoritària.

Finalment, la conservació entre els factors de transcripció humans i de ratolí és força alta, tant pel que fa a la seqüència com dels nivells d'expressió. Pel que fa a l'expressió dels factors de transcripció ortòlegs no es veuen diferències significatives ni atenent a la seva possibilitat de generar diverses isoformes, ni a la seva capacitat de generar noves isoformes –específiques de l'espècie.

7 Un mètode per cercar esdeveniments equivalents

7.1 Introducció

Tal com s'ha comentat anteriorment, el coneixement de l'*splicing* alternatiu dels gens i la seva funció té una gran importància en biomedicina, ja que aquest mecanisme s'ha relacionat amb múltiples malalties (Caceres and Kornblihtt, 2002; Venables, 2006; Xing and Lee, 2006), molts cops per culpa de l'expressió d'isoformes en un teixit o moment que no tocava (Venables, 2004). Per tant, s'ha estudiat la possibilitat d'utilitzar l'*splicing* alternatiu com una diana terapèutica (Hagiwara, 2005), intentant corregir els seus errors (Cartegni and Krainer, 2003; Garcia-Blanco et al., 2004). En aquesta aproximació és imprescindible tenir un coneixement extens de les isoformes (Cuperlovic-Culf et al., 2006), car actuar sobre la isoforma equivocada pot tenir conseqüències contraproduents (Akgul et al., 2004).

D'altra banda, la biomedicina utilitza constantment models animals per intentar reproduir en altres animals els desordres que causen malalties en humans (Gibson et al., 2005; Newman et al., 2006) i també per poder fer estudis in vivo de l'*splicing* alternatiu (Chauhan et al., 2004; Mereau et al., 2007). Però per tenir uns bons models no és suficient que l'investigador utilitzi un organisme amb *splicing* alternatiu, sinó que aquest ha de ser equivalent en el sistema d'interès i el model (veure Figura 40).

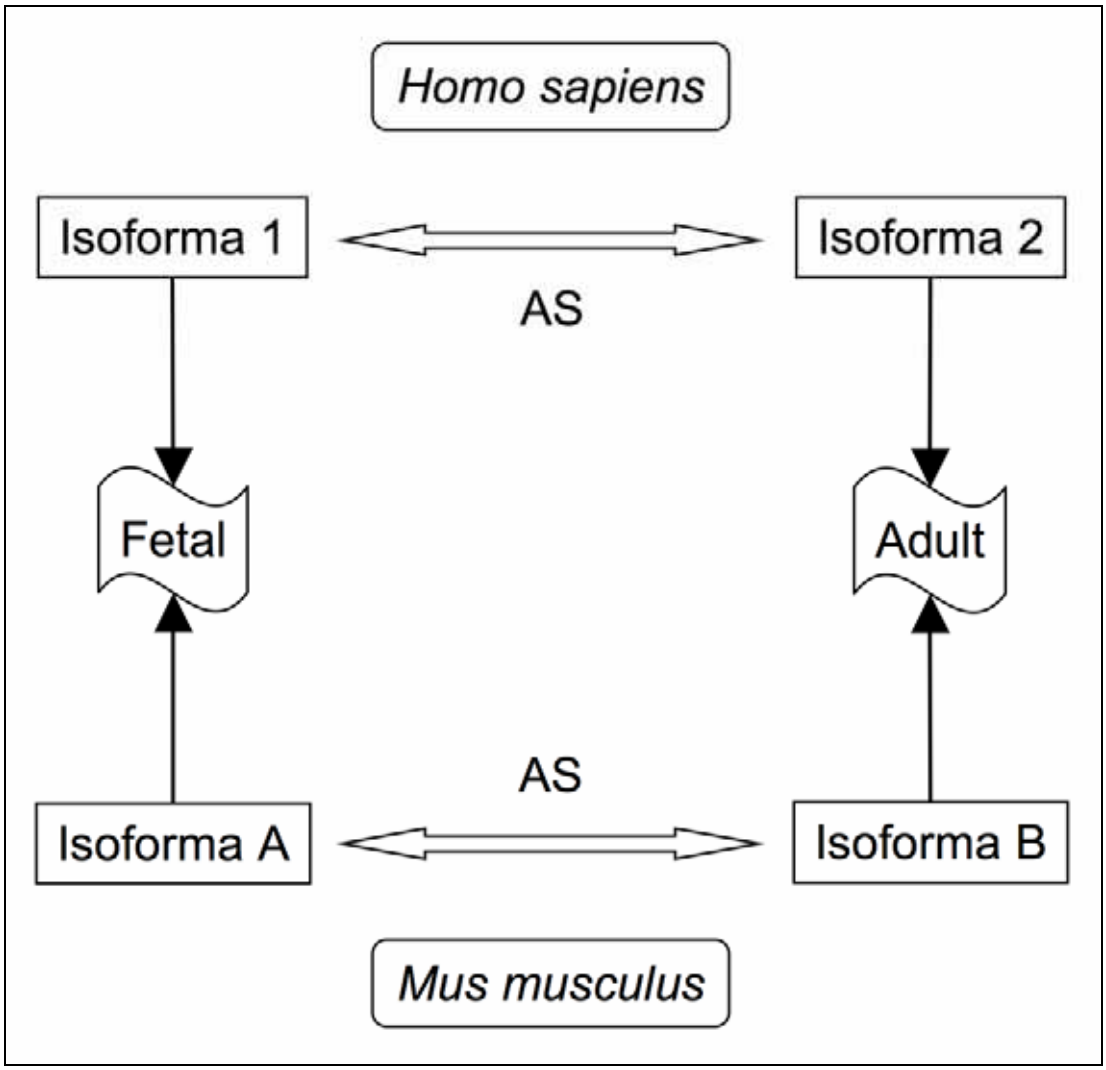

Figura 40. Equivalència entre esdeveniments equivalents. En aquest cas, la isoforma 1 humana i la isoforma A de ratolí són equivalents perquè s'expressen ambdues al fetus; mentrestant, la isoforma 2 humana i la isoforma B murina s'expressen en l'adult. L'equivalència particular de les isoformes comporta l'equivalència de l'esdeveniment d'*splicing* alternatiu definit com els canvis entre les isoformes 1 i 2 en l'humà i les isoformes A i B en el ratolí.

Tot això posa de manifest la importància de disposar d'anotacions funcionals de l'*splicing* alternatiu, que ens han de permetre pujar un nivell en la comprensió de la jerarquia biològica, a diferència de les anotacions funcionals tradicionals que es fan a nivell de proteïnes.

Per facilitar la feina dels investigadors experimentals i ajudar en l'anotació funcional de les isoformes, nosaltres proposem un mètode de predicció que permet detectar esdeveniments equivalents d'*splicing* alternatiu. El nostre mètode, que anomenem SPLASH, s'aplica a esdeveniments formats per dues isoformes i, per tant, no cobreix tot el patró d'*splicing* alternatiu d'un gen, ja que encara que per moltes proteïnes només

s'hagin definit dues isoformes, n'hi ha un percentatge no menyspreable que en tenen unes quantes més i, si mirem a les isoformes predites, aquestes darreres són majoria (Neverov et al., 2005) (veure Taula 15). Addicionalment, hi ha també casos extrems, com el gen *Dscam* de *D. melanogaster* que és el cas paradigmàtic de diversitat d'isoformes, ja que, en teoria, en pot tenir més de 38.000 (Schmucker et al., 2000). En aquests casos, es requeriria l'aplicació reiterada del nostre algoritme, encara que aquest problema no s'ha tractat a la present tesi.

Nombre d'isoformes per proteïna	Isoformes a SwissProt (Boeckmann et al., 2003)	Isoformes predites (Neverov et al., 2005)
2	65.4%	~32%
3	17.5%	~10%
4	8.8%	~17%
5 i més	8.2%	~41%

Taula 15. Distribució d'isoformes per proteïna.

7.2 Mètode de predicció

L'objectiu d'aquest mètode de predicció és respondre una qüestió força senzilla: coneixent un esdeveniment d'*splicing* alternatiu que genera dues isoformes d'una proteïna és possible trobar altres esdeveniments d'*splicing* alternatiu, en la mateixa espècie o en altres, que es puguin considerar equivalents a aquest?

Tanmateix, la manera de resoldre el problema no n'és tant de senzilla i s'ha plasmat en un protocol que requereix l'ús combinat de diverses tècniques bioinformàtiques en quatre passos principals (veure Figura 41): 1) Cerca de candidats amb BLAST (Altschul et al., 1990); 2) Construcció de tots els possibles esdeveniments candidats; 3) Puntuació dels esdeveniments candidats mitjançant l'ús de xarxes neuronals; 4) Filtració dels resultats. En els apartats següents es descriuen els passos realitzats per a la posada a punt d'aquest protocol: cerca de candidats, construcció i puntuació dels possibles esdeveniments i obtenció de la predicció.

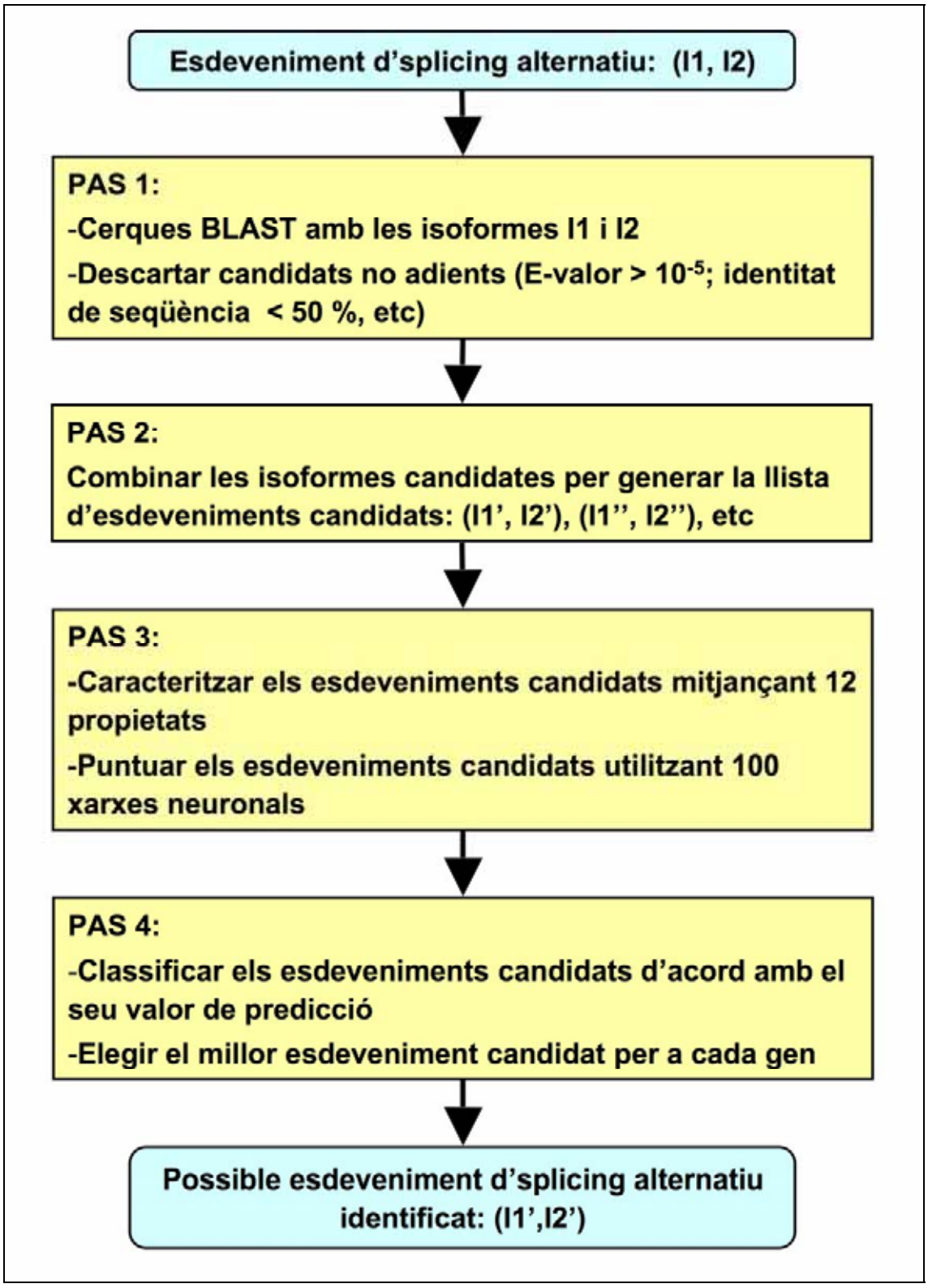

Figura 41. Esquema del mètode de predicció. A partir de les dues isoformes d'un esdeveniment problema es fan sengles cerques amb BLAST; es construeixen tots els esdeveniments equivalents possibles i es calculen els índexs de la relació; els vectors es passen a la xarxa neuronal i amb els filtres posteriors es decideix quines relacions són equivalents.

Primer, es parteix d'un esdeveniment d'*splicing* alternatiu del qual se sap, almenys, la seqüència d'una de les isoformes i els canvis per reconstruir l'altra. Posteriorment, per a cadascuna d'aquestes isoformes es realitza una cerca amb el programa BLAST (Altschul et al., 1990) en una base de dades que conté les diverses isoformes de cada proteïna. Per a cada proteïna, només la isoforma amb la millor puntuació en la valoració de l'alineament es manté com a candidata. En cas d'haver-hi diverses isoformes amb la mateixa puntuació, no es descarten cap d'aquestes. A més, cada isoforma és alineada amb els seus candidats i es descarten tots aquells que tenen una identitat global i local –de la zona d'*splicing* alternatiu- menor que el 50%.

A partir de les isoformes candidates obtingudes al pas anterior, es construeixen tots els esdeveniments equivalents possibles a partir de les isoformes candidates, tenint en compte que les isoformes han de ser diferents però del mateix gen (veure Figura 41). Tots aquests esdeveniments candidats són caracteritzats amb l'ajuda d'uns índexs d'identitat i de longitud. Això permet analitzar-los amb l'ajuda d'un conjunt de xarxes neuronals que donen una predicció global –promig de totes les prediccions particulars- de la fortalesa d'aquella relació d'equivalència. Aquesta predicció pren un valor en l'interval [0,1], considerant-se el valors superiors a 0.5 com a resultats positius (relació d'equivalència) i els altres com a negatius (no equivalència).

Finalment, per obtenir la predicció de quins esdeveniments són veritablement equivalents, s'apliquen els següents criteris als càlculs obtinguts amb les xarxes neuronals: per a cada esdeveniment equivalent ambigu –aquells on es recuperava més d'un candidat a ser equivalent amb l'ajuda de BLAST, s'escull aquell que té un nombre més gran de xarxes neuronals que donen l'equivalència com a bona i, en cas que n'hi hagi més d'un amb el mateix nombre, s'escull aquell que té una predicció global més alta (veure Figura 41).

El mètode té un cert grau de flexibilitat ja que en funció de les pròpies necessitats, l'usuari pot modificar certs paràmetres per restringir més o menys les prediccions d'equivalència, aquelles on es consideren els dos esdeveniments d'*splicing* alternatiu com a equivalents. Els paràmetres per defecte són el màxim de laxes possible.

Una manera d'augmentar la restricció per obtenir una predicció és exigir un valor de predicció global mínim, o exigint que almenys un cert percentatge de les xarxes neuronals considerin l'equivalència com a certa.

7.3 Disseny del mètode de predicció

El desenvolupament del mètode de predicció requerí abordar els punts que es descriuen en els següents apartats: l'origen de les dades, l'obtenció d'un conjunt de parelles d'esdeveniments equivalents, la caracterització dels esdeveniments mitjançant uns paràmetres d'identitat i mida, l'entrenament de la xarxa neuronal, l'elecció d'uns criteris de selecció i el càlcul d'unes figures de mèrit per avaluar la fiabilitat de les prediccions.

7.3.1 Origen de les dades

El desenvolupament de qualsevol mètode de predicció requereix un conjunt de dades que permetin tant el càlcul dels paràmetres associats al mètode com el contrast de la seva capacitat predictiva. En els nostre cas, aquest conjunt s'obtingué a partir de la versió 43 de la base de dades SwissProt (Boeckmann et al., 2003) mitjançant un senzill protocol de mineria de dades per recollir totes les entrades amb informació sobre *splicing* alternatiu. SwissProt només dóna una seqüència de proteïna per entrada, normalment la de la isoforma més llarga, i l'acompanya amb informació addicional sobre variacions (*splicing* alternatiu, polimorfismes...).

La manera d'obtenir les dades requerides fou seleccionant totes les entrades de la base de dades amb el camp "FT VARSPLIC", indicatiu d'un canvi en la seqüència de referència a causa de l'*splicing* alternatiu (cal indicar que, posteriorment, la base de dades ha etiquetat aquest camp com a "FT VAR_SEQ"). Així obtinguérem un conjunt d'esdeveniments equivalents, formats per la isoforma de referència i cadascuna de les isoformes alternatives que són a la base de dades. Després es van processar les dades per regenerar les seqüències de les isoformes alternatives a partir de la seqüència de referència i la informació donada a la base de dades, d'una manera anàloga a la proposada per Kersey i col·laboradors (Kersey et al., 2000). Aquest protocol de treball va permetre obtenir un conjunt de més de 14000 seqüències d'isoformes que van passar a ser la nostra base de dades d'isoformes. Addicionalment, per a cada esdeveniment es va guardar la informació sobre el mecanisme d'*splicing*: nombre d'insercions/delecions i substitucions que s'havien d'aplicar a la isoforma de referència per obtenir l'isoforma alternativa.

7.3.2 Obtenció d'un conjunt de parelles d'esdeveniments equivalents

Per obtenir les parelles d'esdeveniments equivalents vam utilitzar un protocol prèviament emprat al nostre grup (Valenzuela et al., 2004).

Primer, es va utilitzar el programa de cerca BLAST (Altschul et al., 1990) per fer una cerca de tots contra tots en la base de dades d'isoformes, descartant-se tots aquells candidats amb un e-valor més gran que 10^{-50}. A partir d'aquí, les isoformes candidates a ser equivalents van ser alineades amb l'algoritme d'alineament global proposat per Neddleman i Wunsch (Needleman and Wunsch, 1970), utilitzant la matriu BLOSUM62 (Henikoff and Henikoff, 1992) com a matriu de substitució d'aminoàcids.

Després, a partir dels alineaments obtinguts, es calcularen els següents índexs d'identitat (Valenzuela et al., 2004): l'índex d'identitat global, que indica el percentatge de residus alineats que són idèntics; l'index d'identitat local simple, que se centra en els residus que formen part de la zona d'*splicing* alternatiu; l'índex d'identitat local complex, que té en compte la zona d'*splicing* de les dues seqüències (veure Figura 42). De fet, l'índex d'identitat local complex és tan sols una variant més restrictiva de l'índex d'identitat local simple.

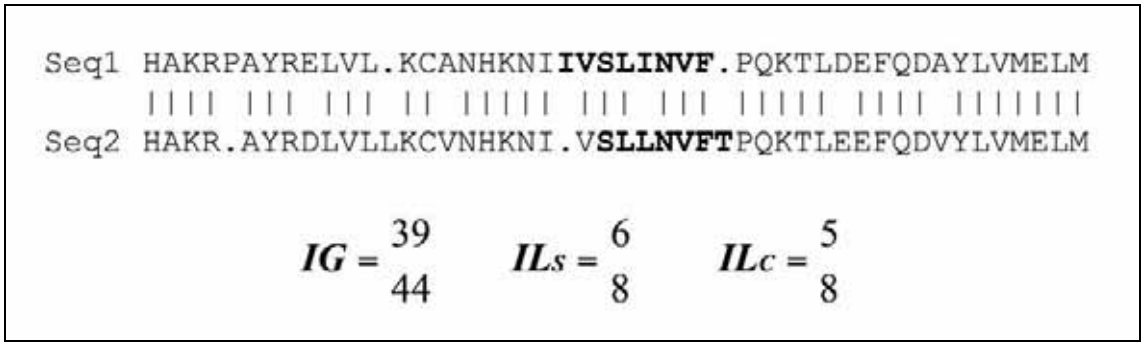

Figura 42. Exemple de càlcul dels índexs d'identitat. En aquest cas, la seqüència 1 (Seq1) és la que s'ha utilitzat com a referència, mentre la seqüència 2 (Seq2) és una de les candidates recuperades amb BLAST. En negreta es marquen les regions d'*splicing* alternatiu de les dues proteïnes. L'índex d'identitat global (IG) és el quocient entre el nombre de residus idèntics i el nombre de residus alineats. L'índex d'identitat local simple (IG_S) és el quocient entre el nombre de residus idèntics en la zona de l'*splicing* alternatiu de Seq1 i el nombre de residus afectats per l'*splicing* alternatiu de Seq1. L'índex d'identitat local complex (IG_C) és el quocient entre el nombre de residus idèntics que formen part de la zona d'*splicing* de les dues proteïnes (Seq1 i Seq2) i el nombre de residus afectats per l'*splicing* alternatiu de Seq1.

Cada isoforma de cada esdeveniment d'*splicing* alternatiu tenia un conjunt de seqüències candidates per ser isoformes equivalents. S'agruparen totes les candidates que pertanyien al mateix gen i, d'entre elles, s'escollí com a equivalent la que tenia els índexs d'identitat global i local més alts. Aquesta elecció es féu per totes les isoformes; per tant, per assolir una veritable relació d'equivalència entre les seqüències proteiques s'havia de complir un criteri de millor candidat en les dues direccions. A partir d'aquí es reconstruïren els esdeveniments equivalents, descartant aquells on apareixien ambigüitats.

Durant tot aquest protocol s'utilitzaren diverses restriccions per evitar incongruències: els canvis causats per l'*splicing* alternatiu havien de ser com a mínim de deu residus de longitud, per excloure errors de seqüenciació (Kondrashov and Koonin, 2001); i els percentatges d'identitat global i local entre les isoformes equivalents no podien ser inferiors al 50% (Valenzuela et al., 2004)(veure Figura 43), per excloure candidats provinents de gens no homòlegs al gen d'interés.

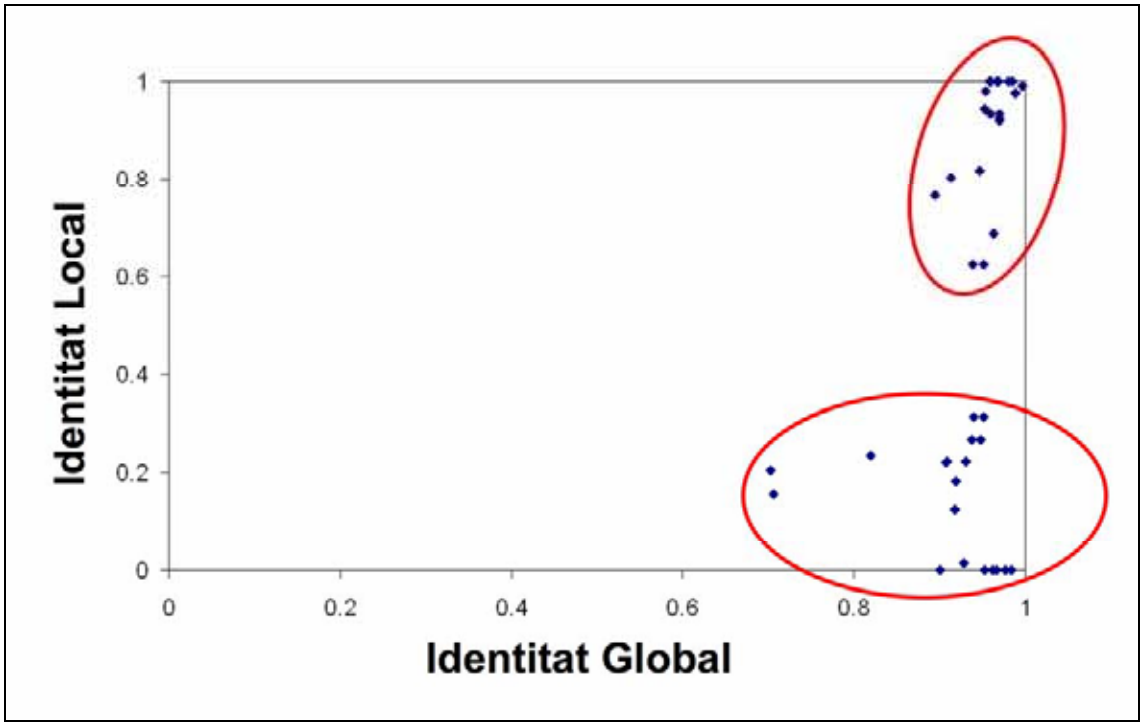

Figura 43. Diversos exemples d'alineaments entre isoformes equivalents i no equivalents. Tots els alineaments donen una identitat global per sobre de 0.6. En canvi, només les isoformes equivalents tenen una identitat local alta (el·lipse superior); per contra, els alineaments entre isoformes no equivalents donen unes identitats locals baixes (el·lipse inferior). En aquest cas s'ha utilitzat la identitat local simple.

A causa de la metodologia utilitzada, tots els possibles esdeveniments equivalents

havien de contenir les isoformes de referència de SwissProt, car aquestes eren necessàries per saber la longitud i localització dels canvis. Així mateix, només es van considerar les equivalències entre gens de diferent espècie.

En cas d'haver-hi esdeveniments complexos formats per més d'una inserció, deleció o substitució, o per una barreja d'aquestes, la longitud de l'esdeveniment es va comptabilitzar com la suma dels diversos canvis i les seves identitats locals, com els promigs ponderats de tots els canvis.

Finalment, també s'utilitzà com a restricció el mecanisme de l'*splicing* alternatiu. Es decidí descartar els esdeveniments equivalents trobats automàticament que no tenien el mateix nombre de substitucions i/o insercions/delecions.

Aquesta cerca automàtica va ser, posteriorment, revisada manualment. En la revisió, es van utilitzar criteris d'identitat de seqüència, localització dels canvis dins de la seqüència proteica i d'evidències funcionals: expressió diferencial (Ball et al., 1995; Jackson and Parham, 1988; Scorilas et al., 2002; Wada et al., 1992) i activitat biològica (Matsuda et al., 1992; O'Connor et al., 1998; Ogawa et al., 1994). Es van descartar, substituir o afegir noves equivalències segons era necessari. Així, s'obtingueren 473 parells diferents d'esdeveniments d'*splicing* alternatiu –els esdeveniments pertanyien a 321 proteïnes diferents de 17 espècies diferents, des d'adenovirus fins a mamífers.

Les equivalències descartades en la inspecció manual de les dades -86 en total- es van guardar per posar a prova el mètode de predicció amb posterioritat.

7.3.3 Paràmetres

Cada parella d'esdeveniments equivalents es pot caracteritzar amb un conjunt de paràmetres que n'estableixen l'equivalència i en base als quals les xarxes neuronals faran la seva predicció. Inspirats per Kondrashov i Koonin (Kondrashov and Koonin, 2001) i allò que havíem après en anteriors treballs (Valenzuela et al., 2004), nosaltres utilitzàrem els índexs d'identitat global i local simple i el quocient de longituds entre les diverses isoformes, ja que ens permeten analitzar tant els esdeveniments formats per insercions/delecions, com aquells constituïts per substitucions.

Es calcularen quatre índexs d'identitat global, quatre d'identitat local simple i quatre índexs de relació de mides per a cada parella d'esdeveniments equivalents. Aquests quatre índexs per a cada variable s'obtingueren comparant, en cada possible relació

d'equivalència, les isoformes equivalents entre elles i aquestes amb les que no ho són (veure Figura 44).

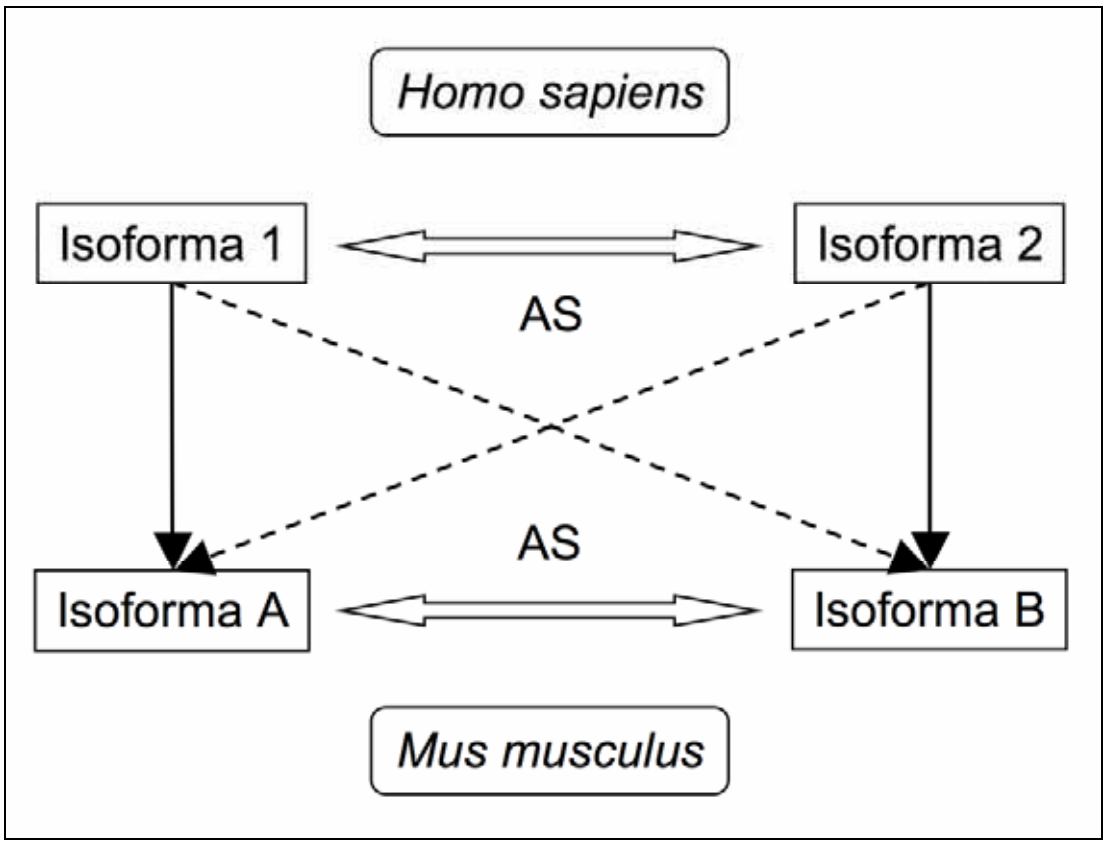

Figura 44. Esquema de les comparacions per caracteritzar els esdeveniments equivalents. Les dobles fletxes indiquen els esdeveniments d'*splicing* alternatiu. Les fletxes negres indiquen les comparacions fetes i la direcció. En totes les comparacions mostrades, les isoformes humanes es prenen com a referència. La fletxa contínua indica una comparació entre isoformes equivalents, mentre la fletxa discontínua indica una comparació entre isoformes no equivalents.

Els índexs d'identitat eren els mateixos que s'havien utilitzat en el protocol d'obtenció de parelles equivalents. Finalment, el darrer índex era el quocient entre la longitud de les dues seqüències.

Així, doncs, s'obtingueren uns vectors amb dotze paràmetres per caracteritzar totes les parelles d'esdeveniments equivalents.

7.3.4 Xarxa neuronal

El nostre problema s'ajusta perfectament a un problema convencional de reconeixement de patrons. Aquests problemes s'han intentat resoldre amb diverses eines (xarxes neuronals, *Support Vector Machines*...). El nostre grup es va decantar per l'ús de les

xarxes neuronals per l'experiència en aquest camp (de la Cruz et al., 2002; Ferrer-Costa et al., 2004) i la seva provada utilitat en l'estudi i resolució de problemes biològics (Garth et al., 1996; Jonsson et al., 1997).

Una xarxa neuronal és una eina d'intel·ligència artificial utilitzada habitualment en els problemes de reconeixement de patrons, que imita el funcionament del cervell (Basheer and Hajmeer, 2000). En el nostre cas, s'utilitzà una xarxa neuronal de tipus *feed-forward* (Rumelhart et al., 1995), amb una capa d'entrada, una capa oculta amb dos neurones i una capa de sortida (veure Figura 45). El vector d'entrada consistia en els dotze paràmetres abans esmentats i els estats de sortida eren 0 quan la relació d'equivalència era negativa i 1 quan era positiva.

Primer de tot, la xarxa ha de ser entrenada per què aprengui a classificar els patrons. Això es fa de manera supervisada, presentant-li uns vectors de dades amb el corresponent estat de sortida (Shepherd et al., 1999). A partir d'aquí, la xarxa començarà a iterar l'aprenentatge d'aquests vectors per trobar una manera de separar els diversos estats mostrats, de manera que acabarà optimitzant uns pesos, normalment inicialitzats a l'atzar, que li permetin construir el millor model matemàtic possible. L'optimització dels pesos es féu amb el mètode de gradients conjugats escalats (Basheer and Hajmeer, 2000; Shepherd et al., 1999) durant un màxim de 500 iteracions. Per a cada predicció, la xarxa calcula un valor de sortida entre 0 i 1. Si el valor és superior a 0.5, el pren com a 1, si no, com a 0.

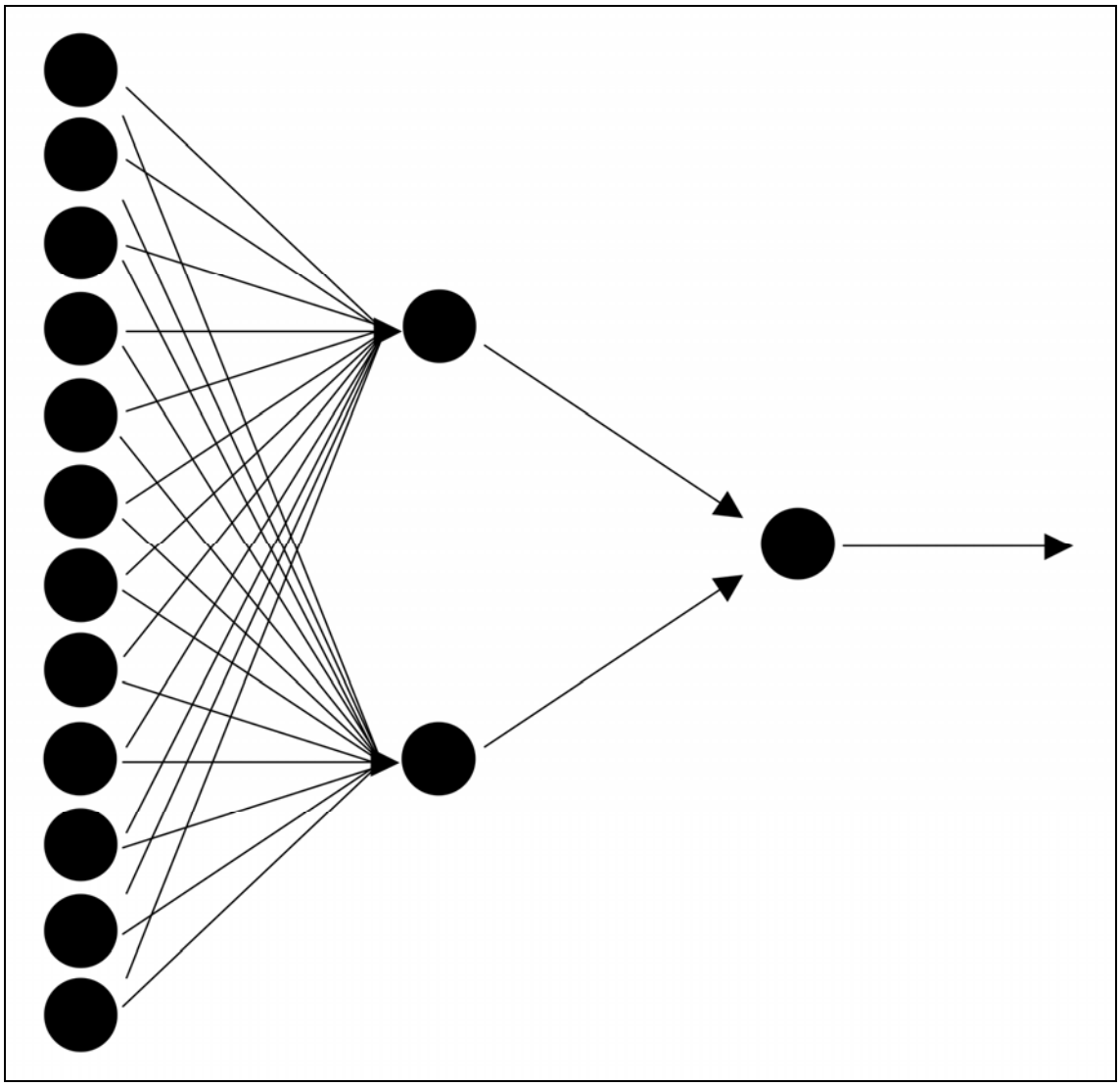

Figura 45. Esquema de la xarxa neuronal utilitzada.

7.3.5 Validació creuada

Per comprovar la fiabilitat de les eines basades en aprenentatge s'acostuma a utilitzar la tècnica de la validació creuada (Krishnan and Westhead, 2003). En aquesta tècnica, el conjunt de dades es divideix a l'atzar en dos subconjunts –entrenament i test. El primer subconjunt s'utilitza per entrenar la xarxa neuronal, mentre que l'altre permet mesurar la capacitat de predicció que ha adquirit.

En el nostre cas, els esdeveniments es van agrupar en funció dels gens i es dividiren en dos subconjunts per realitzar la validació creuada. La divisió es féu a l'atzar i mantenint tots els esdeveniments d'*splicing* alternatiu d'un mateix gen agrupats. La separació en funció dels gens es va fer per evitar que dades corresponents a un mateix gen estiguessin repartides entre els dos conjunts, ja que això pot provocar una

sobreestimació de la capacitat predictiva de la xarxa (Krishnan and Westhead, 2003).

7.3.6 Entrenament de les xarxes neuronals

En l'entrenament supervisat de la xarxa se li han de presentar els vectors de paràmetres amb el seu corresponent estat de sortida (Shepherd et al., 1999). Per tant, s'havia de tenir un conjunt d'esdeveniments positius i un altre conjunt de negatius. Els esdeveniments positius eren els 473 parells d'esdeveniments que s'havien trobat amb el protocol explicat anteriorment i després s'havien revisat manualment. Per tenir un conjunt d'esdeveniments inequívocament negatius s'optà per fer-ho a partir dels esdeveniments positius, invertint l'equivalència de les isoformes o substituint una de les isoformes equivalents per una altra isoforma. D'aquesta manera es van construir prop de 5000 parells negatius.

A causa de la diferència de mida entre els conjunts de vectors positius i negatius es va aplicar un protocol estàndard per remostrejar els conjunts d'entrenament (Basheer and Hajmeer, 2000; Heckerling et al., 2003): s'augmentà la quantitat de positius fins a obtenir una proporció 1:1, mitjançant la duplicació a l'atzar dels vectors positius – protocol repetit 100 cops per a cada conjunt d'entrenament. D'aquesta manera s'obtingueren 200 conjunts d'entrenament diferents, que es van posar a prova amb els dos subconjunts originals, sense remostrejar (veure Figura 46). Així doncs, s'entrenaren 200 xarxes neuronals diferents. Cada xarxa donava una predicció particular que després es promitjava per obtenir una predicció global única.

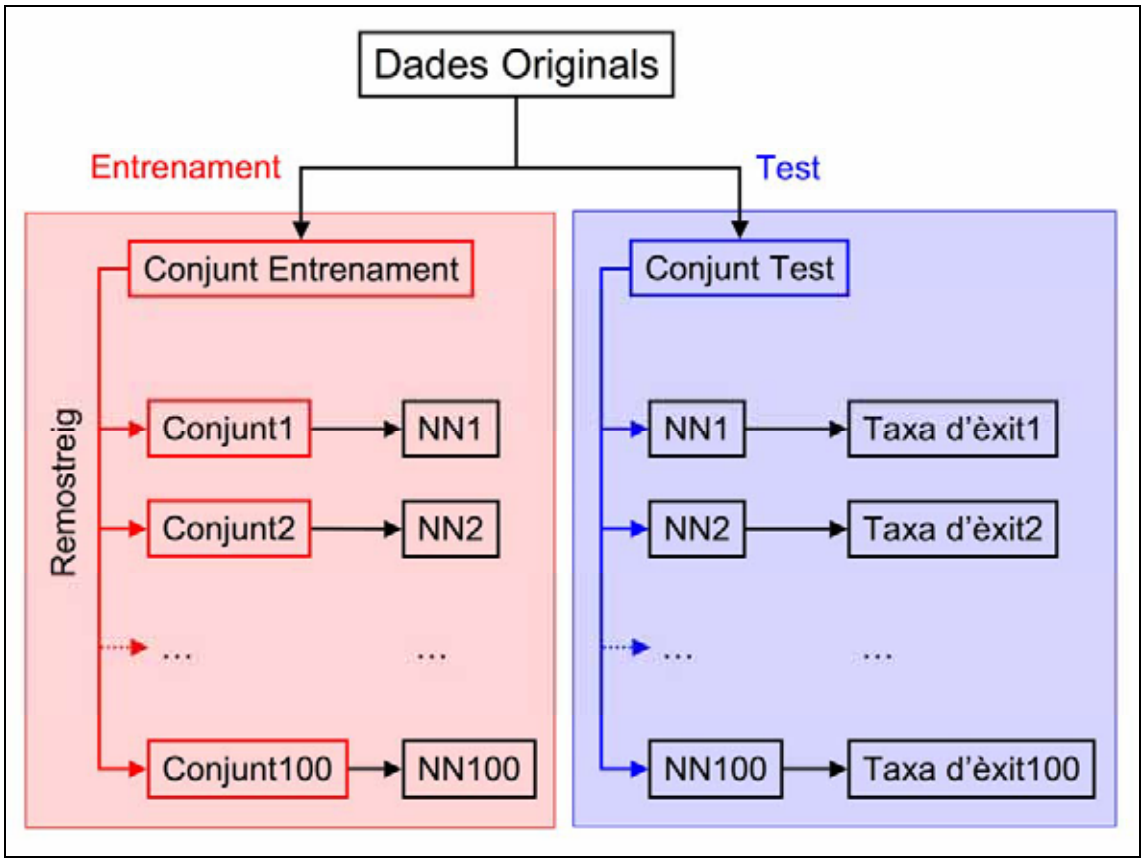

Figura 46. Esquema de l'entrenament de les xarxes neuronals. Les dades originals són separades en dos conjunts (entrenament i test). El conjunt d'entrenament és remostrejat per equilibrar el nombre de patrons positius i negatius. La xarxa s'entrena amb cadascun dels subconjunts d'entrenament resultants i és provada amb el conjunt de test. Cada prova genera unes figures de mèrit que seran promitjades per calcular la taxa d'èxit de les xarxes.

7.3.7 Criteris de selecció

Pot passar que, per a un esdeveniment d'*splicing* alternatiu, les xarxes acabin predint com a equivalent més d'un esdeveniment candidat. Evidentment, això no és possible i comporta una ambigüitat de la predicció. Per solucionar aquest problema es va decidir utilitzar uns criteris de selecció a posteriori. Després d'obtenir les prediccions de les xarxes, quan hi hagués un problema d'ambigüitat en la predicció, s'escolliria un sol dels esdeveniments com a equivalent.

Es van provar dos criteris de selecció diferents: el promig de valors de sortida –la predicció global de les xarxes- i el percentatge de xarxes amb prediccions positives. Òbviament, en els dos casos s'escollí el valor més alt. Finalment, es decidí utilitzar el percentatge de xarxes amb prediccions positives com a criteri principal i l'altre com a secundari.

7.3.8 Figures de mèrit de les xarxes

Per avaluar la capacitat predictiva de les xarxes neuronals i la fiabilitat del mètode s'utilitzaren diverses figures de mèrit corrents: l'exactitud (Equació 7) –mesura l'èxit global de la predicció, la precisió (Equació 8) –mesura la capacitat de diferenciar entre positius i negatius-, la sensibilitat (Equació 9) –mesura la capacitat d'identificar els positius- i l'especificitat (Equació 10) –mesura la capacitat d'identificar els negatius.

$$\textit{Exactitud} = \frac{tp + tn}{tp + tn + fp + fn} \qquad \text{(Equació 7)}$$

$$\textit{Precisió} = \frac{tp}{tp + fp} \qquad \text{(Equació 8)}$$

$$\textit{Sensibilitat} = \frac{tp}{tp + fn} \qquad \text{(Equació 9)}$$

$$\textit{Especificitat} = \frac{tn}{tn + fp} \qquad \text{(Equació 10)}$$

on tp són les prediccions positives que s'han encertat; tn, les prediccions negatives que s'han encertat; fp, les prediccions positives que s'han fallat i fn, les prediccions negatives que s'han fallat.

7.3.9 Tests del mètode de predicció

Es feren diverses proves per calcular el rendiment del protocol de predicció (veure Figura 41).

En primer lloc, es buscaren esdeveniments equivalents per aquells esdeveniments que s'havien utilitzat en l'entrenament de les 200 xarxes neuronals (veure Test 1 a la Taula 16). En aquest cas, tots els esdeveniments tenien almenys un equivalent. Cada esdeveniment fou predit sols amb 100 xarxes, per evitar utilitzar les mateixes xarxes de l'entrenament en la predicció. Per calcular les figures de mèrit del mètode s'utilitzaren tan sols els esdeveniments positius i negatius de l'entrenament de la xarxa, sense tenir en compte si hi havia altres prediccions.

Després, els esdeveniments pels quals es buscava equivalents foren aquells que s'havien descartat en la inspecció ocular de les dades (veure Test 2 a la Taula 16). Per fer la mesura del rendiment, s'utilitzaren els esdeveniments positius de l'entrenament de la xarxa i com a negatius els descartats. Com abans, no es tingueren en compte noves

prediccions.

Finalment, es féu una prova modificant la base de dades d'isoformes. Es buscaren equivalències pels esdeveniments de l'entrenament de les xarxes, però abans s'havien eliminat una o les dues isoformes equivalents de la base de dades. Així, qualsevol predicció d'equivalència que es fes seria un error (veure Test 3 a la Taula 16).

Test 1	Positius	Selecció de casos recuperats amb el protocol ad hoc i la inspecció manual
	Negatius	Altres combinacions que es poden fer modificant l'ordre o alguna isoforma de les parelles equivalents
Test 2	Positius	Selecció de casos recuperats amb el protocol ad hoc i la inspecció manual
	Negatius	Casos recuperats amb el protocol ad hoc i descartats durant la inspecció manual
Test 3	Positius	No n'hi ha
	Negatius	Selecció de casos recuperats amb el protocol ad hoc i la inspecció manual

Taula 16. Tests del mètode de predicció.

7.4 Resultats dels tests

A continuació es mostren els resultats pels diferents tests, tant per les xarxes neuronals com pel mètode sencer. En el cas de les xarxes, férem els tests 1 i 2 per avaluar la seva capacitat predictiva: el test 1 és el construït per entrenar, mentre que el test 2 té casos negatius més complicats. D'altra banda, per avaluar el mètode, es feren els tests 1 i 3: el test 1 ens permet veure la capacitat de trobar els veritables positius, quan sabem segur que n'hi ha, mentre que el test 3 mesura l'habilitat per descartar falsos positius. En l'anàlisi del nostre mètode, hem comparat els resultats amb un mètode control consistent en elegir el millor candidat de cada cerca BLAST.

7.4.1 Capacitat predictiva de les xarxes neuronals

La següent taula (Taula 17) ens presenta les figures de mèrit obtingudes en l'entrenament de les xarxes neuronals. Aquestes mesures estan calculades a partir de les prediccions fetes sobre el conjunt de test.

	Test 1	Test 2
Exactitud	0.89±0.01	0.82±0.02
Precisió	0.46±0.06	0.83±0.01
Sensibilitat	0.94±0.02	0.94±0.02
Especificitat	0.88±0.01	0.47±0.05

Taula 17. Figures de mèrit de les xarxes neuronals.

Com es veu, el mètode és molt sensible però poc precís, és a dir, detecta gairebé totes les relacions d'equivalència veritables, però la meitat de les prediccions positives són errònies.

La precisió de les prediccions també es va calcular utilitzant criteris de selecció a posteriori. Així, quan s'utilitza algun dels criteris de selecció –valor promig de les xarxes o percentatge de prediccions positives, la precisió de les prediccions augmenta molt (veure Taula 18). La causa d'aquest augment cal buscar-la en què no totes les prediccions positives tenen la mateixa fiabilitat.

Test 1	Precisió
Promig de valors de sortida	0.90±0.02
Percentatge de prediccions positives	0.97±0.01

Taula 18. Precisió aplicant criteris de selecció.

Això ens indica que les xarxes tenen una gran utilitat en el nostre problema, però que s'han d'utilitzar criteris de selecció a posteriori.

7.4.2 Poder predictiu del mètode

Per contrastar el poder predictiu de la metodologia explicada amb anterioritat (veure Figura 41), es va decidir comparar les figures de mèrit d'SPLASH amb les d'un altre protocol més senzill: quedar-se amb el millor candidat per a cada cerca amb BLAST.

La Taula 20 ens ensenya el percentatge d'esdeveniments que obtenen alguna predicció. Només s'han tingut en compte els esdeveniments que se sap que tenen esdeveniments equivalents.

Test 1	SPLASH	Mètode control
Percentatge de prediccions	0.94±0.01	0.99±0.01

Taula 19. Percentatge de prediccions.

Evidentment, el percentatge de prediccions d'SPLASH mai podrà ser superior al de BLAST, car aquest darrer és utilitzat en el mètode de predicció d'SPLASH en una etapa inicial. No obstant això, el que és interessant de remarcar en aquest punt és que la pèrdua de prediccions és molt petita després d'utilitzar tot el nostre protocol (veure Figura 41 pel protocol).

La Taula 20 mostra els resultats tenint en compte els mateixos esdeveniments utilitzats en l'entrenament de la xarxa. Per tant, a l'hora de calcular les figures de mèrit, es consideren com a negatius tots els esdeveniments falsos utilitzats en l'entrenament de les xarxes. Concordant amb els resultats mostrats a la Taula 19, la sensibilitat és més alta pel mètode de control. En canvi, les figures que també tenen en compte els negatius –especificitat, precisió i exactitud- són superiors per SPLASH.

Test 1	SPLASH	Mètode control
Exactitud	0.99±0.00	0.96±0.04
Precisió	0.98±0.01	0.73±0.31
Sensibilitat	0.93±0.02	0.97±0.01
Especificitat	1.00±0.00	0.96±0.05

Taula 20. Figures de mèrit del mètode de predicció.

És interessant remarcar que tant sols en un 25% dels casos en que no s'obté la predicció correcta, se n'obté una d'errònia. És a dir, en un 75% d'aquests casos el que passa és que no s'obté cap mena de predicció –ja sigui perquè no es recuperen candidats amb BLAST o perquè les xarxes neuronals els valoren desfavorablement.

Finalment, la Taula 21 mostra els resultats del Test 3: els percentatge de prediccions positives obtingudes quan, abans d'utilitzar el mètode, s'eliminen de la base de dades una o les dues isoformes equivalents. Evidentment, aquestes prediccions són incorrectes. Observem que el nostre mètode es comporta millor que el control basat en l'ús de BLAST.

Test 3	SPLASH	Mètode control
Manca d'una isoforma	0.81±0.04	0.71±0.04
Manca de dues isoformes	0.91±0.02	0.88±0.05

Taula 21. Especificitat del mètode de predicció.

7.5 Discussió dels possibles errors

Es féu un control exhaustiu tant de les prediccions de les xarxes neuronals com de les obtingudes amb tot el protocol (veure Figura 41) per detectar possibles errors o biaixos en el mètode.

7.5.1 Xarxes neuronals

Un possible error podia ser a causa de la identitat global entre les isoformes equivalents. La Taula 22 ens mostra que, per bé que a mesura que la identitat entre les isoformes baixa també ho fa la precisió de les prediccions, aquest descens és petit i fins a identitats del 60% podem considerar les prediccions força fiables.

Identitat promig entre isoformes equivalents	Precisió
90%	0.98±0.01
80%	0.96±0.03
70%	0.93±0.01
60%	0.92±0.12

Taula 22. Precisió depenent de la identitat.

Una altra font de problemes podia ser el mecanisme de l'esdeveniment d'*splicing* alternatiu. Dividirem el conjunt d'esdeveniments en insercions/delecions, substitucions i esdeveniments complexes –aquells formats per insercions/delecions i substitucions alhora.

La Taula 23 ens mostra com la precisió per les substitucions és gairebé perfecta, mentre en el cas de les insercions/delecions i els esdeveniments complexes hi havia un percentatge d'errors minso, però més gran.

Mecanisme de l'esdeveniment	Precisió
Insercions/delecions	0.97±0.00
Substitucions	0.99±0.01
Complexes	0.95±0.07

Taula 23. Precisió depenent del tipus d'esdeveniment d'*splicing*.

Els resultats de la Taula 23 ens feren pensar que era probable que el fet d'haver de comparar dos delecions devia ser més difícil. Per tant, analitzàrem la precisió de les prediccions tenint en compte la mida de les insercions/delecions. Es consideraren com a insercions o delecions de mida petita aquelles fins a 30 residus, les de mida més gran es catalogaren com a grans. La Taula 24 ens mostra que la mida de la part delecionada afecta la qualitat de la predicció, que no obstant això continua sent molt acurada.

Mida de la inserció/deleció	Precisió
Petites	0.99±0.00
Grans	0.94±0.01

Taula 24. Precisió depenent de la mida de la inserció/deleció.

En resum, les xarxes neuronals utilitzades en el nostre mètode prediuen molt bé totes les substitucions i les insercions i/o delecions de mida petita.

A mesura que la identitat entre les seqüències baixa també ho fa la precisió de les prediccions.

7.5.2 Mètode de predicció

Igual que s'observava en les xarxes, a mesura que la identitat entre les isoformes va baixant, també ho fan la precisió i la sensibilitat del mètode (Taula 25), però fins i tot a identitats mitjanes (60%) els resultats són força satisfactoris.

Identitat promig entre isoformes equivalents	Exactitud	Precisió	Sensibilitat
90% (N=305)	0.99±0.00	0.96±0.02	0.95±0.01
80% (N=89)	0.98±0.02	0.89±0.06	0.89±0.06
70% (N=25)	0.93±0.04	0.83±0.05	0.83±0.05
60% (N=13)	0.95±0.06	0.83±0.24	0.83±0.24

Taula 25. Precisió depenent de la identitat.

A la Taula 26 podem veure com la sensibilitat de les insercions/delecions es més baixa que en les substitucions o els esdeveniments complexes. Això afecta, òbviament, a la precisió. A diferència del que passava en les xarxes neuronals (veure Taula 23), sembla que aquí el fet de tenir diversos canvis en un mateix esdeveniment aporta fiabilitat al mètode.

Mecanisme de l'esdeveniment	Exactitud	Precisió	Sensibilitat
Insercions/delecions (N=248)	0.98±0.01	0.91±0.03	0.90±0.02
Substitucions (N=147)	0.99±0.01	0.96±0.04	0.95±0.03
Complexes (N=78)	1.00±0.00	0.99±0.02	0.99±0.02

Taula 26. Precisió depenent del tipus d'esdeveniment d'*splicing*.

Finalment, la Taula 27 ens mostra que no hi ha diferències ni en la precisió ni en la sensibilitat a causa de la mida de les insercions i/o delecions.

Mida de la inserció/deleció	Exactitud	Precisió	Sensibilitat
Petites (N=145)	0.98±0.00	0.90±0.01	0.90±0.00
Grans (N=103)	0.98±0.01	0.91±0.05	0.91±0.05

Taula 27. Precisió depenent de la mida de la inserció/deleció.

Aquests resultats es veuen reforçats pel fet que dels casos en que l'error porta a una predicció errònia, tots són en esdeveniments amb una sola deleció i, pràcticament, la meitat en delecions de mida petita.

Per acabar aquest anàlisi d'errors es miraren detingudament aquell 25% de casos negatius del test 2 que obtenien una predicció d'equivalència. Es veié que les identitats global i local entre les isoformes equivalents era molt alta (94.4% i 73.4%, respectivament). A més a més, sorprenentment, aquí ens trobàrem que prop de dos terços dels errors eren en esdeveniments amb substitucions i que un 10% eren esdeveniments complexes. Així doncs, els falsos positius del mètode apareixen bàsicament en esdeveniments on se substitueixen fragments bastant semblants i que no tenen cap esdeveniment equivalent.

7.6 Discussió

La identificació de les isoformes que participen en determinats processos biològics – naturals o aberrants- és un tema molt important en biomedicina (Cuperlovic-Culf et al.,

2006). No obstant això, actualment, se sap molt poc de les funcions o característiques de la majoria de variants d'*splicing*. Nosaltres hem desenvolupat un mètode que permet la identificació d'esdeveniments d'*splicing* alternatiu homòlegs o equivalents (veure Figura 40). Aquest mètode, que està basat en un ús combinat de cerques de seqüències similars i xarxes neuronals per a valorar les possibles equivalències, té un poder predictiu prou bo –tant pel que fa a la precisió, com a l'especificitat- per fer pensar que s'ha iniciat un pas cap al desenvolupament de protocols d'anotació funcional automàtica dels esdeveniments d'*splicing*.

RESUM

8 Resum

Des del descobriment del mecanisme d'*splicing* ja es va postular l'existència de l'*splicing* alternatiu i les seves possibles implicacions en l'evolució de la funció gènica. Tanmateix, no ha estat fins els darrers anys que aquest fenomen ha pres protagonisme, vist sovint com un altre nivell de control per generar variabilitat en el proteoma. Així, quan hi ha errors, apareixen isoformes patològiques, les quals s'han relacionat amb moltes i diverses malalties.

A nivell d'àcids nucleics, l'*splicing* alternatiu es pot donar per diverses vies, que resulten en canvis a nivell proteic entre les isoformes. Aquests canvis, al seu torn, poden donar variacions estructurals, causa de les diferències funcionals. La bioinformàtica ha participat en l'estudi de l'*splicing* alternatiu a tots aquests nivells. Així, ha estat utilitzada com una eina de suport en projectes de seqüenciació genòmica i de microxips. Per altra banda, també s'ha utilitzat per estudiar la complexitat del proteoma –tant a nivell de freqüència, com de la conservació dels canvis- i l'impacte estructural i funcional de l'*splicing* alternatiu.

El treball presentat en aquesta tesi es basa en l'estudi dels efectes estructurals i funcionals en les proteïnes causats per l'*splicing* alternatiu. D'aquesta manera, hem analitzat el rol de l'*splicing* alternatiu com a font de variabilitat proteica, n'hem estudiat la conservació interespecífica dels efectes, ens hem centrat en una família funcional de proteïnes per analitzar-ne les variants d'*splicing* i hem cercat un protocol per a la identificació d'esdeveniments d'*splicing* alternatiu equivalents.

El primer punt que hem analitzat és la relació entre l'*splicing* alternatiu i la duplicació gènica, que són dos processos implicats en la diversificació del proteoma. Darrerament, s'ha trobat una anticorrelació entre la presència de variants d'*splicing* i duplicats i s'han descobert uns casos particulars d'una possible intercanviabilitat funcional entre els dos fenòmens. Per tot plegat, ens hem decidit a comparar les dues fonts de variabilitat proteica, tant des del punt de vista genòmic com proteòmic. Allò què hem vist és que la presència d'*splicing* alternatiu i duplicació gènica estan inversament relacionades, però s'ha descartat la hipòtesi d'intercanviabilitat funcional perquè els gens amb *splicing* alternatiu i amb duplicats tenen distribucions funcionals similars i els efectes proteics que provoquen els dos fenòmens són molt diferents. Per resoldre aquesta paradoxa, nosaltres proposem fixar-nos en l'escenari evolutiu i l'equilibri de la dosi gènica com a

paràmetres essencials per explicar l'anticorrelació.

Posteriorment, hem analitzat si els efectes de l'*splicing* alternatiu sobre la modulació funcionals són semblants en diferents espècies. Per començar, hem caracteritzat els efectes sobre les proteïnes de l'*splicing* alternatiu en quatre espècies diferents i, després, hem comparat esdeveniments equivalents o homòlegs. Els nostres resultats mostren que en espècies diferents hi ha una gran conservació pel que fa a la manera que l'*splicing* alternatiu modula la funció proteica. Per tant, sembla que les diferències de complexitat entre els organismes no estan causades per un ús diferencial dels mecanismes de l'*splicing* alternatiu a l'hora de modificar la funció de les proteïnes, sinó que haurem de parar esment a altres possibilitats.

Per acabar l'estudi dels efectes de l'*splicing* alternatiu en la variabilitat proteica i funcional, ens hem centrat en els factors de transcripció. Aquestes proteïnes estan construïdes de forma modular i l'*splicing* alternatiu genera isoformes amb diferents habilitats. En el nostre treball, ens hem interessat pel mecanisme per generar diverses isoformes reguladores de la transcripció i la seva conservació. Així, veiem que per respondre a la multiplicitat d'estímuls, l'*splicing* alternatiu modifica uns dominis més sovint que uns altres, però ho fa d'una manera poc precisa, afectant fragments de dominis i regions properes. A més, la conservació estructural i funcional dels efectes de l'*splicing* alternatiu és molt alta.

Finalment, hem decidit proposar un mètode per a la cerca d'esdeveniments homòlegs d'*splicing* alternatiu. L'objectiu ha estat ajudar en els camps de la biomedicina i la farmacogenòmica, on és molt important conèixer perfectament el rol de l'*splicing* alternatiu i de cada isoforma en particular. El mètode treballa a partir d'un esdeveniment d'*splicing* alternatiu i, a partir d'aquí, intenta trobar altres esdeveniments que siguin homòlegs o equivalents. Observant les figures de mèrit dels tests que hem realitzat, creiem que el mètode funciona amb una bona fiabilitat, fet que ens porta a pensar que pot ajudar en l'anotació funcional de les isoformes i en l'elecció de models animals adequats.

CONCLUSIONS

9 Conclusions

Les principals conclusions que s'extreuen del treball realitzat en aquesta tesi són les següents:

- Existeix una anticorrelació a nivell genòmic entre l'*splicing* alternatiu i la duplicació gènica, que no es pot explicar per la preferència per un dels dos mecanismes de certes famílies funcionals.
- Les insercions/delecions que genera l'*splicing* alternatiu en la seqüència de proteïnes són, en general, de mida molt superior a les que es generen després de la duplicació gènica i, a més, no acostumen a afectar les mateixes regions de la seqüència.
- Les substitucions de residus entre les variants d'*splicing* tendeixen a involucrar canvis importants de les propietats físico-químiques dels aminoàcids i a estar concentrades en petits fragments de la seqüència. En canvi, els duplicats tenen substitucions menys dràstiques i espargides per tota la proteïna.
- Els mecanismes globals d'increment de la diversitat proteica per raó de l'*splicing* alternatiu estan conservats en diferents espècies.
- Les isoformes homòlogues tenen un alt grau de conservació de les propietats físico-quimiques originades per l'*splicing* alternatiu.
- L'*splicing* alternatiu actua selectivament sobre els dominis funcionals dels factors de transcripció, però, en canvi els seus efectes són poc precissos.
- Els factors de transcripció ortòlegs tenen una gran conservació estructural i funcional, fins i tot independent de la seva capacitat per generar diverses isoformes o isoformes específiques.
- És possible tenir un mètode automàtic que permeti trobar esdeveniments equivalents d'*splicing* alternatiu, per ajudar en l'anotació funcional i en les recerques biomèdiques i farmacogenòmiques.

BIBLIOGRAFIA

10 Bibliografia

Abdel-Rahman, A., Shetty, A. K., and Abou-Donia, M. B. (2002). Disruption of the Blood-Brain Barrier and Neuronal Cell Death in Cingulate Cortex, Dentate Gyrus, Thalamus, and Hypothalamus in a Rat Model of Gulf-War Syndrome. Neurobiology of Disease *10*, 306-326.

Aerbajinai, W., Ishihara, T., Arahata, K., and Tsukahara, T. (2002). Increased expression level of the splicing variant of SIP1 in motor neuron diseases. Int J Biochem Cell Biol *34*, 699-707.

Akgul, C., Moulding, D. A., and Edwards, S. W. (2004). Alternative splicing of Bcl-2-related genes: functional consequences and potential therapeutic applications. Cell Mol Life Sci *61*, 2189-2199.

Altschmied, J. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *161*, 259-267.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Arney, K. L., and Fisher, A. G. (2004). Epigenetic aspects of differentiation. J Cell Sci *117*, 4355-4363.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Ast, G. (2004). How did alternative splicing evolve? *5*, 773-782.

Auboeuf, D., Honig, A., Berget, S. M., and O'Malley, B. W. (2002). Coordinate regulation of transcription and splicing by steroid receptor coregulators. Science *298*, 416-419.

Azubel, M., Habib, N., Sperling, R., and Sperling, J. (2006). Native spliceosomes assemble with pre-mRNA to form supraspliceosomes. J Mol Biol *356*, 955-966.

Baek, D., and Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *102*, 12813-11288.

Ball, C. L., Hunt, S. P., and Robinson, M. S. (1995). Expression and localization of

alpha-adaptin isoforms. J Cell Sci *108 (Pt 8)*, 2865-2875.

Barrier, M., Robichaux, R. H., and Purugganan, M. D. (2001). Accelerated regulatory gene evolution in an adaptive radiation. Proc Natl Acad Sci U S A *98*, 10208-10213.

Basheer, I. A., and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods *43*, 3-31.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L.*, et al.* (2004). The Pfam protein families database. Nucleic Acids Res *32*, D138-141.

Batsche, E., Yaniv, M., and Muchardt, C. (2006). The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat Struct Mol Biol *13*, 22-29.

Beaumont, C., Porcher, C., Picat, C., Nordmann, Y., and Grandchamp, B. (1989). The mouse porphobilinogen deaminase gene. Structural organization, sequence, and transcriptional analysis. J Biol Chem *264*, 14829-14834.

Benmoyal-Segal, L., Vander, T., Shifman, S., Bryk, B., Ebstein, R. P., Marcus, E. L., Stessman, J., Darvasi, A., Herishanu, Y., Friedman, A., and Soreq, H. (2005). Acetylcholinesterase/paraoxonase interactions increase the risk of insecticide-induced Parkinson's disease. Faseb J *19*, 452-454.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol *229*, 1065-1082.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2007). GenBank. Nucleic Acids Res *35*, D21-25.

Bentley, D. L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol *17*, 251-256.

Berget, S. (1995). Exon recognition in vertebrate splicing. J Biol Chem *270*, 2411 - 2414.

Berget, S. M., and Sharp, P. A. (1977). A spliced sequence at the 5[prime]-terminus of adenovirus late mRNA. *29*, 332-344.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T.*, et al.* (2006). Ensembl 2006. Nucleic Acids Res *34*, D556-561.

Black, D. L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell *103*, 367-370.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry *72*, 291-336.

Blanchette, M., Green, R. E., Brenner, S. E., and Rio, D. C. (2005). Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila 10.1101/gad.1314205. Genes Dev *19*, 1306-1314.

Blencowe, B. J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends in Biochemical Sciences *25*, 106-110.

Blencowe, B. J. (2006). Alternative splicing: new insights from global analyses. Cell *126*, 37-47.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I.*, et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res *31*, 365-370.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Bomze, H. M., and Lopez, A. J. (1994). Evolutionary conservation of the structure and expression of alternatively spliced Ultrabithorax isoforms from Drosophila. *136*, 965-977.

Boue, S., Vingron, M., Kriventseva, E., and Koch, I. (2002). Theoretical analysis of alternative splice forms using computational methods. Bioinformatics *18 Suppl 2*, S65-73.

Bracco, L., and Kearsey, J. (2003). The relevance of alternative RNA splicing to pharmacogenomics. Trends in Biotechnology *21*, 346-353.

Brenner, T., Hamra-Amitay, Y., Evron, T., Boneva, N., Seidman, S., and Soreq, H. (2003). The role of readthrough acetylcholinesterase in the pathophysiology of myasthenia gravis. Faseb J *17*, 214-222.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett *474*, 83 - 86.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat Genet *30*, 29 - 30.

Brinkman, B. M. N. (2004). Splice variants as cancer biomarkers. Clinical Biochemistry Special Issue: Recent Advances in Cancer Biomarkers *37*, 584-594.

Brudno, M., Gelfand, M., Spengler, S., Zorn, M., Dubchak, I., and Conboy, J. (2001). Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res *29*, 2338 - 2348.

Buee, L., Bussiere, T., Buee-Scherrer, V., Delacourte, A., and Hof, P. R. (2000). [tau] Protein isoforms, phosphorylation and role in neurodegenerative disorders. *33*, 95-130.

Buratti, E., and Baralle, F. E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. *24*, 10505-10514.

Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res *28*, 4364-4375.

Bustamante, C. D., Nielsen, R., and Hartl, D. L. (2002). A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *19*, 110-117.

Caceres, J. F., and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet *18*, 186-193.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N*., et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet *22*, 231-238.

Cartegni, L., and Krainer, A. R. (2003). Correction of disease-associated exon skipping by synthetic exon-specific activators. Nat Struct Biol *10*, 120-125.

Castle, J., Garrett-Engele, P., Armour, C., Duenwald, S., Loerch, P., Meyer, M., Schadt, E., Stoughton, R., Parrish, M., Shoemaker, D., and Johnson, J. (2003). Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biology *4*, R66.

Caudevilla, C., Serra, D., Miliar, A., Codony, C., Asins, G., Bach, M., and Hegardt, F. G. (1998). Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. Proc Natl Acad Sci U S A *95*, 12185-12190.

Clark, F., and Thanaraj, T. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet *11*, 451 - 464.

Clark, T., Sugnet, C., and Ares, M. (2002). Genome-wide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907 - 910.

Cline, M. S., Blume, J., Cawley, S., Clark, T. A., Hu, J.-S., Lu, G., Salomonis, N., Wang, H., and Williams, A. (2005). ANOSVA: a statistical method for detecting splice variation from expression data
10.1093/bioinformatics/bti1010. Bioinformatics *21*, i107-115.

Cohen, O., Erb, C., Ginzberg, D., Pollak, Y., Seidman, S., Shoham, S., Yirmiya, R., and Soreq, H. (2002). Neuronal overexpression of "readthrough" acetylcholinesterase is associated with antisense-suppressible behavioral impairments. Mol Psychiatry *7*, 874-885.

Cowper, A. E., Caceres, J. F., Mayeda, A., and Screaton, G. R. (2001). Serine-Arginine (SR) Protein-like Factors That Antagonize Authentic SR Proteins and Regulate Alternative Splicing
10.1074/jbc.M103967200. J Biol Chem *276*, 48908-48914.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J. S. (2000). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat Genet *24*, 340-341.

Cuperlovic-Culf, M., Belacel, N., Culf, A. S., and Ouellette, R. J. (2006). Data analysis of alternative splicing microarrays. Drug Discovery Today *11*, 983-990.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M., and Clamp, M. (2004). The Ensembl automatic gene annotation system. Genome Res *14*, 942-950.

Cusack, B. P., and Wolfe, K. H. (2005). Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *22*, 2198-2208.

Chauhan, A. K., Iaconcig, A., Baralle, F. E., and Muro, A. F. (2004). Alternative splicing of fibronectin: a mouse model demonstrates the identity of in vitro and in vivo systems and the processing autonomy of regulated exons in adult mice. Gene *324*, 55-63.

Chen, L. L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. *24*, 4271-4280.

Choi, H. S., Chung, M., Tzameli, I., Simha, D., Lee, Y. K., Seol, W., and Moore, D. D. (1997). Differential transactivation by two isoforms of the orphan nuclear hormone receptor CAR. J Biol Chem *272*, 23565-23571.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003). Evolution of the protein repertoire. Science *300*, 1701-1703.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. Embo J *5*, 823-826.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5[prime] ends of adenovirus 2 messenger RNA. *12*, 1-8.

Darreh-Shori, T., Hellstrom-Lindahl, E., Flores-Flores, C., Guan, Z. Z., Soreq, H., and Nordberg, A. (2004). Long-lasting acetylcholinesterase splice variations in

anticholinesterase-treated Alzheimer's disease patients
doi:10.1046/j.1471-4159.2003.02230.x. Journal of Neurochemistry *88*, 1102-1113.

Davis, C. A., Grate, L., Spingola, M., and Ares, M., Jr (2000). Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast
10.1093/nar/28.8.1700. Nucl Acids Res *28*, 1700-1706.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, M. O. Dayhoff, ed. (Washington, DC, National Biomedical Research Foundation), pp. 345.

de la Cruz, X., Hutchinson, E. G., Shepherd, A., and Thornton, J. M. (2002). Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci U S A *99*, 11157-11162.

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

Deutsch, M., and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res *27*, 3219-3228.

Dominguez, M., Ferres-Marco, D., Gutierrez-Avino, F. J., Speicher, S. A., and Beneyto, M. (2004). Growth and specification of the eye are controlled independently by Eyegone and Eyeless in Drosophila melanogaster. Nat Genet *36*, 31-39.

Dorn, R., Reuter, G., and Loewendorf, A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila
10.1073/pnas.151268698. PNAS *98*, 9724-9729.

Dralyuk, I., Brudno, M., Gelfand, M. S., Zorn, M., and Dubchak, I. (2000). ASDB: database of alternatively spliced genes. Nucleic Acids Res *28*, 296-297.

Durbin, R., Eddy, S., Krogh, A., and Mitchinson, G. (1998). Biological Sequence Analysis (Cambridge, Cambridge University Press).

Eng, L. (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *23*, 67-76.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., Madhusudhan, M. S., Yerkovich, B., and Sali, A. (2003). Tools for comparative protein structure modeling and analysis. Nucleic Acids Res *31*, 3375-3380.

Etzold, T., Ulyanov, A., and Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. Methods Enzymol *266*, 114-128.

Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A., and Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res *32*, 269-278.

Fehlbaum, P., Guihal, C., Bracco, L., and Cochet, O. (2005). A microarray configuration to quantify expression levels and relative abundance of splice variants 10.1093/nar/gni047. Nucl Acids Res *33*, e47-.

Ferranti, P., Lilla, S., Chianese, L., and Addeo, F. (1999). Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein. J Protein Chem *18*, 595-602.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol *315*, 771-786.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. Proteins *57*, 811-819.

Foulkes, N. S., Mellstrom, B., Benusiglio, E., and Sassone-Corsi, P. (1992). Developmental switch of CREM function during spermatogenesis: from antagonist to activator. Nature *355*, 80-84.

Frederikse, P. H., and Ren, X. O. (2002). Lens defects and age-related fiber cell degeneration in a mouse model of increased AbetaPP gene dosage in Down syndrome. Am J Pathol *161*, 1985-1990.

Frishman, D., and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. Proteins *27*, 329-335.

Furnham, N., Ruffle, S., and Southan, C. (2004). Splice variants: a homology modeling approach. Proteins *54*, 596-608.

Garcia, J., Gerber, S. H., Sugita, S., Sudhof, T. C., and Rizo, J. (2004). A conformational switch in the Piccolo C2A domain regulated by alternative splicing. Nat Struct Mol Biol *11*, 45-53.

Garcia-Blanco, M. A., Baraniak, A. P., and Lasda, E. L. (2004). Alternative splicing in disease and therapy. Nat Biotechnol *22*, 535-546.

Garth, A. D. N., Rollins, D. K., Zhu, J., and Chen, V. C. P. (1996). Evaluation of model discrimination techniques in artificial neural networks with application to grain drying. Intelligent Engineering Systems Through Artificial Neural Networks *6*, 939-950.

Garzon, J., Rodriguez-Diaz, M., Lopez-Fando, A., and Sanchez-Blazquez, P. (2001). RGS9 proteins facilitate acute tolerance to mu-opioid effects. Eur J Neurosci *13*, 801-811.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P. M., Green, M. R., Riva, S., and Biamonti, G. (2005). Cell Motility Is Controlled by SF2/ASF through Alternative Splicing of the Ron Protooncogene. Molecular Cell *20*, 881-890.

Gibson, C. W., Kulkarni, A. B., and Wright, J. T. (2005). The use of animal models to explore amelogenin variants in amelogenesis imperfecta. Cells Tissues Organs *181*, 196-201.

Gilbert, W. (1978). Why genes in pieces? *271*, 501.

Goedert, M., Wischik, C. M., Crowther, R. A., Walker, J. E., and Klug, A. (1988). Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. Proc Natl Acad Sci U S A *85*, 4051-4055.

Goldstrohm, A. C., Greenleaf, A. L., and Garcia-Blanco, M. A. (2001). Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. Gene *277*, 31-47.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. Technometrics *7*, 247-254.

Grabowski, P. J., and Black, D. L. (2001). Alternative RNA splicing in the nervous system. Prog Neurobiol *65*, 289-308.

Grandchamp, B., De Verneuil, H., Beaumont, C., Chretien, S., Walter, O., and Nordmann, Y. (1987). Tissue-specific expression of porphobilinogen deaminase. Two isoenzymes from a single gene. Eur J Biochem *162*, 105-110.

Granneman, J. G., Zhai, Y., Zhu, Z., Bannon, M. J., Burchett, S. A., Schmidt, C. J., Andrade, R., and Cooper, J. (1998). Molecular characterization of human and rat RGS 9L, a novel splice variant enriched in dopamine target regions, and chromosomal localization of the RGS 9 gene. Mol Pharmacol *54*, 687-694.

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197-1211.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *17*, 100-107.

Graveley, B. R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *123*, 65-73.

Green, R., Lewis, B., Hillman, R., Blanchette, M., Lareau, L., Garnett, A., Rio, D., and Brenner, S. (2003). Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics *19*,

I118 - I121.

Gupta, S., Zink, D., Korn, B., Vingron, M., and Haas, S. A. (2004). Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. BMC Genomics *5*, 72.

Haas, B. J., Volfovsky, N., Town, C. D., Troukhan, M., Alexandrov, N., Feldmann, K. A., Flavell, R. B., White, O., and Salzberg, S. L. (2002). Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol *3*, RESEARCH0029.

Hagiwara, M. (2005). Alternative splicing: a new drug target of the post-genome era. Biochim Biophys Acta *1754*, 324-331.

Hanioka, N., Kimura, S., Meyer, U. A., and Gonzalez, F. J. (1990). The human CYP2D locus associated with a common genetic defect in drug oxidation: a G1934----A base change in intron 3 of a mutant CYP2D6 allele results in an aberrant 3' splice recognition site. Am J Hum Genet *47*, 994-1001.

Hastings, M. L., and Krainer, A. R. (2001). Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol *13*, 302-309.

Hastings, M. L., Resta, N., Traum, D., Stella, A., Guanti, G., and Krainer, A. R. (2005). An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *12*, 54-59.

Hastings, M. L., Wilson, C. M., and Munroe, S. H. (2001). A purine-rich intronic element enhances alternative splicing of thyroid hormone receptor mRNA. Rna-a Publication of the Rna Society *7*, 859-874.

He, W., Cowan, C. W., and Wensel, T. G. (1998). RGS9, a GTPase accelerator for phototransduction. Neuron *20*, 95-102.

Heckerling, P. S., Gerber, B. S., Tape, T. G., and Wigton, R. S. (2003). Prediction of Community-Acquired Pneumonia Using Artificial Neural Networks 10.1177/0272989X03251247. Med Decis Making *23*, 112-121.

Hefferon, T. W., Groman, J. D., Yurk, C. E., and Cutting, G. R. (2004). A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing 10.1073/pnas.0400182101. PNAS *101*, 3504-3509.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-10919.

Hicks, M. J., Yang, C. R., Kotlajich, M. V., and Hertel, K. J. (2006). Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns. PLoS Biol *4*, e147.

Hiller, M., Huse, K., Platzer, M., and Backofen, R. (2005). Creation and disruption of protein features by alternative splicing -- a novel mechanism to modulate function. Genome Biol *6*, R58.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004). Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *36*, 1255-1257.

Hirano, M., and Noda, T. (2004). Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. Gene *342*, 165-177.

Hisatomi, H., Nagao, K., Kawakita, M., Matsuda, T., Hirata, H., Yamamoto, S., Nakamoto, T., Harasawa, H., Kaneko, N., Hikiji, K., and Tsukada, Y. (2002). Detection of circulating prostate tumor cells: alternative spliced variant of PSM induced false-positive result. Int J Mol Med *10*, 619-622.

Hong, Y. S., Kim, S. Y., Bhattacharya, A., Pratt, D. R., Hong, W. K., and Tainsky, M. A. (1995). Structure and function of the HOX A1 human homeobox gene cDNA. Gene *159*, 209-214.

Horiuchi, T., and Aigaki, T. (2006). Alternative trans-splicing: a novel mode of pre-mRNA processing. Biol Cell *98*, 135-140.

Horiuchi, T., Giniger, E., and Aigaki, T. (2003). Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, lola. Genes Dev *17*, 2496-2501.

Hovmoller, S., and Zhou, T. (2004). Why are both ends of the polypeptide chain on the outside of proteins? Proteins *55*, 219-222.

Hsu, T., Gogos, J. A., Kirsh, S. A., and Kafatos, F. C. (1992). Multiple zinc finger forms resulting from developmentally regulated alternative splicing of a transcription factor gene. Science *257*, 1946-1950.

Hu, G., Madore, S., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. (2001). Predicting splice variant from DNA chip expression data. Genome Res *11*, 1237 - 1245.

Huang, Y. H., Chen, Y. T., Lai, J. J., Yang, S. T., and Yang, U. C. (2002). PALS db: Putative Alternative Splicing database. Nucleic Acids Res *30*, 186-190.

Huminiecki, L., and Wolfe, K. H. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res *14*, 1870-1879.

Hurst, L. D., Pal, C., and Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet *5*, 299-310.

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A.*, et al.* (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. Nature *393*, 702-705.

Huxford, T., Mishler, D., Phelps, C. B., Huang, D. B., Sengchanthalangsy, L. L., Reeves, R., Hughes, C. A., Komives, E. A., and Ghosh, G. (2002). Solvent exposed non-contacting amino acids play a critical role in NF-kappaB/IkappaBalpha complex formation. J Mol Biol *324*, 587-597.

Hymowitz, S. G., Compaan, D. M., Yan, M., Wallweber, H. J., Dixit, V. M., Starovasnik, M. A., and de Vos, A. M. (2003). The crystal structures of EDA-A1 and EDA-A2: splice variants with distinct receptor specificity. Structure *11*, 1513-1520.

IHGSC (2001). Initial sequencing and analysis of the human genome. *409*, 860-921.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2004). Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. Nucleic Acids Res *32*, 5096-5103.

Inestrosa, N. C., Alvarez, A., Perez, C. A., Moreno, R. D., Vicente, M., Linker, C., Casanueva, O. I., Soto, C., and Garrido, J. (1996). Acetylcholinesterase Accelerates Assembly of Amyloid-[beta]-Peptides into Alzheimer's Fibrils: Possible Role of the Peripheral Site of the Enzyme. Neuron *16*, 881-891.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Ishimura-Oka, K., Nakamuta, M., Chu, M. J., Sullivan, M., Chan, L., and Oka, K. (1995). Partial structure of the mouse glucokinase gene. Genomics *29*, 751-754.

Isoe-Wada, K., Urakami, K., Wakutani, Y., Adachi, Y., Arai, H., Sasaki, H., and Nakashima, K. (1999). Alteration in brain presenilin-1 mRNA expression in sporadic Alzheimer's disease. Eur J Neurol *6*, 163-167.

Jackson, A. P., and Parham, P. (1988). Structure of human clathrin light chains. Conservation of light chain polymorphism in three mammalian species. J Biol Chem *263*, 16688-16695.

Jensen, K., Dredge, B., Stefani, G., Zhong, R., Buckanovich, R., Okano, H., Yang, Y., and Darnell, R. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359 - 371.

Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X., and Li, Y. (2001). AsMamDB: an alternative splice database of mammals. Nucleic Acids Res *29*, 260-263.

Jirajaroenrat, K., Pongjaroenkit, S., Krittanai, C., Prapanthadara, L., and Ketterman, A. J. (2001). Heterologous expression and characterization of alternatively spliced glutathione S-transferases from a single Anopheles gene. Insect Biochem Mol Biol *31*, 867-875.

Johnson, J. M. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *302*, 2141-2144.

Jonsson, A., Winquist, F., Schnurer, J., Sundgren, H., and Lundstrom, I. (1997). Electronic nose for microbial quality classification of grains. International Journal of Food Microbiology
Contributions to Methods in Food Mycology *35*, 187-193.

Jordan, I. K., Marino-Ramirez, L., and Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. Gene *345*, 119-126.

Jurica, M. S., and Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell *12*, 5-14.

Kadener, S., Cramer, P., Nogues, G., Cazalla, D., de la Mata, M., Fededa, J. P., Werbajh, S. E., Srebrow, A., and Kornblihtt, A. R. (2001). Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. Embo J *20*, 5759-5768.

Kafri, R., Bar-Even, A., and Pilpel, Y. (2005). Transcription control reprogramming in genetic backup circuits. Nat Genet *37*, 295-299.

Kalnina, Z., Zayakin, P., Silina, K., and Line, A. (2005). Alterations of pre-mRNA splicing in cancer. Genes Chromosomes Cancer *42*, 342-357.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G.*, et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res *14*, 331-342.

Kan, Z., Rouchka, E. C., Gish, W. R., and States, D. J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res *11*, 889-900.

Kersey, P., Hermjakob, H., and Apweiler, R. (2000). VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. Bioinformatics *16*, 1048-1049.

Kim, G. J., Cheon, Y. H., Park, M. S., Park, H. S., and Kim, H. S. (2001). Generation of protein lineages with new sequence spaces by functional salvage screen. Protein Eng *14*, 647-654.

Kim, H., Klein, R., Majewski, J., and Ott, J. (2004). Estimating rates of alternative splicing in mammals and invertebrates. *36*, 915
916-916
917.

Kim, N., Alekseyenko, A. V., Roy, M., and Lee, C. (2007). The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res *35*, D93-98.

Kirchhausen, T., Scarmato, P., Harrison, S. C., Monroe, J. J., Chow, E. P., Mattaliano, R. J., Ramachandran, K. L., Smart, J. E., Ahn, A. H., and Brosius, J. (1987). Clathrin light chains LCA and LCB are similar, polymorphic, and share repeated heptad motifs. Science *236*, 320-324.

Kirschbaum-Slager, N., Lopes, G. M., Galante, P. A., Riggins, G. J., and de Souza, S. J. (2004). Splicing factors are differentially expressed in tumors. Genet Mol Res *3*, 512-520.

Kirschbaum-Slager, N., Parmigiani, R. B., Camargo, A. A., and de Souza, S. J. (2005). Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data
10.1152/physiolgenomics.00237.2004. Physiol Genomics *21*, 423-432.

Kondrashov, F. A., and Koonin, E. V. (2001). Origin of alternative splicing by tandem exon duplication. *10*, 2661-2669.

Kondrashov, F. A., and Koonin, E. V. (2003). Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *19*, 115-119.

Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000). The Impact of Comparative Genomics on Our Understanding of Evolution. Cell *101*, 573-576.

Kopelman, N. M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *37*, 588-589.

Kornblihtt, A. R. (2005). Promoter usage and alternative splicing. Current Opinion in Cell Biology
Nucleus and gene expression *17*, 262-268.

Kornblihtt, A. R. (2006). Chromatin, transcript elongation and alternative splicing. *13*, 5-7.

Kreahling, J., and Graveley, B. R. (2004). The origins and implications of Aluternative splicing. Trends Genet *20*, 1-4.

Krecic, A. M., and Swanson, M. S. (1999). hnRNP complexes: composition, structure,

and function. Current Opinion in Cell Biology *11*, 363-371.

Krishnan, V. G., and Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics *19*, 2199-2209.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. Trends Genet *19*, 124-128.

Kultz, D. (1998). Phylogenetic and functional classification of mitogen- and stress-activated protein kinases. J Mol Evol *46*, 571-588.

Labrador, M., Mongelard, F., Plata-Rengifo, P., Baxter, E. M., Corces, V. G., and Gerasimova, T. I. (2001). Protein encoding by both DNA strands. *409*, 1000.

Laity, J. H., Chung, J., Dyson, H. J., and Wright, P. E. (2000a). Alternative splicing of Wilms' tumor suppressor protein modulates DNA binding activity through isoform-specific DNA-induced conformational changes. Biochemistry *39*, 5341-5348.

Laity, J. H., Dyson, H. J., and Wright, P. E. (2000b). Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein. Proc Natl Acad Sci U S A *97*, 11932-11935.

Laoide, B. M., Foulkes, N. S., Schlotter, F., and Sassone-Corsi, P. (1993). The functional versatility of CREM is determined by its modular structure. Embo J *12*, 1179-1191.

LaRosa, G. J., and Gudas, L. J. (1988). Early retinoic acid-induced F9 teratocarcinoma stem cell gene ERA-1: alternate splicing creates transcripts for a homeobox-containing protein and one lacking the homeobox. Mol Cell Biol *8*, 3906-3917.

Latchman, D. S. (1996a). Activation and repression of gene expression by POU family transcription factors. Philos Trans R Soc Lond B Biol Sci *351*, 511-515.

Latchman, D. S. (1996b). Inhibitory transcription factors. Int J Biochem Cell Biol *28*, 965-974.

Latchman, D. S. (1996c). The Oct-2 transcription factor. Int J Biochem Cell Biol *28*, 1081-1083.

Laurencikiene, J., Kallman, A. M., Fong, N., Bentley, D. L., and Ohman, M. (2006). RNA editing and alternative splicing: the importance of co-transcriptional coordination. EMBO Rep *7*, 303-307.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S. F., and Lee, C. (2004).

Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data
10.1093/nar/gnh173. Nucl Acids Res *32*, e180-.

Lee, C., Atanelov, L., Modrek, B., and Xing, Y. (2003). ASAP: the Alternative Splicing Annotation Project. Nucleic Acids Res *31*, 101-105.

Lee, C., and Roy, M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. Genome Biology *5*, 231.

Lee, C., and Wang, Q. (2005). Bioinformatics analysis of alternative splicing. Brief Bioinform *6*, 23-33.

Letunic, I., Copley, R. R., and Bork, P. (2002). Common exon duplication in animals and its role in alternative splicing. *11*, 1561-1567.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. Nucleic Acids Res *32*, D142-144.

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3[prime] splice-site selection in Alu exons. *300*, 1288-1291.

Lewis, B. P., Green, R. E., and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *100*, 189-192.

Li, B. L., Li, X. L., Duan, Z. J., Lee, O., Lin, S., Ma, Z. M., Chang, C. C., Yang, X. Y., Park, J. P., Mohandas, T. K.*, et al.* (1999). Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. J Biol Chem *274*, 11060-11071.

Li, L., and Howe, G. A. (2001). Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway. Plant Mol Biol *46*, 409-419.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics *17*, 282-283.

Lian, Y., and Garner, H. R. (2005). Evidence for the regulation of alternative splicing via complementary DNA sequence repeats
10.1093/bioinformatics/bti180. Bioinformatics *21*, 1358-1364.

Liao, B. Y., and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol *23*, 530-540.

Lister, J. A., Close, J., and Raible, D. W. (2001). Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential. Dev Biol *237*, 333-344.

Liu, S., and Altman, R. B. (2003). Large scale study of protein domain distribution in the context of alternative splicing. Nucleic Acids Res *31*, 4828-4835.

Lopez, A. J. (1995). Developmental role of transcription factor isoforms generated by alternative splicing. Dev Biol *172*, 396-411.

Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *32*, 279-305.

Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigo, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *579*, 1900-1903.

Lorson, C. L., Hahnen, E., Androphy, E. J., and Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci U S A *96*, 6307-6311.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155.

Lynch, M., and Conery, J. S. (2003). The origins of genome complexity. Science *302*, 1401-1404.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C., and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res *32*, D235-239.

Magen, A., and Ast, G. (2005). The importance of being divisible by three in alternative splicing. Nucleic Acids Res *33*, 5574-5582.

Magnuson, M. A., and Shelton, K. D. (1989). An alternate promoter in the glucokinase gene is active in the pancreatic beta cell. J Biol Chem *264*, 15936-15942.

Makova, K. D., and Li, W. H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res *13*, 1638-1645.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499-506.

Maniatis, T., and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature *418*, 236-243.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y.,

Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z.*, et al.* (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res *33*, D192-196.

Margolin, J. F., Friedman, J. R., Meyer, W. K., Vissing, H., Thiesen, H. J., and Rauscher, F. J., 3rd (1994). Kruppel-associated boxes are potent transcriptional repression domains. Proc Natl Acad Sci U S A *91*, 4509-4513.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct *29*, 291-325.

Masel, J. (2006). Cryptic genetic variation is enriched for potential adaptations. *172*, 1985-1991.

Matlin, A. J., Clark, F., and Smith, C. W. J. (2005). Understanding alternative splicing: Towards a cellular code. Nature Reviews Molecular Cell Biology *6*, 386-398.

Matsuda, M., Tanaka, S., Nagata, S., Kojima, A., Kurata, T., and Shibuya, M. (1992). Two species of human CRK cDNA encode proteins with distinct biological activities. Mol Cell Biol *12*, 3482-3489.

Mereau, A., Le Sommer, C., Lerivray, H., Lesimple, M., and Hardy, S. (2007). Xenopus as a model to study alternative splicing in vivo. Biol Cell *99*, 55-65.

Meshorer, E., and Soreq, H. (2002). Pre-mRNA splicing modulations in senescence. Aging Cell *1*, 10-16.

Meyer, I. M., and Miklos, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs 10.1093/nar/gki923. Nucl Acids Res *33*, 6338-6348.

Mills, A. A. (2005). p53: link to the past, bridge to the future. Genes Dev *19*, 2091-2099.

Mironov, A., Fickett, J., and Gelfand, M. (1999). Frequent alternative splicing of human genes. Genome Res *9*, 1288 - 1293.

Mockler, T. C., and Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics *85*, 1-15.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13 - 19.

Modrek, B., and Lee, C. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss. *34*, 177-180.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes 10.1093/nar/29.13.2850. Nucl Acids Res *29*, 2850-2859.

Mount, S., and Steitz, J. (1983). Lessons from mutant globins. Nature *303*, 380-381.

Mucchielli-Giorgi, M. H., Hazout, S., and Tuffery, P. (1999). PredAcc: prediction of solvent accessibility. Bioinformatics *15*, 176-177.

Munch, C., Ebstein, M., Seefried, U., Zhu, B., Stamm, S., Landwehrmeyer, G. B., Ludolph, A. C., Schwalenstocker, B., and Meyer, T. (2002). Alternative splicing of the 5'-sequences of the mouse EAAT2 glutamate transporter and expression in a transgenic model for amyotrophic lateral sclerosis. J Neurochem *82*, 594-603.

Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., and Gotoh, O. (2006). Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics *22*, 1211-1216.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Neu-Yilik, G., Gehring, N. H., Hentze, M. W., and Kulozik, A. E. (2004). Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. Genome Biol *5*, 218.

Neverov, A. D., Artamonova, II, Nurtdinov, R. N., Frishman, D., Gelfand, M. S., and Mironov, A. A. (2005). Alternative splicing and protein function. BMC Bioinformatics *6*, 266.

Newman, M., Musgrave, F. I., and Lardelli, M. (2006). Alzheimer disease: Amyloidogenesis, the presenilins and animal models. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *In Press, Corrected Proof*.

Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? Bioessays *25*, 1147-1149.

Nogues, G., Kadener, S., Cramer, P., Bentley, D., and Kornblihtt, A. R. (2002). Transcriptional activators differ in their abilities to control alternative splicing. J Biol Chem *277*, 43110-43114.

Nurtdinov, R. N., Artamonova, I. I., Mironov, A. A., and Gelfand, M. S. (2003). Low conservation of alternative splicing patterns in the human and mouse genomes. *12*, 1313-1320.

Oakley, A. J., Harnnoi, T., Udomsinprasert, R., Jirajaroenrat, K., Ketterman, A. J., and Wilce, M. C. (2001). The crystal structures of glutathione S-transferases isozymes 1-3 and 1-4 from Anopheles dirus species B. Protein Sci *10*, 2176-2185.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res *33*, D476-480.

O'Connor, L., Strasser, A., O'Reilly, L. A., Hausmann, G., Adams, J. M., Cory, S., and Huang, D. C. (1998). Bim: a novel member of the Bcl-2 family that promotes apoptosis. Embo J *17*, 384-395.

Offman, M. N., Nurtdinov, R. N., Gelfand, M. S., and Frishman, D. (2004). No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. BMC Bioinformatics *5*, 41.

Ogawa, S., Toyoshima, H., Kozutsumi, H., Hagiwara, K., Sakai, R., Tanaka, T., Hirano, N., Mano, H., Yazaki, Y., and Hirai, H. (1994). The C-terminal SH3 domain of the mouse c-Crk protein negatively regulates tyrosine-phosphorylation of Crk associated p130 in rat 3Y1 cells. Oncogene *9*, 1669-1678.

Okumura, M., Kondo, S., Ogata, M., Kanemoto, S., Murakami, T., Yanagida, K., Saito, A., and Imaizumi, K. (2005). Candidates for tumor-specific alternative splicing. Biochemical and Biophysical Research Communications *334*, 23-29.

Ornitz, D. M., Xu, J., Colvin, J. S., McEwen, D. G., MacArthur, C. A., Coulier, F., Gao, G., and Goldfarb, M. (1996). Receptor specificity of the fibroblast growth factor family. J Biol Chem *271*, 15292-15297.

Pacheco, T. R., Gomes, A. Q., Barbosa-Morais, N. L., Benes, V., Ansorge, W., Wollerton, M., Smith, C. W., Valcarcel, J., and Carmo-Fonseca, M. (2004). Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS. J Biol Chem *279*, 27039-27049.

Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., and Blencowe, B. J. (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet *21*, 73-77.

Pan, Q., Saltzman, A. L., Kim, Y. K., Misquitta, C., Shai, O., Maquat, L. E., Frey, B. J., and Blencowe, B. J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev *20*, 153-158.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., and Morris, Q. D. (2004). Revealing Global Regulatory Features of Mammalian Alternative Splicing Using a Quantitative Microarray Platform. Molecular Cell *16*, 929-941.

Papp, B., Pal, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194-197.

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein

structures. J Mol Biol *224*, 461-471.

Peneff, C., Ferrari, P., Charrier, V., Taburet, Y., Monnier, C., Zamboni, V., Winter, J., Harnois, M., Fassy, F., and Bourne, Y. (2001). Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture. Embo J *20*, 6191-6202.

Pieples, K., and Wieczorek, D. F. (2000). Tropomyosin 3 increases striated muscle isoform diversity. Biochemistry *39*, 8291-8297.

Pospisil, H., Herrmann, A., Bortfeldt, R. H., and Reich, J. G. (2004). EASED: Extended Alternatively Spliced EST Database. Nucleic Acids Res *32*, D70-74.

Rahman, Z., Gold, S. J., Potenza, M. N., Cowan, C. W., Ni, Y. G., He, W., Wensel, T. G., and Nestler, E. J. (1999). Cloning and characterization of RGS9-2: a striatal-enriched alternatively spliced product of the RGS9 gene. J Neurosci *19*, 2016-2026.

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M.*, et al.* (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome 10.1101/gr.4200206. Genome Res *16*, 11-19.

Reinach, F. C., and MacLeod, A. R. (1986). Tissue-specific expression of the human tropomyosin gene involved in the generation of the trk oncogene. Nature *322*, 648-650.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004). Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res *3*, 76-83.

Roberts, G. C., and Smith, C. W. (2002). Alternative splicing: combinatorial output from the genome. Curr Opin Chem Biol *6*, 375-383.

Robertson, J., Doroudchi, M. M., Nguyen, M. D., Durham, H. D., Strong, M. J., Shaw, G., Julien, J. P., and Mushynski, W. E. (2003). A neurotoxic peripherin splice variant in a mouse model of ALS. J Cell Biol *160*, 939-949.

Robichaud, G. A., Nardini, M., Laflamme, M., Cuperlovic-Culf, M., and Ouellette, R. J. (2004). Human Pax-5 C-terminal Isoforms Possess Distinct Transactivation Properties and Are Differentially Modulated in Normal and Malignant B Cells 10.1074/jbc.M407171200. J Biol Chem *279*, 49956-49963.

Rogina, B., and Upholt, W. B. (1995). The chicken homeobox gene Hoxd-11 encodes two alternatively spliced RNA species. Biochem Mol Biol Int *35*, 825-831.

Rohrbach, S., Muller-Werdan, U., Werdan, K., Koch, S., Gellerich, N. F., and Holtz, J. (2005). Apoptosis-modulating interaction of the neuregulin/erbB pathway with

antracyclines in regulating Bcl-xS and Bcl-xL in cardiomyocytes. Journal of Molecular and Cellular Cardiology *38*, 485-493.

Roy, M., Xu, Q., and Lee, C. (2005). Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms
10.1093/nar/gki792. Nucl Acids Res *33*, 5026-5033.

Ruben, S. M., Narayanan, R., Klement, J. F., Chen, C. H., and Rosen, C. A. (1992). Functional characterization of the NF-kappa B p65 transcriptional activator and an alternatively spliced derivative. Mol Cell Biol *12*, 444-454.

Rumelhart, D., Durbin, R., Golden, R., and Chauvin, Y. (1995). Backpropagation: Theory, architectures and applications. Backpropagation: The Basic Theory, 1-34.

Russell, R. B., and Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol *244*, 332-350.

Sakharkar, M. K., and Kangueane, P. (2004). Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. BMC Bioinformatics *5*, 67.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Sato, N., Hori, O., Yamaguchi, A., Lambert, J. C., Chartier-Harlin, M. C., Robinson, P. A., Delacourte, A., Schmidt, A. M., Furuyama, T., Imaizumi, K.*, et al.* (1999). A novel presenilin-2 splice variant in human Alzheimer's disease brain tissue. J Neurochem *72*, 2498-2505.

Scorilas, A., Levesque, M. A., Ashworth, L. K., and Diamandis, E. P. (2002). Cloning, physical mapping and structural characterization of the human alpha(A)-adaptin gene. Gene *289*, 191-199.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky, S. L. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. Cell *101*, 671-684.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A *95*, 5857-5864.

Schultz, J., Ponting, C. P., Hofmann, K., and Bork, P. (1997). SAM as a protein interaction domain involved in developmental regulation. Protein Sci *6*, 249-253.

Schweizer, A., Valdenaire, O., Nelbock, P., Deuschle, U., Dumas Milne Edwards, J. B., Stumpf, J. G., and Loffler, B. M. (1997). Human endothelin-converting enzyme (ECE-

1): three isoforms with distinct subcellular localizations. Biochem J *328 (Pt 3)*, 871-877.

Semon, M., and Duret, L. (2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol *23*, 1715-1723.

Shai, O., Morris, Q. D., Blencowe, B. J., and Frey, B. J. (2006). Inferring global levels of alternative splicing isoforms using a generative model of microarray data. Bioinformatics *22*, 606-613.

Shakhnovich, B. E., and Koonin, E. V. (2006). Origins and impact of constraints in evolution of gene families. Genome Res *16*, 1529-1536.

Shan, L., Vincent, J., Brunzelle, J. S., Dussault, I., Lin, M., Ianculescu, I., Sherman, M. A., Forman, B. M., and Fernandez, E. J. (2004). Structure of the murine constitutive androstane receptor complexed to androstenol: a molecular basis for inverse agonism. Mol Cell *16*, 907-917.

Sharp, P. A. (1994). Split genes and RNA splicing. Cell *77*, 805-815.

Shen, H., and Green, M. R. (2004). A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. Mol Cell *16*, 363-373.

Shen, H., Kan, J. L., and Green, M. R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Mol Cell *13*, 367-376.

Shepherd, A. J., Gorse, D., and Thornton, J. M. (1999). Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci *8*, 1045-1055.

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., Loerch, P. M., Leonardson, A., Lum, P. Y., Cavet, G*., et al.* (2001). Experimental annotation of the human genome using microarray technology. *409*, 922-927.

Shortle, D., and Sondek, J. (1995). The emerging role of insertions and deletions in protein engineering. Curr Opin Biotechnol *6*, 387-393.

Simard, M. J., and Chabot, B. (2002). SRp30c Is a Repressor of 3' Splice Site Utilization. Mol Cell Biol *22*, 4001-4010.

Sirand-Pugnet, P., Durosay, P., d'Orval, B. C., Brody, E., and Marie, J. (1995). [beta]-Tropomyosin Pre-mRNA Folding Around a Muscle-specific Exon Interferes with Several Steps of Spliceosome Assembly. Journal of Molecular Biology *251*, 591-602.

Smith, C. W. J., and Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *25*, 381-388.

Sondek, J., and Shortle, D. (1990). Accommodation of single amino acid insertions by the native state of staphylococcal nuclease. Proteins *7*, 299-305.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins *28*, 405-420.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. *12*, 1060-1067.

Spingola, M., Grate, L., Haussler, D., and Ares, M., Jr. (1999). Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. Rna *5*, 221-234.

Srebrow, A., and Kornblihtt, A. R. (2006). The connection between splicing and cancer. J Cell Sci *119*, 2635-2641.

Srinivasan, K., Shiue, L., Hayes, J. D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L. R., Bryant, J., Smith, M., and Rommelfanger, C. (2005). Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods Post-transcriptional Regulation of Gene Expression *37*, 345-359.

Stamm, S. (2002). Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. Hum Mol Genet *11*, 2409-2416.

Stamm, S., Riethoven, J. J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N. L., and Thanaraj, T. A. (2006). ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res *34*, D46-55.

Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., and Zhang, M. (2000). An alternative-exon database and its statistical analysis. DNA Cell Biol *19*, 739 - 756.

Stetefeld, J., Alexandrescu, A. T., Maciejewski, M. W., Jenny, M., Rathgeb-Szabo, K., Schulthess, T., Landwehr, R., Frank, S., Ruegg, M. A., and Kammerer, R. A. (2004). Modulation of agrin function by alternative splicing and Ca2+ binding. Structure *12*, 503-515.

Stetefeld, J., and Ruegg, M. A. (2005). Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci *30*, 515-521.

Stevens, S. W., Ryan, D. E., Ge, H. Y., Moore, R. E., Young, M. K., Lee, T. D., and Abelson, J. (2002). Composition and functional characterization of the yeast spliceosomal penta-snRNP. Mol Cell *9*, 31-44.

Stoilov, P., Meshorer, E., Gencheva, M., Glick, D., Soreq, H., and Stamm, S. (2002). Defects in Pre-mRNA Processing as Causes of and Predisposition to Diseases doi:10.1089/104454902320908450. DNA and Cell Biology *21*, 803-818.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P. E.*, et al.* (2004). A Gene Expression Map for the Euchromatic Genome of Drosophila melanogaster 10.1126/science.1101312. Science *306*, 655-660.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G.*, et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Su, Z., Wang, J., Yu, J., Huang, X., and Gu, X. (2006). Evolution of alternative splicing after gene duplication. *16*, 182-189.

Sugnet, C. W. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *2*, e4.

Sugnet, C. W., Kent, W. J., Ares, M., and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. 66-77.

Sun, H., and Chasin, L. A. (2000). Multiple splicing defects in an intronic false exon. Mol Cell Biol *20*, 6414-6425.

Takenaga, K., Nakamura, Y., Kageyama, H., and Sakiyama, S. (1990). Nucleotide sequence of cDNA for nonmuscle tropomyosin 5 of mouse fibroblast. Biochim Biophys Acta *1087*, 101-103.

Talavera, D., Vogel, C., Orozco, M., Teichmann, S. A., and de la Cruz, X. (2007). The (In)dependence of Alternative Splicing and Gene Duplication. PLoS Comput Biol *3*, e33.

Taneri, B., Snyder, B., Novoradovsky, A., and Gaasterland, T. (2004). Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. Genome Biol *5*, R75.

Tanizawa, Y., Matsutani, A., Chiu, K. C., and Permutt, M. A. (1992). Human glucokinase gene: isolation, structural characterization, and identification of a microsatellite repeat polymorphism. Mol Endocrinol *6*, 1070-1081.

Taylor, J. K., Zhang, Q. Q., Wyatt, J. R., and Dean, N. M. (1999). Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides. Nature Biotechnology *17*, 1097-1100.

Thanaraj, T. A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids

Res *29*, 2581-2593.

Thanaraj, T. A., Clark, F., and Muilu, J. (2003). Conservation of human alternative splice events in mouse. *31*, 2544-2552.

Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J. J., Le Texier, V., and Muilu, J. (2004). ASD: the Alternative Splicing Database. Nucleic Acids Res *32*, D64-69.

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol *333*, 863-882.

Topham, C. M., Srinivasan, N., and Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng *10*, 7-21.

Ule, J., Jensen, K., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212 - 1215.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M*., et al.* (2005). Nova regulates brain-specific splicing to shape the synapse. *37*, 844-852.

Valdenaire, O., Lepailleur-Enouf, D., Egidy, G., Thouard, A., Barret, A., Vranckx, R., Tougard, C., and Michel, J. B. (1999). A fourth isoform of endothelin-converting enzyme (ECE-1) is generated from an additional promoter molecular cloning and characterization. Eur J Biochem *264*, 341-349.

Valdenaire, O., Rohrbacher, E., and Mattei, M. G. (1995). Organization of the gene encoding the human endothelin-converting enzyme (ECE-1). J Biol Chem *270*, 29794-29798.

Valentine, J. W. (2000). Two genomic paths to the evolution of complexity in bodyplans. Paleobiology, 513-519.

Valenzuela, A., Talavera, D., Orozco, M., and de la Cruz, X. (2004). Alternative splicing mechanisms for the modulation of protein function: conservation between human and other species. J Mol Biol *335*, 495-502.

van Driel, R., Fransz, P. F., and Verschure, P. J. (2003). The eukaryotic genome: a system regulated at different hierarchical levels. J Cell Sci *116*, 4067-4075.

Venables, J. P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Venables, J. P. (2006). Unbalanced alternative splicing and its significance in cancer. BioEssays *28*, 378-386.

Venter, J. C. (2001). The sequence of the human genome. *291*, 1304-1351.

Wada, K., Yokotani, N., Hunter, C., Doi, K., Wenthold, R., and Shimasaki, S. (1992). Differential Expression of Two Distinct Forms of mRNA Encoding Members of a Dipeptidyl Aminopeptidase Family
10.1073/pnas.89.1.197. PNAS *89*, 197-201.

Wang, H., Hubbell, E., Hu, J.-s., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M. A., Ares, M., Kulp, D. C., and Haussler, D. (2003a). Gene structure-based splice variant deconvolution using a microarry platform
10.1093/bioinformatics/btg1044. Bioinformatics *19*, i315-322.

Wang, P., Yan, B., Guo, J. T., Hicks, C., and Xu, Y. (2005). Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc Natl Acad Sci U S A *102*, 18920-18925.

Wang, W. (2005). Origin and evolution of new exons in rodents. *15*, 1258-1264.

Wang, Z., Lo, H., Yang, H., Gere, S., Hu, Y., Buetow, K., and Lee, M. (2003b). Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res *63*, 655 - 657.

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., and Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Watahiki, A., Waki, K., Hayatsu, N., Shiraki, T., Kondo, S., Nakamura, M., Sasaki, D., Arakawa, T., Kawai, J., Harbers, M.*, et al.* (2004). Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *1*, 233-239.

Wen, F., Li, F., Xia, H., Lu, X., Zhang, X., and Li, Y. (2004). The impact of very short alternative splicing on protein structures and functions in the human genome. Trends Genet *20*, 232-236.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res *31*, 28-33.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol *297*, 233-249.

Witzgall, R., O'Leary, E., Leaf, A., Onaldi, D., and Bonventre, J. V. (1994). The Kruppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. Proc Natl Acad Sci U S A *91*, 4514-4518.

Wong, G. K., Passey, D. A., and Yu, J. (2001). Most of the human genome is

transcribed. Genome Res *11*, 1975-1977.

Wootton, J. C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol *266*, 554-571.

Xie, H., Zhu, W. Y., Wasserman, A., Grebinskiy, V., Olson, A., and Mintz, L. (2002). Computational analysis of alternative splicing using EST tissue information. Genomics *80*, 326-330.

Xie, X., Gu, Y., Fox, T., Coll, J. T., Fleming, M. A., Markland, W., Caron, P. R., Wilson, K. P., and Su, M. S. (1998). Crystal structure of JNK3: a kinase implicated in neuronal apoptosis. Structure *6*, 983-991.

Xing, Y., and Lee, C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *102*, 13526-13531.

Xing, Y., and Lee, C. (2006). Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. Nat Rev Genet *7*, 499-509.

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res *30*, 3754 - 3766.

Yanai, I., Graur, D., and Ophir, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. Omics *8*, 15-24.

Yeakley, J. M., Fan, J. B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., and Fu, X. D. (2002). Profiling alternative splicing on fiber-optic arrays. Nat Biotechnol *20*, 353-358.

Yeh, B. K., Igarashi, M., Eliseenkova, A. V., Plotnikov, A. N., Sher, I., Ron, D., Aaronson, S. A., and Mohammadi, M. (2003). Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. Proc Natl Acad Sci U S A *100*, 2266-2271.

Yeo, G., Holste, D., Kreiman, G., and Burge, C. (2004). Variation in alternative splicing across human tissues. Genome Biology *5*, R74.

Zarich, N., Oliva, J. L., Jorge, R., Santos, E., and Rojas, J. M. (2000). The isoform-specific stretch of hSos1 defines a new Grb2-binding domain. Oncogene *19*, 5872-5883.

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D., Hayashizaki, Y., Gaasterland, T., Group, R. G., and members, G. (2003). Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res *13*, 1290 - 1300.

Zhang, K., Howes, K. A., He, W., Bronson, J. D., Pettenati, M. J., Chen, C., Palczewski, K., Wensel, T. G., and Baehr, W. (1999). Structure, alternative splicing, and expression of the human RGS9 gene. Gene *240*, 23-34.

Zhang, L., Yang, N., Mohamed-Hadley, A., Rubin, S. C., and Coukos, G. (2003). Vector-based RNAi, a novel tool for isoform-specific knock-down of VEGF and anti-angiogenesis gene therapy of cancer. Biochemical and Biophysical Research Communications *303*, 1169-1178.

Zhang, X. H., and Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *18*, 1241-1250.

Zheng, C. L., Kwon, Y. S., Li, H. R., Zhang, K., Coutinho-Mansfield, G., Yang, C., Nair, T. M., Gribskov, M., and Fu, X. D. (2005). MAASE: an alternative splicing database designed for supporting splicing microarray applications. Rna *11*, 1767-1776.

Zhou, Y., Zhou, C., Ye, L., Dong, J., Xu, H., Cai, L., Zhang, L., and Wei, L. (2003a). Database and analyses of known alternatively spliced genes in plants. Genomics *82*, 584-595.

Zhou, Y. Q., He, C., Chen, Y. Q., Wang, D., and Wang, M. H. (2003b). Altered expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: generation of different splicing RON variants and their oncogenic potential. Oncogene *22*, 186-197.