Faculdade de Medicina
Universidade de Lisboa

Bioinformatics Studies on the Complexity of Splicing and Gene Expression

Nuno Luís Barbosa Morais
Doutoramento em Ciências Biomédicas,
especialidade de Ciências Funcionais

Tese orientada por:
Doutor Samuel Aparício (University of Cambridge, UK)
Professora Doutora Maria do Carmo Fonseca

2006

Aos meus Pais

A impressão desta dissertação foi aprovada pela Comissão Coordenadora do Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 20 de Julho de 2006.

A impressão deste documento foi financiada pela Fundação para a Ciência e a Tecnologia (Fundo Social Europeu - 3º Programa Quadro) - Bolsa PRAXIS XXI SFRH/BD/2914/2000.

As opiniões expressas nesta publicação são da exclusiva responsabilidade do seu autor.

LaTeX 2_ε– August 17, 2006

Contents

Prefácio

Nesta dissertação são apresentados os resultados do trabalho de investigação desenvolvido entre os anos de 2001 e 2006 na Faculdade de Medicina da Universidade de Lisboa, sob orientação da Professora Doutora Maria do Carmo Fonseca, no Departamento de Oncologia da Universidade de Cambridge, Reino Unido (Janeiro de 2003 a Julho de 2005), sob orientação do Doutor Samuel Aparício, e no Laboratório Europeu de Biologia Molecular (EMBL), em Heidelberg - Alemanha (Fevereiro a Maio de 2002), sob orientação do Doutor Juan Valcárcel.

Este trabalho teve como objectivo central a identificação e a caracterização de mecanismos de complexidade da expressão génica, através de abordagens bioinformáticas. A análise incidiu particularmente no *splicing* do pre-mRNA, tanto ao nível dos seus elementos reguladores em *trans* (os chamados factores de *splicing*) como ao nível dos reguladores em *cis* (nomeadamente os sítios de ligação dos factores ao RNA).

Procurou-se perceber se o *splicing* em vertebrados beneficiou de novos mecanismos ou apenas do refinamento dos ancestrais. Tentou-se também avaliar e distinguir as diferenças na evolução dos diversos componentes da maquinaria de *splicing*.

Este trabalho visou também a distinção e identificação de elementos reguladores de *splicing* ao nível da sequência do RNA. A análise bioinformática incidiu particularmente no reconhecimento dos diversos motivos de ligação dos factores de *splicing*. Procurou-se o estabelecimento das repercussões funcionais (ao nível do *splicing* alternativo e de processos celulares tão importantes como o metabolismo do mRNA ou a apoptose) de variações na abundância e na sequência daqueles sinais.

O trabalho envolveu ainda a participação em projectos de *microarrays*, ferramenta poderosa em estudos de compexidade e na resolução das questões descritas, uma vez

que permite a avaliação de padrões de expressão génica à escala genómica.

A dissertação está dividida em seis capítulos.

O primeiro é introdutório e começa por abordar o processo de *splicing* na génese do RNA mensageiro (secção 1.1). Descreve-se a maquinaria e os mecanismos moleculares do *splicing*, assim como a regulação de *splicing* alternativo. Segue-se uma secção (1.2) descritiva da dinâmica dos mecanismos de evolução dos genomas em vertebrados. São também abordadas as ferramentas bioinformáticas associadas ao estudo das questões biológicas descritas (secção 1.3). Faz-se ainda uma descrição sumária da tecnologia dos *microarrays* e da sua importância como instrumento poderoso na análise da expressão génica (secção 1.4). Conclui-se o primeiro capítulo com a discussão dos objectivos fundamentais do trabalho (secção 1.5).

O segundo capítulo apresenta os resultados originais obtidos no estudo da evolução dos factores de *splicing* em eucariotas, sob a forma de artigo publicado [Barbosa-Morais et al., 2006]. Parte destes resultados foram utilizados num trabalho de colaboração publicado, sobre a diversidade do factor $U2AF^{35}$ [Pacheco et al., 2004].

No terceiro capítulo, apresentado como artigo de revisão [Mollet et al., 2006], analisa-se a evolução das características estruturais das famílias de proteínas relacionadas com o factor U2AF e são discutidas as implicações da sua diversidade na regulação do *splicing*.

O quarto capítulo é dedicado ao estudo dos elementos reguladores de *splicing* em *cis*. Descreve-se o desenvolvimento de um programa informático destinado a prever, em sequências de RNA, sítios de ligação de factores de *splicing*, nomeadamente proteínas SR e hnRNPs (secção 4.1). O programa foi incluido como ferramenta na "bancada virtual" do projecto Alternative Splicing Database, publicada em [Stamm et al., 2006]. Resume-se também o trabalho de busca de motivos de ligação dos factores PTB e $U2AF^{65}$ a mRNAs seleccionados experimentalmente (combinando imunoprecipitação com *microarrays*) como interagindo com aqueles factores (secção 4.2). Este trabalho foi realizado em colaboração com Margarida Gama Carvalho e originou um artigo submetido [Gama-Carvalho et al., 2006]. No quarto capítulo são ainda sumarizados os resultados de trabalho desenvolvido no EMBL em estudos de regulação da apoptose por *splicing* alternativo (secção 4.3) e de características dos sítios

de *splicing* intrónicos (secções 4.4 e 4.5).

O quinto capítulo resume a aplicação das ferramentas bioinformáticas de anotação de sequências, como colaboração em projectos de *microarrays*, dos quais resultaram várias publicações [Naderi et al., 2004; Teschendorff et al., 2006b; Naderi et al., 2006; Teschendorff et al., 2005; Teschendorff et al., 2006a; Chin et al., 2006; Ibrahim et al., 2006].

No sexto e último capítulo faz-se uma discussão integrada de todo o trabalho e respectivos resultados. São apresentadas as suas conclusões finais e perspectivas futuras.

Como previsto no ponto 4 do Artigo 15° do Regulamento de Doutoramentos da Universidade de Lisboa, a presente dissertação foi redigida em língua inglesa e contém um resumo alargado em língua portuguesa (Sumário). As justificações para esta escolha são de ordem diversa. Por um lado, grande parte do trabalho de investigação foi desenvolvido em laboratórios estrangeiros cuja língua oficial é a inglesa. Por outro lado, é provável a participação de cientistas estrangeiros no júri das Provas de Doutoramento. A língua inglesa garante ainda maior facilidade de difusão do documento pela comunidade científica internacional e foi a usada nos artigos científicos resultantes do trabalho descrito nesta dissertação.

Como anteriormente mencionado, os resultados do trabalho individual e de colaboração descrito nesta dissertação são apresentados em artigos científicos (publicados [•], aceites para publicação [•] e submetidos para publicação [○]), para os quais a contribuição individual do Doutorando foi de índole diversa:

• **Barbosa-Morais NL**, Carmo-Fonseca M, Aparicio S. "Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion." *Genome Res.*, 2006 Jan;16(1):66-77 [Barbosa-Morais et al., 2006]
Contribuição individual: concepção e execução dos métodos e toda a análise de resultados; escrita integral (com supervisão, contribuição e correcção por parte dos orientadores/co-autores) do artigo.

• Teschendorff AE, Naderi A, **Barbosa-Morais NL**, Caldas C. "PACK: Profile

Analysis using Clustering and Kurtosis to find molecular classifiers in cancer". *Bioinformatics*, 2006 May 8 [Teschendorff et al., 2006a]
Contribuição individual: recolha e anotação bioinformática cruzada dos resultados de diferentes estudos de microarrays *usados no teste do método descrito.*

• Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, **Barbosa-Morais NL**, Thanaraj TA. "ASD: a bioinformatics resource on alternative splicing". *Nucleic Acids Res.*, 2006 Jan 1;34(Database issue):D46-55 [Stamm et al., 2006]
Contribuição individual: autoria do programa Splicing Rainbow, incluido na "bancada virtual" descrita no artigo.

• Teschendorff AE, Wang Y, **Barbosa-Morais NL**, Brenton JD, Caldas C. "A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data". *Bioinformatics*, 2005 Jul 1;21(13):3025-33 [Teschendorff et al., 2005]
Contribuição individual: recolha e anotação bioinformática cruzada dos resultados de diferentes estudos de microarrays *usados no teste do método descrito.*

• Pacheco TR, Gomes AQ, **Barbosa-Morais NL**, Benes V, Ansorge W, Wollerton M, Smith CW, Valcarcel J, Carmo-Fonseca M. "Diversity of vertebrate splicing factor U2AF^{35}: identification of alternatively spliced U2AF1 mRNAS". *J Biol Chem*, 2004 Jun 25;279(26):27039-49 [Pacheco et al., 2004]
Contribuição individual: análise bioinformática de sequências.

• Naderi A, Ahmed AA, **Barbosa-Morais NL**, Aparicio S, Brenton JD, Caldas C. "Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling". *BMC Genomics*, 2004 Jan 30;5(1):9 [Naderi et al., 2004]
Contribuição individual: bioinformática (análise e anotação de sequências).

• Chin SF, Wang Y, Thorne NP, Teschendorff AE, Pinder SE, Vias M, **Barbosa-Morais NL**, Roberts I, Naderi A, Garcia M, Iyer NJ, Kranjac T, Robertson J, Ruffalo T, Aparicio S, Tavare S, Ellis I, Brenton J, Caldas C. "Using array-CGH to define

molecular potraits of primary breast cancer". *Oncogene*, in press 2006 [Chin et al., 2006]
Contribuição individual: anotação bioinformática cruzada dos clones representados nos microarrays *usados.*

• Ibrahim AEK, Thorne NP, Baird K, **Barbosa-Morais NL**, Tavare S, Collins VP, Wyllie AH, Arends MJ, Brenton JD. "MMASS: an optimised array-based method for assessing CpG island methylation". *Nucleic Acids Research*, in press 2006 [Ibrahim et al., 2006]
Contribuição individual: derivação de sequências e anotação bioinformática dos clones representados nos microarrays *usados; previsão da distribuição de sítios de restrição; optimização automática da combinação de enzimas de restrição.*

• Naderi A, Teschendorff AE, Pinder SE, **Barbosa-Morais NL**, Paish CE, Ellis IO, Brenton JD, Caldas C. "Microarray Expression Signature predicts the outcome of Postmenopausal patients with Breast Cancer". *Oncogene*, in press 2006 [Naderi et al., 2006]
Contribuição individual: recolha, preparação e anotação bioinformática cruzada dos resultados de diferentes estudos de microarrays *usados no teste das previsões feitas com base nas assinaturas de expressão geradas.*

○ Gama-Carvalho M, **Barbosa-Morais NL**, Brodsky AS, Silver P, Carmo-Fonseca M. "Genome wide identification of functionally distinct subsets of cellular mRNAs associated with the mammalian splicing factors U2AF^{65} and PTB". *Submitted* 2006 [Gama-Carvalho et al., 2006]
Contribuição individual: análise e anotação bioinformáticas de sequências (busca de motivos de ligação ao RNA para os factores de splicing *em estudo) e participação no tratamento estatístico dos respectivos resultados.*

○ Mollet I, **Barbosa-Morais NL**, Andrade J, Carmo-Fonseca M. "Diversity of human U2AF splicing factors". (Review) *Submitted* 2006 [Mollet et al., 2006]
Contribuição individual: análise evolutiva; colaboração no estudo da estrutura de domínios funcionais.

◦ Teschendorff AE, Naderi A, **Barbosa-Morais NL**, Pinder SE, Ellis IO, Aparicio S, Brenton JD, Caldas C. "A consensus molecular prognostic classifier for ER positive breast cancer". *Submitted* 2006 [Teschendorff et al., 2006b]
Contribuição individual: recolha, preparação e anotação bioinformática cruzada dos resultados de diferentes estudos de microarrays *usados na meta-análise descrita.*

Abbreviations

A - adenosine
ALPS - autoimmune lymphoproliferative syndrome
API - application programming interface
ASCII - American Standard Code for Information Interchange
ATP - adenosine triphosphate
BAC - bacterial artificial chromosomes
BLAST - Basic Local Alignment Search Tool
BBP - branch point binding protein
BP - branch point
C - cytidine
cDNA - complementary DNA
C-terminal - carboxy-terminal
CDC - cell division cycle
CELF - CUG-BP and ETR3-like factor
CGH - comparative genomic hybridization
CLK - CDC-like kinase
CNE - conserved non-coding element
CTD - carboxy-terminal domain
CUG-BP - CUG-binding protein
DDC - duplication-degeneration-complementation
DNA - desoxyribonucleic acid
ELAV - embryonic lethal abnormal visual
EMBL - European Molecular Biology Laboratory
ESE - exonic splicing enhancer
ESS - exonic splicing silencer
EST - expressed sequence tag
FasL - Fas ligand
FAST - Fas-activated serine/threonine kinase
FIR - FUSE-binding protein-interacting repressor
G - guanosine
GMP - guanosine monophosphate
HMM - Hidden Markov Model
hn - heterogeneous nuclear
I - inosine
ISE - intronic splicing enhancer
ISS - intronic splicing silencer
IUPAC - International Union of Pure and Applied Chemistry
KH - K-homology
KSRP - KH-type splicing regulatory protein
LINE - long interspersed nucleotide elements

LS - least squares
LTR - long terminal repeats
m - messenger
ME - minimum evolution
ML - Maximum-Likelihood
MMASS - Microarray-based Methylation Assessment of Single Samples
MP - Maximum parsimony
Mya - million years ago
Myr - million years
n - neuronal
N - any nucleotide
NJ - Neighbor-Joining
NMD - nonsense-mediated decay
NPC - nuclear pore complexes
ORF - open reading frame
OTU - operational taxonomic units
p - phosphodiester bond
PACK - Profile Analysis using Clustering and Kurtosis
PCA - principal component analysis
PCB - poly-C binding protein
poly-A - polyadenylation
PTB - polypyrimidine tract-binding protein
PUF60 - poly U binding Factor-60kDa
Py - polypyrimidine
R - purine
RNA - ribonucleic acid
RNAP - RNA polymerase
RNP - ribonucleoprotein
RRM - RNA recognition motif
RS - arginine/serine rich
RT - reverse transcription
S - strong hydrogen bonding
sFas - soluble Fas
SMART - Simple Modular Architecture Research Tool
sn - small nuclear
SNAP - Synonymous Nonsynonymous Analysis Program
SNP - single nucleotide polymorphism
SQL - Structured Query Language
SR - serine/arginine rich
SRPK - SR-protein-specific kinase
ss - splice site
T - thymidine
U - uridine
U2AF - U2 snRNP auxiliary factor
UCSC - University of California - Santa Cruz
UHM - U2AF homology motif
UPGMA - unweighted pair-group method using arithmetic averages
USER - untranslated sequence elements for regulation
UTR - untranslated region
Y - pyrimidine

List of Figures

List of Tables

Acknowledgements

The work presented in this dissertation was kindly supported by:

▷ Fundação para a Ciência e a Tecnologia, Portugal / European Social Fund (3rd Framework Programme) - Fellowship PRAXIS XXI SFRH/BD/2914/2000

▷ EMBO (European Molecular Biology Organization) Short-Term Fellowship (ASTF 44-2005)

▷ Wellcome Trust, UK

▷ European Comission - Project RNOMICS (QLG2-CT-2001-01554)

▷ Cancer Research UK (grants to Simon Tavaré)

This work would have not been possible without the help of several people to whom I want to express my gratitude.

Thanks to Professor Maria do Carmo Fonseca for giving a physicist the opportunity to succeed in biology, by hosting me in her research group. I am deeply grateful for her scientific excellence, interest, unconditional support and friendship over these years. I am also thankful to her for being provided the privilege of working in top laboratories with other top scientists.

I am very grateful to Dr. Sam Aparício for taking me as a member of his lab in Cambridge and introducing me to the wonders of genomic comparison and "state of the art" bioinformatics. His intellectual brilliance and guidance have been fundamental in my scientific education. I wish to thank him for his selfless friendship, for always stimulating my intellectual freedom and encouraging interaction and collaboration with other interesting scientists around.

I owe a debt of gratitude to Dr. Juan Valcárcel for his generosity and patience as a perfect host in EMBL, for introducing me to the world of alternative splicing and

all the aspects of its regulation, for his permanent interest in my work, for important comments on the Genome Research manuscript and for consistent friendship and support.

I would like to acknowledge all my dear colleagues and friends at the Faculty of Medicine in Lisbon. They have strongly helped me to acquire biological language and intellectual framework. They always made me feel one of the group, despite my absence. Thank you Margarida Gama Carvalho for great friendship, the many exciting brainstorms and the collaboration. Thank you also for the invitation to teach in your PGDB courses, which proved to be a fantastic experience. Thanks to friend Teresa Raquel Pacheco for the fruitful collaboration and for all her interest and support. Thanks to Inês Mollet for complicit friendship and for sharing the struggles of solely bioinformaticians. I acknowledge Filipa Gallo's dedicated friendship, collaboration and precious help with compiling the splicing factors list. Thanks to Anita Gomes and Sandra Caldeira for sharing their scientific interests and projects with me. I am indebted to Professor João Ferreira for his permanent interest in my work, for recommending me to the Carmo Fonseca lab and for making my first real contact with Cell Biology such a pleasant experience. Thanks to Søren Steffensen for his help with phylogeny. Special thanks to José Braga and José Rino for many favors and companionship.

I would to thank my dear colleagues and friends at the Cancer Genomics Program (Hutchison/MRC Institute, Cambridge) for three of the best years of my life. I feel thankful for all the thrilling scientific interactions and for the friendly and familiar atmosphere. Thanks to my groupmates Cristian Brocchieri (also a great housemate!), Damian Yap, Katrin Mooslehner and Simone Polvani for support and companionship. I acknowledge Ali Naderi, Andrew Teschendorff, Ashraf Ibrahim, Suet-Feung Chin and Carlos Caldas for fruitful collaborations. Thanks to Bin Liu and James Brenton, for their patience and help with Linux, and to Paul Edwards for precious advice on the writing of this thesis. I thank friend Natalie Thorne for collaborations, support and mentorship. Thanks to Simon Tavaré and the whole Computational Biology Group for their support, their help in statistical issues, and for making the last stage of this adventure and the transition into microarrays a pleasant and exciting experience.

I wish to acknowledge my dear friend Ahmed Ahmed for many joyful and precious moments of complicity and brainstorming. Special thanks to Maria Garcia for the sharing, the empathy, the coffees and for the wonderful feeling of meeting a best friend at work everyday. Thanks to all my many friends in Cambridge for making me feel at home away from home.

I would like to thank Ângela Relógio for her friendship and for the splicing factors database. I am also grateful to my colleagues in the Valcárcel group, namely José Maria for sharing his projects and Luís Soares for fruitful discussions, and to all my other friends in EMBL for making those months in Heidelberg a great time. Thanks to Christine Gemünd for the splicing factors database. Special thanks to Caroline Lemerle for all her help with Perl scripting.

A big thank you to all my friends at IMCB Singapore for two fantastic weeks (August 2002), namely Elia Stupka, a perfect host, for great support, Alan Christoffels for help with phylogeny and Shawn Hoon for stimulating discussions on splicing regulatory elements.

I would like to acknowledge José Leal for friendship, mentorship and support and everybody in the Teichmann group (LMB), namely Sarah, for exciting scientific interaction.

I thank the ASD Team, specially Alphonse Thanaraj, for the collaboration and for providing a public home to the Splicing Rainbow.

Thanks to Christopher W. J. Smith for critical discussions and Carol Featherston for editing of the Genome Research manuscript.

I acknowledge people from the CRG in Barcelona, namely Claudia Ben-Dov, Roderic Guigó and Eduardo Eyras, for their interest, support and important discussions.

Thanks to everybody I have scientifically interacted with over the last years (too many to be named here) for their intangible but precious contributions to my work.

Muito obrigado a todos os meus entes queridos, Família e Amigos, pelo afecto inexorável e pela tolerância para com as minhas ausências de emigrante, provando-me que a distância pode mesmo ser só geográfica. Um agradecimento especial à linda Ana, cujos mimos, estímulo e dedicação foram factores de motivação e concentração

na fase final desta etapa.

O meu último agradecimento é para os meus Pais Isabel e Luís, a quem esta dissertação é dedicada, pelo Amor incondicional, pelo apoio permanente, por sempre terem criado as condições emocionais e logísticas para o meu sucesso e por me terem educado sobre pilares de liberdade, responsabilidade e honestidade.

Sumário

Palavras-chave: expressão génica, *splicing*, evolução, bioinformática, *microarrays*

A descodificação de um genoma eucariota no seu proteoma envolve a produção intermédia de um RNA mensageiro (mRNA). Nas células eucariotas, o mRNA é gerado e processado no núcleo em várias etapas complexas antes de ser exportado para o citosol, onde é traduzido em proteína. As etapas de processamento incluem modificações das extremidades 5' e 3', remoção de sequências não codificantes (intrões) e frequentemente edição de RNA. Quando comparados com procariotas, os eucariotas beneficiam de níveis extra de regulação da expressão génica (transcrição, processamento de pre-mRNA, exportação de mRNA, degradação de mRNA), nomeadamente devido à separação espacial entre os locais em que o mRNA é gerado (núcleo) e traduzido (citoplasma). O mesmo gene pode originar RNAs e proteínas diferentes, o que aumenta o potencial codificante de um genoma e constitui uma base para a complexidade do organismo. Por exemplo, os metazoários beneficiam de grande diversidade na especialização celular e isso exige um controlo apertado da expressão de um subgrupo particular desses genes para cada tipo de célula ou para cada estádio do desenvolvimento celular. Para além disso, a resposta de uma célula a factores fisiológicos e ambientais requer uma regulação dos produtos génicos dependente de sinais extracelulares [Gama-Carvalho, 2002; Relógio, 2002; Orphanides and Reinberg, 2002].

A maioria dos genes em eucariotas exibe uma estrutura de sequências codificantes (exões) intercaladas por sequências não codificantes, normalmente mais longas (intrões). A reacção de processamento do pre-mRNA que permite a remoção dos intrões chama-se *splicing*. O *splicing* é levado a cabo pelo spliceossoma, um

grande complexo macromolecular que se monta nos sítios de *splicing* (i.e. junções intrão/exão), processando-os. As proteinas integrantes da maquinaria molecular responsável pelo *splicing* do pre-mRNA são designadas por factores de *splicing*. O *splicing* é um aspecto particularmente crucial da regulação génica. O processamento do pre-mRNA tem que ocorrer com extrema precisão para que a mensagem codificada no mRNA e traduzida no ribossoma seja a correcta [Burge et al., 1999]. Por outro lado, o potencial codificante e a versatilidade funcional de cada gene são aumentados por formas alternativas de *splicing* e a consequente produção de isoformas proteicas diferentes, muitas vezes específicas de um determinado tecido ou estágio de desenvolvimento. O *splicing* alternativo tem um papel fundamental em processos como a apoptose, a diferenciação do sistema nervoso, o direccionamento axonal, a excitação e a contracção celulares, etc [Hastings and Krainer, 2001; Lopez, 1998; Smith and Valcarcel, 2000; Black, 2003]. Deficiências nos mecanismos de *splicing*, tais como mutações em sequências reguladoras e alteraçõs nos níveis de factores de *splicing*, podem originar splicing aberrante e causar doenças [Krawczak et al., 1992; Cartegni et al., 2002].

A conclusão de projectos de sequenciação de genomas inteiros permitiu à comunidade científica colocar, ao nível da sequência, muitas questões relevantes sobre a complexidade dos organismos. A sequenciação foi acompanhada pelo desenvolvimento de poderosas ferramentas de anotação genómica que permitiram tomar a sequência de DNA e acrescentar-lhe vários níveis de interpretação biológica. Um genoma pode ser anotado ao nível dos nucleótidos, ao nível das proteínas e ao nível dos processos [Stein, 2001; Brent, 2005].

A análise de um genoma individual fornece informação importante sobre a sua estrutura mas menos sobre a sua função. Há uma grande variedade de genomas sequenciados pelo que a genómica comparativa pode ser usada na anotação funcional. A comparação de genomas permite detectar sequências conservadas que podem corresponder a "assinaturas" funcionais, por força da selecção evolutiva. [Miller et al., 2004].

Nesse contexto, propusemo-nos contribuir para o estudo da evolução da complexidade da expressão génica e comparar os genes que codificam proteínas constitu-

intes da maquinaria de *splicing* em diversas espécies eucariotas representativas e, dessa forma, avaliar as especificidades funcionais de cada linhagem. Até há poucos anos, o estudo da história evolutiva da maquinaria de *splicing* era limitado pela ausência de sequências de genomas completos, que só recentemente começaram a estar disponíveis. Por exemplo, as sequenciação e anotação dos genomas do peixe-balão japonês *Fugu rubripes* [Aparicio et al., 2002] e da ascídia *Ciona intestinalis* [Dehal et al., 2002] permitem agora preencher esse vazio com os ramos fiduciais dos vertebrados distantes e dos cordados, respectivamente. Proporciona-se assim a oportunidade de pesquisar exaustivamente os factores de *splicing* nestas espécies e dessa forma alargar o nosso conhecimento sobre a sua evolução. Para o efeito, concebemos e desenvolvemos um protocolo computacional destinado a identificar e proceder à anotação semi-automática de factores de *splicing* em espécies de eucariotas representativas [Barbosa-Morais et al., 2006]. O estudo centrou-se em famílias de proteínas cuja função no *splicing* está confirmada por evidência experimental. Inspeccionámos visualmente 1894 proteínas, das quais 224 foram corrigidas manualmente.

A análise descrita mostra a conservação generalizada das proteínas constituintes da base estrutural do spliceossoma, nomadamente snRNPs (*small nuclear ribonucleoproteins*) e proteínas Sm, ao longo da linhagem eucariótica. Essa conservação contrasta com expansões selectivas de famílias de proteínas que se sabe estarem envolvidas na regulação do *splicing*, nomeadamente de proteínas SR (nucleares, ricas em serinas e argininas) em metazoários e hnRNPs (*heterogeneous nuclear ribonucleoproteins*, predominantemente proteínas nucleares de ligação ao RNA e com grande diversidade de actividades celulares) em vertebrados. Também observámos que as famílias de cinases CLK e SRPK (responsáveis pela fosforilação de proteínas SR) e a família de proteínas reguladoras de *splicing* CUG-BP/CELF se encontram expandidas em vertebrados. Para além disso, verificámos a existência de vários genes de factores de *splicing* monoexónicos em mamíferos, o que sugere que a complexidade da maquinaria de *splicing* naquela classe terá beneficiado do fenómeno de retrotransposição (processo pelo qual um mRNA, ou fragmentos dele, sofre transcrição reversa RNA→DNA e é inserido no DNA cromossomal).

Revimos e analisámos a evolução das famílias de proteínas relacionadas com o

factor U2AF, caracterizando a conservação da sua estrutura funcional e discutindo as implicações da sua diversidade na regulação do *splicing*. Os nossos estudos revelam, mais uma vez, algumas expansões selectivas em vertebrados e posteriores duplicações em linhagens específicas. Estes eventos sugerem que as proteínas U2AF e semelhantes terão funções únicas e de alta especificidade, ao nível do controlo da expressão génica em organismos complexos.

As ferramentas de análise e anotação de sequências podem também ser utilizadas na busca de sinais reguladores não codificantes na sequência de mRNA (elementos reguladores em *cis*) que actuam como alvos para os complexos proteicos envolvidos na biogénese do mRNA (reguladores em *trans*). Nesse sentido, desenvolvemos um programa, a que se deu o nome de Splicing Rainbow, destinado a prever potenciais sítios de ligação de factores de *splicing*, nomeadamente proteínas SR e hnRNPs, ao RNA. O programa envolveu, com base numa exaustiva recolha bibliográfica, a compilação de mais de 50 motivos e foi colocado *online*, como parte da *ASD Workbench* [Stamm et al., 2006].

Usámos a mesma abordagem na busca comparativa de motivos de ligação dos factores $U2AF^{65}$ (subunidade de 65 kDa do factor auxiliar do U2 snRNP) e PTB (hnRNP de ligação a tractos poli-pirimidínicos) em sequências completas de mRNA, discriminando região codificante e UTRs. Experiências de imunoprecipitação de RNA e *microarrays*[1], à escala genómica, identificaram mRNAs que interagem com os ditos factores. A classificação desses mRNAs em grupos de *Gene Ontology*[2] sugere que cada factor está associado com populações de mRNA funcionalmente coerentes (o $U2AF^{65}$ a mRNAs envolvidos na regulação da transcrição e do ciclo celular; a PTB a trancritos associados a transporte intracelular e compartimentos citoplasmáticos). A análise bioinformática, que visou perceber se as populações têm elementos de distinção nas respectivas sequências, mostra uma densidade significativamente maior de motivos

[1] *Microarrays* são pequenos suportes sólidos nos quais milhares de sequências codificadoras de genes ou transcritos (sondas) estão imobilizadas ou ligadas de forma organizada em posições conhecidas. As amostras podem ser DNA, cDNA ou oligonucleótidos. O DNA é impresso, depositado ou sintetizado diretamente no suporte.

[2] O projecto *Gene Ontology* visa fornecer um vocabulário dinâmico e controlado para descrever as características de um gene e respectivos produtos em qualquer organismo.

de ligação nas populações de mRNAs associados aos factores do que nas populações não associadas usadas como controlo. De um modo geral, os resultados do trabalho apoiam o modelo de interacção diferenciada entre populações de mRNA funcionalmente relacionadas e proteínas de ligação ao RNA de acção reguladora específica, através de elementos de sequência reguladores não traduzidos [Gama-Carvalho et al., 2006].

Demos ainda várias outras aplicações às ferramentas bioinformáticas de análise de reguladores de *splicing* em *cis*. Com base na busca dos hipotéticos motivos de ligação do factor TIA-1, procurámos intrões em genes envolvidos em apoptose com *splicing* alternativo susceptível de ser regulado pelo dito factor. Procurámos caracterizar o comportamento do spliceossoma no processamento de intrões com sítio de *splicing* a 3' de sequência "AGAG" (uma vez que "AG" é o consenso para sítio de *splicing* a 3' de um intrão). Fizemos uma análise de sequências, à escala genómica, procurando exemplos de intrões potencialmente co-processados pelos dois tipos de spliceossoma[3] - não encontrámos qualquer intrão de características "híbridas". Descobrimos associações entre nucleótidos específicos na composição dos sítios de *splicing* intrónicos a 5'.

A compreensão dos mecanismos de expressão génica passa não só pela caracterização qualitativa dos factores envolvidos e dos produtos génicos mas também pela determinação efectiva da quantidade de mRNA transcrito num dado sistema. Os *microarrays* são assim uma ferramenta poderosa em estudos de complexidade, já que permitem a avaliação de padrões de expressão génica à escala genómica. A descrição detalhada dos genes representados num *microarray* é uma componente informativa fundamental para este tipo de experiências. Nesse contexto, propusemo-nos adaptar as ferramentas bioiformáticas de análise e anotação de sequências (desenvolvidas no âmbito do estudo dos reguladores de *splicing*) a projectos de *microarrays*. Essas colaborações envolveram trabalhos de natureza vária: anotação transcriptómica no

[3] A maioria dos intrões é processado por um splicessoma caracterizado por cinco snRNPs (U1, U2, U4, U5 e U6), para além de outros factores proteicos. No entanto, nos Metazoários há uma classe rara de intrões processados por uma maquinaria de *splicing* distinta, composta por quatro snRNPs (U11, U12, U4atac and U6atac) diferentes mas funcionalmente análogos aos bem caracterizados U1, U2, U4 e U6 snRNPs, respectivamente (o U5 snRNP é partilhado pelos dois tipos de spliceossoma).

desenvolvimento de um protocolo de optimização dos processos de purificação envolvidos na obtenção de RNA para *arrays* de expressão e que se provou ser gerador de dados altamente reprodutíveis [Naderi et al., 2004]; anotação conjunta e cruzada de diferentes conjuntos de dados de *microarrays* de expressão, numa meta-análise destinada à derivação de genes com padrões de expressão adequados ao seu uso no prognóstico em cancro da mama [Teschendorff et al., 2006b]; anotação cruzada e mapeamento genómico de clones de *arrays* envolvidos em diversos estudos de *array-CGH*, na definição de perfis moleculares de cancros da mama primários [Chin et al., 2006]; anotação genómica das sequências de sondas e respectiva estimação do número de sítios de restrição na optimização de combinação de enzimas, como parte de um método para identificação de perfis de metilação de ilhas CpG [4] à escala genómica [Ibrahim et al., 2006].

De um modo geral, este trabalho evidencia o enorme potencial de abordagens computacionais na resolução de questões fundamentais da Biologia Celular e Molecular, nomeadamente ao nível da expressão dos genes. O estudo da maquinaria de *splicing* em eucariotas constitui um contributo importante para a compreensão da evolução do *splicing* e da sua relação com a complexidade dos organismos. Os resultados apoiam uma forte relação entre as evoluções dos factores reguladores e do *splicing* alternativo, sugerindo que aqueles terão influenciado a pressão selectiva sobre os sítios de *splicing* e outros elementos reguladores em *cis*. Nesse domínio, este trabalho levanta questões susceptíveis de lançar novas linhas de investigação, nomeadamente ao nível da correlação entre a especificidade funcional dos factores por determinadas isoformas e os respectivos sítios de ligação, em diferentes espécies e tipos celulares. Finalmente, mostrámos ser possível integrar, no mesmo âmbito de ferramentas bioinformáticas, estudos completos de evolução, funcionais e de expressão, à escala genómica.

[4] Ilhas CpG são regiões do DNA na zona do promotor de um gene com uma grande concentração de pares citosina-guanina (ligadas por um fosfodiéster). Contrariamente aos pares CpG nas regiões codificantes de um gene, a maioria dos pares constituintes das ilhas CpG não são metilados se os genes são expressos. Esta observação sugere que a metilação de sítios CpG no promotor de um gene pode inibir a sua expressão.

Abstract

Keywords: gene expression, pre-mRNA splicing, evolution, bioinformatics, microarrays

The decoding of an eukaryote genome into its proteome involves the production of an intermediate mRNA, whose biogenesis pathway comprises several complex steps. Indeed gene expression in eukaryotes is known to be regulated at several levels through an integrated and co-ordinated network of interactions. Splicing is a particularly crucial aspect of gene regulation, as the processing of pre-mRNA to mature mRNA must be very precise to ensure that the correct message is translated. Moreover the coding potential and functional versatility of genes is increased by alternative splicing pathways that generate different protein isoforms, very often in a developmental state or tissue-specific way.

Splicing is carried out by the spliceosome, a large macromolecular complex that assembles onto special sequences at the intron/exon junctions. Although more than 200 human spliceosomal and splicing-associated proteins are known, the evolution of the splicing machinery had not been previously studied extensively. The recent near-complete sequencing and annotation of distant vertebrate and chordate genomes provides the opportunity for an exhaustive comparative analysis of splicing factors across eukaryotes.

I have developed a semi-automated computational pipeline to identify and annotate splicing factors in representative species of eukaryotes. My analysis shows a general conservation of the core splicosomal proteins across the eukaryotic lineage, contrasting with selective expansions of protein families known to play a role in the regulation of splicing, most notably of SR proteins in metazoans and of heterogeneous

nuclear ribonucleoproteins (hnRNP) in vertebrates. I also observed vertebrate-specific expansion of the CLK and SRPK kinases (which phosphorylate SR proteins), and the CUG-BP/CELF family of splicing regulators. Furthermore I report several intronless genes amongst splicing proteins in mammals, suggesting that retrotransposition contributed to the complexity of the mammalian splicing apparatus.

We have reviewed the conserved structural features that characterize the U2AF protein families, discussing the potential implications of their diversity for splicing regulation. My evolutionary studies reveal that some U2AF families also benefited from vertebrate-specific expansion and subsequent lineage-specific duplications, suggesting unique and highly specific functions for those proteins, in relation to control of gene expression in complex organisms.

My work also focused on the identification of splicing cis-regulatory elements and their functionality. Hence I developed a program to predict putative binding sites for splicing factors, namely SR proteins and hnRNPs. This approach was followed in the recognition of putative sequence elements discriminating mRNA populations experimentally shown to interact with splicing factors U2AF and PTB. We performed a comparative sequence motif search for consensus U2AF and PTB binding sites in full length transcripts and our results suggest differential interaction between functionally related mRNA populations and specific regulatory RNA-binding proteins, through untranslated sequence regulatory elements. Likewise, I used sequence motif recognition to search for introns in apoptosis-related genes whose alternative splicing could be regulated by factor TIA-1. I have looked at splice site features too. I have tried to characterize the spliceosome's behaviour when processing introns with "AGAG" 3' splice sites. I have performed a genome-wide search for introns potentially co-processed by major and minor spliceosomes - I found no "hybrid" intron. I have found some associations between specific nucleotides in the composition of intronic 5' splice sites.

I have applied our sequence analysis and annotation tools to several microarray projects. Microarrays are a powerful tool in complexity studies as they allow the evaluation of genome-wide patterns of gene expression. I have contributed to studies that involve goals as diverse as the optimization of RNA purification steps, the iden-

tification of robust prognostic meta-gene sets for outcome in breast cancer, the use of array-CGH to define molecular portraits of primary breast cancers, or the profiling of CpG islands.

In general, this work puts into evidence the use and potential of bioinformatics in approaching fundamental questions in Molecular and Cell Biology, namely at the gene expression level. By studying the complete machinery of splicing across eukaryotes, I have given an important contribution on the evolution of splicing and its relation to the complexity of organisms. My results indicate a strong link between the evolutions of regulatory factors and alternative splicing, suggesting that the former influenced the selective pressure on splice sites and other cis-regulatory elements. Furthermore, this work raises some questions susceptible of triggering new lines of research, namely on correlating functional specificity of individual factors for their splice isoforms with the cognate recognition sequences in different species or cell types. Finally, I show that complete genome-wide studies on evolution, function and expression can be integrated in one consistent bioinformatics framework.

Chapter 1

Introduction

The decoding of an eukaryote genome into its proteome involves the production of an intermediate messenger ribonucleic acid (mRNA). In eukaryotic cells, the mRNA is generated and processed in the cell nucleus through several complex steps before being exported to the cytosol, where translation into protein takes place. Processing steps include 5' and 3' ends modification, removal of non-coding sequences (splicing) and sometimes RNA editing. When compared to prokaryotes, eukaryotes benefit from extra levels of gene expression regulation (transcription, pre-mRNA processing, mRNA export, mRNA degradation, translation regulation, posttranslational modifications), namely due to the spatial separation between the sites of mRNA generation (nucleus) and translation (cytoplasm). Indeed different RNAs and proteins can be produced from the same gene, which increases the coding potential of an eukaryotic genome and constitutes a basis for complexity. For instance metazoans benefit from great diversity in cell specialization and this requires a tight control of the expression of a particular subset of these genes for each cell lineage or developmental state. Moreover, the response of a cell to physiological and environmental factors requires a regulation of gene products dependent on intra and extracellular signals [Gama-Carvalho, 2002; Relógio, 2002; Orphanides and Reinberg, 2002].

Most eukaryotic genes exhibit a structure of coding sequences (exons) interrupted by non coding sequences, generally longer (introns). The reaction of intron removal in the processing of pre-mRNA is called splicing. Splicing is a particularly crucial aspect

of gene regulation and this processing of pre-mRNA to mature mRNA must occur with extreme precision to ensure that the correct message is translated at the ribosome. Additionally, the coding potential and functional versatility of genes is increased by alternative splicing pathways that generate different protein isoforms, very often in a developmental state or tissue-specific way. Alternative splicing plays a key role in the regulation of developmental and cellular processes like apoptosis, nervous system differentiation, sex determination, axon guidance, cell excitation and contraction, etc [Hastings and Krainer, 2001; Lopez, 1998; Smith and Valcarcel, 2000; Black, 2003]. Defects in splicing mechanisms, such as mutations in regulatory sequences and changes in levels of the proteins comprising the splicing machinery, can lead to aberrant splicing and cause disease [1].

The completion of whole-genome-sequencing projects has allowed scientists to address, at the sequence level, many relevant questions on organisms complexity. Sequencing was accompanied by the development of very powerful and useful genome annotation tools. It is now possible to take the raw DNA sequence and add layers of interpretation providing its biological significance. For example, the analysis of complete sets of translated open reading frames (ORFs) in mammals revealed a surprising abundance of alternatively spliced RNAs. There are several levels at which a genome can be annotated: nucleotide-level, protein-level and process-level [Stein, 2001; Brent, 2005].

The analysis of an individual genome provides important information on its structure but less on its function. Genomes have been sequenced for a wide range of organ-

[1] More than a decade ago, a first survey of point mutations in mRNA splice junctions led to an estimate of more than 15% of mutations causing disease involve splicing [Krawczak et al., 1992]. This is clearly an underestimate as it does not consider silent mutations in coding regions that, for example, affect exonic splicing enhancers and may therefore have an effect on the translated product [Cartegni et al., 2002; Pagani and Baralle, 2004]. Indeed systematic studies showed that genomic variants affecting splicing (many of which not affecting the consensus splice sites) are involved in near 50% of ataxia telangiectasia and neurofibromatosis type 1 cases [Teraoka et al., 1999; Ars et al., 2000]. Splicing defects are connected with other pathologies as diverse as thalassemia, myotonic dystrophy, Menke disease, occipital horn syndrome, familial disautonomia, cystic fibrosis, Frasier syndrome, retinitis pigmentosa and some forms of cancer [Faustino and Cooper, 2003; Hastings and Krainer, 2001; Nissim-Rafinia and Kerem, 2002; Sharp, 1994; Smith and Valcarcel, 2000].

isms and therefore comparative genomics can be used in the functional annotation. Comparing genomic sequences may allow to find conservation signatures, as functional sequences are subject to evolutionary selection [Miller et al., 2004]. One tool for the study of evolution of gene expression complexity is to compare, across representative species of eukaryotes, the genes encoding proteins that constitute the mRNA processing machinery and hence assess lineage functional specificities. Moreover, sequence analysis tools can be used in the search for non-coding mRNA regulatory signals that act as targets for the protein complexes involved in mRNA biogenesis.

Measuring actual levels of gene expression involves assessing the amount of transcribed mRNA in a given system. Microarrays are thus a powerful state-of-the-art tool in complexity studies, as they provide a genomewide large scale evaluation of gene expression profiles and patterns.

1.1 Pre-mRNA Splicing

1.1.1 mRNA biogenesis pathway

Due to their complexity, the consecutive steps of mRNA biogenesis have initially been studied separately but recent work shows that they are not independent (Figure 1.1) [Orphanides and Reinberg, 2002; Moore, 2005]. Functional links between the protein factors that carry out the different steps in the gene expression pathway have been revealed and physical interactions between the various machineries have also been uncovered. The protein factors responsible for each individual step in gene expression are functionally and sometimes physically connected. Regulation of the pathway from gene to protein is controlled at multiple stages but there are no general rules describing how the pathway is regulated. Different classes of genes are regulated at different stages [Orphanides and Reinberg, 2002; Moore, 2005].

RNA polymerase II (RNAP II), a 12-subunit complex, is responsible for the transcription of mRNA in eukaryotic cells. It catalyzes the DNA-dependent synthesis of mRNA but requires accessory proteins, named general transcription factors, as it can not recognize the promoters of target genes [Woychik and Hampsey, 2002; Orphanides and Reinberg, 2002]. The initiation of transcription involves several steps: assembly

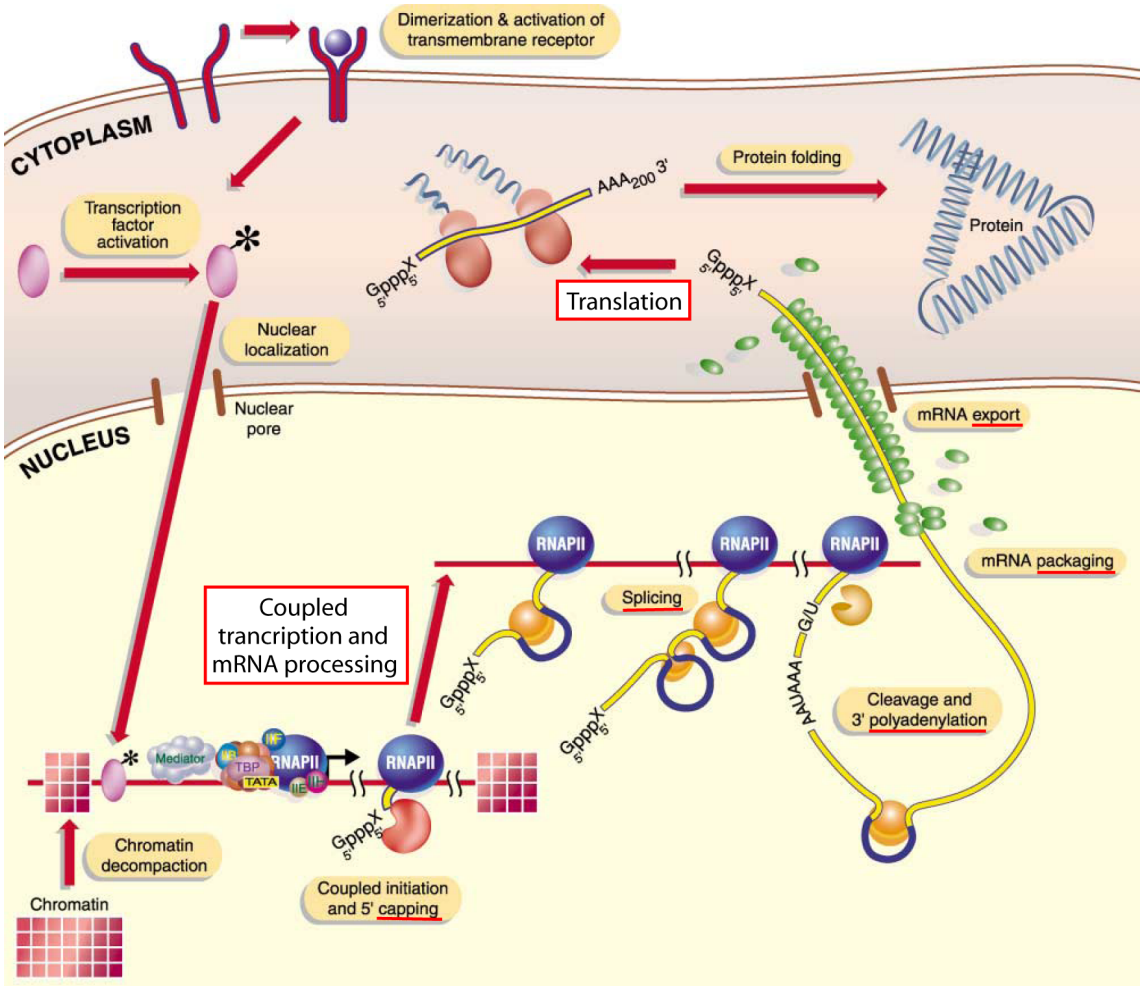

Figure 1.1: Gene Expression

In this current model for the regulation of gene expression, each step is physically and functionally connected to the next (see text for details). (Adapted from [Orphanides and Reinberg, 2002].)

of a so-called pre-initiation complex at the core promoter, separation of the DNA strands at the initiation site ("promoter melting"), formation of the first phosphodiester bonds of the transcript, and disruption of the interactions between RNAP II and the promoter ("promoter clearance"). The elongation stage begins with a massive phosphorilation of the C-terminal domain (CTD) of the large subunit of RNAP II. It is currently accepted that the CTD is a platform for the ordered assembly of the different families of pre-mRNA processing factors, as it has been shown that the CTD is essential for efficient pre-mRNA processing and interacts specifically with all classes of processing machineries. CTD phosphorylation is therefore likely to coordinate the recruitment of pre-mRNA capping, splicing, and 3' processing factors (see below) at different stages in the nascent mRNA synthesis. Transcription termination consists in the release of transcripts from the site of transcription and of RNAP II from the DNA template [Proudfoot et al., 2002]. Currently this process is not fully understood but it has been shown to depend on the existence of a polyadenylation (poly-A) signal and to be triggered by the 3' end processing.

The first pre-mRNA processing step is the addition of a 'cap' at the 5' end of the mRNA to prevent degradation of the transcript by exonucleases. Capping occurs after 20–30 nucleotides have been synthesized and is a three-step reaction: first an RNA 5' triphosphatase hydrolyzes the triphosphate of the first nucleotide to a diphosphate; then a guanylyltransferase catalyzes the addition of a GMP (guanosine monophosphate) to the first nucleotide of the pre-mRNA via an unusual 5'-5' triphosphate linkage; finally a methyltransferase methylates the N7 position of the transferred GMP [Shatkin and Manley, 2000]. This initial cap structure is then recognized by the cap binding complex and it is believed to then play a major role in the stabilization of the mRNA, as it represents an obstacle for 5'-3' exonucleases. It also enhances translation by promoting the engagement of the ribosomal subunits with the mRNA [Proudfoot et al., 2002].

Most of the genes in metazoa are interrupted by long noncoding sequences named introns. Thus, in order to generate a functional message from the DNA template, introns must be spliced out of the RNA copy of the gene. The pre-mRNA splicing mechanism is described in 1.1.2 and below.

All eukaryotic protein encoding mRNAs (with the exception of replication-dependent histone genes in higher eukaryotes) contain a uniform and protective 3' end comprising about 200 adenosine nucleotides. The 3' end processing consists in a two-step reaction: the mRNA is cleaved and then polyadenylated. The formation of the poly-A tail involves the recognition of specific sequences present on the pre-mRNA and the polyadenylation machinery, consisting of at least six multimeric protein factors [Proudfoot et al., 2002; Zhao et al., 1999].

Besides capping, splicing and 3' end processing, there are other types of post-transcriptional modifications broadly defined as RNA editing. The most common in mammals are deamination reactions, like the conversion of C to U and of A to I (inosine or isoleucine - read as G in translation), but insertion or deletion of particular bases have also been reported. These modifications can affect both coding and non-coding (namely intronic) sequences and are suggested to regulate splicing and to have a role in processing and stability of mRNAs [Keegan et al., 2001].

After being synthesized and before being translated, the mRNA must be 'exported' from the nucleus to the cytoplasm. Specialized gates termed nuclear pore complexes (NPC) span the nuclear envelope. It is believed that mRNA transport proteins recognize and bind to a conserved element found in processed transcripts, forming splicing-dependent mRNP complexes, and target them to nuclear pores, while hnRNP proteins (known to be splicing factors - see 1.1.3) retain introns in the nucleus [Reed and Magni, 2001; Moore, 2005]. Indeed it has been suggested that splicing and export are coupled, as some other splicing factors (namely SR proteins - see 1.1.3) are known to be involved in mRNA transport and splicing of pre-mRNAs is known to promote their export [Orphanides and Reinberg, 2002].

Translation of mRNA is not independent of its biogenesis [Gama-Carvalho, 2002]. The efficiency of translation is regulated by proteins that travel with the mRNPs to the cytoplasm and a first round of translation is known to be a quality control step (by detecting mRNAs with premature stop codons)[2]. There are further mechanisms that control the quality of a nascent transcript, preventing the synthesis of spurious

[2]Moreover, work reporting nascent polypeptides in nucleic sites suggests low level translation in the nucleus, coupled with transcription [Iborra et al., 2001].

proteins that could inflict damage to the cell by proofreading the messages and regulating mRNA stability. For instance, mRNA turnover pathways are decisive in gene expression. Nuclear exonucleases are believed, not only to degrade introns, but to contribute to the elimination of inefficiently processed pre-mRNAs and malformed mRNAs [Moore, 2002].

1.1.2 Mechanism of Splicing

The splicing of pre-mRNA involves two transesterification reactions (Figure 1.2) [Burge et al., 1999]. In the first, the 3'-5'-phosphodiester bond at the 5' splice site is attacked by the 2'-hidroxyl group of the conserved intronic adenosine at the branch site. A 2'-5' phosphodiester bond is formed, generating a lariat and a free 5' exon, with a 3'-hidroxyl group. This group then attacks the phosphodiester bond at the 3' splice site, releasing the intron lariat (to be degraded) and ligating the exons[3].

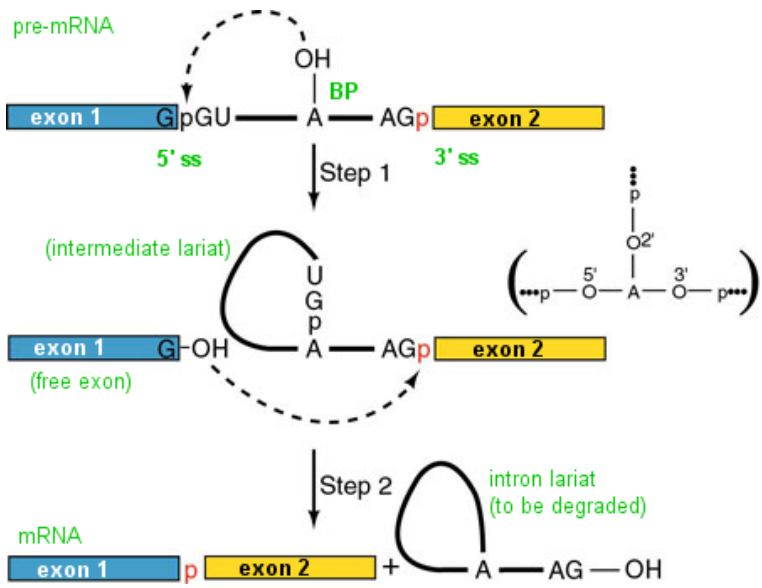

Figure 1.2: Splicing chemical mechanism

5' ss and 3' ss - splice sites; BP - Branch Point; p - phosphodiester bond (see text for details).

[3]Most pre-mRNAs undergo the described mechanism, which can be termed *cis*-splicing. However, in some species (namely Trypanosomes and Nematodes), the process can occur as *trans*-splicing: 5' and 3' splice sites are in different pre-mRNAs, fused during splicing [Nilsen, 2001].

The exact location of introns, exons and sites for the transesterification reactions are defined by conserved elements within the gene sequence [Burge et al., 1999].

In *Saccharomyces cerevisiae* the 5' splice site is defined by the consensus sequence R|GUAUGU, the branch point by UACUA<u>A</u>C and the 3' splice site by CAG|N (preceded by a poly-U tract)[4]. These very specific elements are sufficient for a proper recognition of the splice sites by the splicing machinery and the subsequent intron excision.

In higher eukaryotes, the consensus sequences are more degenerate and therefore less specific (Figure 1.3). Although essential, they are not sufficient for splicing [Green, 1986]. For metazoans, only |GT at the 5' splice site, AG| at the 3' splice site and the branch site A (aproximately 18-40 nucleotides upstream of the 3' splice site) are very conserved. There is also a poly-Y (polypyrimidine) tract (10-20 nucleotides long) upstream of the 3' splice site. As these motifs do not provide full specificity for splice site determination, other sequence elements (intronic and exonic splicing enhancers and silencers) are involved in splice site selection [Blencowe, 2000].

In many vertebrates, insects and plants there is a minor class of introns (aproximately 0.2% of human introns), lacking a poly-Y tract, with slightly different and highly conserved splicing signals: |RUAUCCUUU for 5' splice site, YAS| for 3' splice site and UCCUUA<u>A</u>C for branch point (10-20 nucleotides upstream of the 3' splice site)[5] [Burge et al., 1999; Zhu and Brendel, 2003; Levine and Durbin, 2001; Burge et al., 1998; Tarn and Steitz, 1996; Tarn and Steitz, 1997; Patel and Steitz, 2003]. These introns were originally named AT-AC introns and are spliced out by a different splicing machinery (the minor spliceosome).

[4]R represents a pu**R**ine (A or G), <u>A</u> the branching nucleotide, N any nucleotide and the vertical bar | the splice junction.

[5]R represents a pu**R**ine (A or G), <u>A</u> the branching nucleotide, Y a p**Y**rimidine (C or T), S a **S**trong hydrogen bonding (C or G, following the IUPAC ambiguous nucleotide code) and the vertical bar | the splice junction.

1.1.3 Spliceosome

Splicing is carried out by the spliceosome, a large macromolecular complex that assembles onto special sequences at the intron/exon junctions [Sharp, 1994; Kramer, 1996; Luhrmann et al., 1990].

The estimated number of human spliceosomal and splicing-related proteins has been rapidly increasing due to improvement and sophistication of methods used in their identification. For example, gel filtration has been combined with affinity-chromatography techniques [Reed, 1990], two dimensional gel electrophoresis methods and advanced mass spectrometry [Neubauer et al., 1998; Wilm et al., 1996] now make it possible to analyze very complex peptide mixtures by liquid chromatography coupled with tandem mass spec [Griffin and Aebersold, 2001]. Recently application of these methods has increased the apparent number of spliceosomal proteins [Rappsilber et al., 2002; Washburn et al., 2001]. More than 200 spliceosome-associated proteins are currently known [Burge et al., 1999; Black, 2003; Hartmuth et al., 2002; Zhou et al., 2002; Jurica and Moore, 2003; Neubauer et al., 1998; Rappsilber et al.,

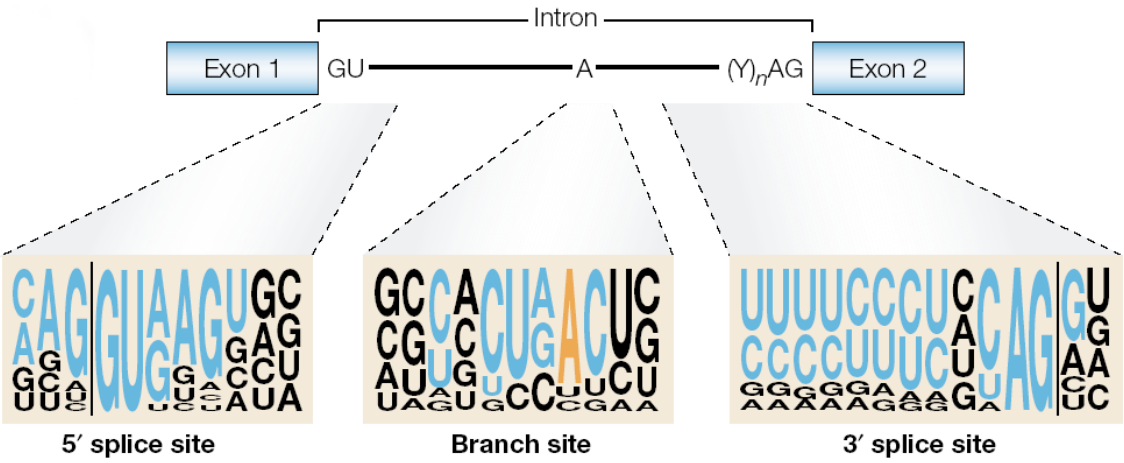

Figure 1.3: Consensus splicing signals

Schematics of a two-exon human pre-mRNA showing conserved intronic motifs. The size of each nucleotide is proportional to its frequency at the corresponding position (in an alignment of conserved intronic sequences). Classical consensus motifs are represented in blue and branch-point A in orange. Vertical lines represent intron/exon boundaries. (Adapted from [Cartegni et al., 2002].)

2002; Nilsen, 2003]. Although fluorescence and biochemical studies suggest a role in splicing for many of the novel proteins identified with sensitive protein detection methods [Rappsilber et al., 2002], many of these newer components are not functionally validated and therefore can not be classified as bona fide splicing factors [Jurica and Moore, 2003].

The spliceosome comprises small nuclear ribonucleoprotein particles (snRNPs) and a collection of protein splicing factors. At present, nine distinct types of snRNPs - U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac - are known to form spliceosomes[6]. Each snRNP is composed of a stable small nuclear RNA (snRNA) bound by a core ring of seven different so-called Sm proteins (B/N, D1, D2, D3, E, F, G)[7]. Each snRNP also contains several specific proteins responsible for cross-linking with pre-mRNA or protein-protein interactions in the spliceosome assembly [Luhrmann et al., 1990; Gozani et al., 1996; Gozani et al., 1998]. For example, the 70 kDa U1 snRNP specific protein comprises one RNA recognition motif (RRM) and two arginine/serine rich (RS) domains for interaction with other proteins (Figure 1.4).

Several DExD/H-box proteins (ATPases/RNA helicases or unwindases) identified both in yeast and mammalian systems are thought to mediate multiple RNA conformational changes occurring throughout the splicing process and hence to be required for the stepwise and dynamic assembly of the spliceosome. It is believed that they are required for the unwinding of short RNA-RNA duplexes that are formed between the different snRNAs or pre-mRNA molecules and for the dissociation of RNA-protein complexes. For instance, the Prp5 and UAP56 helicases, which are DEAD-box proteins (containing the conserved amino acid motif Asp-Glu-Ala-Asp, D-E-A-D in the single letter amino acid code), might facilitate the recruitment of the U2 snRNP. The 100 kDa subunit of the U5 snRNP is also a DEAD-box helicase (domain structure illustrated in Figure 1.4) and has been implicated in the ATP-dependent switch between the initial 5'SS:U1 snRNA duplex and the subsequent 5'SS:U6 snRNA pairing (see subsection 1.1.4). The first trans-esterification reaction requires the DEAH-box

[6]U11, U12, U4atac and U6atac form, together with U5, the minor spliceosome [Tarn and Steitz, 1997; Patel and Steitz, 2003].

[7]Exceptionally U6 snRNP contains a different but similar ring of Sm like (LSm) proteins [Salgado-Garrido et al., 1999; Seraphin, 1995]

protein Prp2 and subsequent steps (i.e. the second trans-esterification reaction, the release of the mature mRNA and the recycling of the spliceosomal components) require other DEAH-box proteins: Prp16, Prp22 and Prp43, respectively [Rocak and Linder, 2004; Ismaili et al., 2001].

Among the most important non-snRNP factors is the U2 snRNP auxiliary factor (U2AF) which in mammals is composed of a 35 and a 65 kDa subunits. $U2AF^{65}$ comprises three RNA recognition motifs (RRM), involved in the binding to the polypyrimidine tract, and an amino-terminal arginine/serine rich (RS) domain (Figure 1.4), associated with the recruitment of U2 snRNP to the branch site [Valcarcel et al., 1996]. $U2AF^{35}$ contains a pseudo-RRM and carboxy-terminal RS domain and is known to interact with the AG dinucleotide at the 3' splice site [8] [Merendino et al., 1999; Wu et al., 1999].

Accessory proteins mediate additional functions in cooperation with the spliceosome. For example, SR (serine/arginine-rich) proteins, highly conserved throughout metazoans, play an important role [Graveley, 2000; Hastings and Krainer, 2001; Longman et al., 2000; Mount and Salz, 2000; Tacke and Manley, 1999] by facilitating spliceosome assembly as mediators of the interaction between snRNPs, or by substrate specific exonic splicing enhancers (ESEs) detection [Blencowe, 2000; Cartegni et al., 2002]. By binding to ESEs, the SR proteins help to recruit $U2AF^{65}$ (through interaction with the RS domain of $U2AF^{35}$) and establish the link between the factors associated with the 3' splice site of an intron and factors associated with the 5' of the following intron (performing exon recognition). The degeneracy of ESE sequences may allow overlap in binding of proteins with antagonist effects, different binding affinities and variable expression levels. SR proteins can also act as 'passive' enhancers, by antagonizing splicing silencers. Specificity is therefore introduced by combinatorial control. This has a key role on the regulation of alternative splicing [Smith and Val-

[8] Introns can be classified as AG independent, if they can undergo de first step of splicing without a conserved 3' splice site, and AG-dependent, if splicing is disrupted when the AG is mutated [Reed, 1989]. A strong poly-Y tract can compensate for the absence of the AG dinucleotide, which indicates that the recognition of the poly-Y tract by $U2AF^{65}$ is sufficient AG-independent introns to be spliced out. In AG-dependent introns, $U2AF^{35}$ proves to be indispensable, mediating the recognition of the 3' splice site through the interaction with the AG.

carcel, 2000]. SR proteins have one or two N-terminal RRMs that interact with the pre-mRNA and a C-terminal RS domain (repetitions of arginine/serine dipeptides) (Figure 1.4) mainly responsible for protein-protein interactions but recently shown to interact also with RNA [Shen and Green, 2004; Shen et al., 2004]. These are features that confer splicing activation properties to SR proteins. SR-related nuclear matrix proteins are known to act as splicing coactivators [Blencowe et al., 1998; Eldridge et al., 1999]. CLK (CDC-like) and SRPK (SR-protein-specific) kinases phosphorylate SR proteins, modulating their function in splicing [Duncan et al., 1998; Prasad et al., 1999; Ngo et al., 2005]. They include a conserved catalytic domain (illustrated for protein Clk2 in Figure 1.4) whose N-terminal extremity has been shown to be involved in ATP binding [Hanks and Hunter, 1995].

Like SR proteins, heterogenous nuclear ribonucleoproteins (hnRNPs), whose structure usually includes one or more RRMs (Figure 1.4) and some auxiliary domains, are important alternative splicing regulators. They are a large group of molecules identified by their association with unspliced mRNA precursors (hnRNA) and are not a single family of related proteins [Krecic and Swanson, 1999]. They bind to many intronic and exonic splicing enhancers and repressors. For example, hnRNP A1 is a splicing factor shown to counteract SR proteins, acting as an exonic repressor. Some genes have tissue-specific splicing patterns that are sensitive to the relative ratio of hnRNP A1 to ASF/SF2 (SR protein and important splicing regulator). hnRNP A1 can also act as an intronic repressor and even autoregulates the splicing of its own transcript [Black, 2003]. Some other hnRNPs are tissue or gene specific splicing regulators. In neurons, alternative splicing of the *c-src* Nl exon is regulated by an intronic enhancer binding a complex of hnRNPs F and H and KSRP (KH-type splicing regulatory protein) and a repressor binding PTB/hnRNP I [Modafferi and Black, 1999]. PTB is also involved in the regulation of alternative splicing for many other genes [Ashiya and Grabowski, 1997; Zhang et al., 1999; Lou et al., 1999; Wollerton et al., 2001]. hnRNPs are also involved in many other cellular activities and regulatory pathways: transcription regulation, telomere maintenance, mRNA translation and turnover, etc [Krecic and Swanson, 1999].

There are many other splicing factors that are tissue or gene specific. TIA-1 is

an RNA-binding protein (comprising 3 RRMs - Figure 1.4) known to be involved in a mechanism of alternative splicing regulation that controls biological processes as important as programmed cell death in humans [Forch et al., 2000]. Nova-1, a hnRNP-related protein, harbors three KH-type RNA-binding domains (Figure 1.4) and is expressed exclusively in neurons within the central nervous system, regulating neuron-specific alternative splicing [Jensen et al., 2000]. Likewise Elav (embryonic lethal abnormal visual) proteins have 3 RRMs (Figure 1.4) and are known to be involved in neuron-specific alternative splicing [Lisbin et al., 2001]. The CUGBP (CUG-binding) and ETR-like protein family (CELF) also includes neuron-specific splicing regulators (e.g. NAPOR [Zhang et al., 2002]). CELF proteins are generally implicated in tissue-specific and developmentally regulated alternative splicing [Ladd et al., 2001].

1.1.4 Spliceosome assembly

In higher eukaryotes, spliceosome assembly is directed by three major conserved sequence elements in the pre-mRNA: the 5'-splice site, the branch point and the 3'-splice site. These sequences are recognized by the snRNPs and protein splicing factors through protein-RNA interactions and snRNA-pre-mRNA base pairing.

Splicing of the vast majority of introns depends on base-pairing interactions involving the recognition of the 5' splice site (which first two nucleotides are usually GT) by U1 snRNP (and base pairing with U1 snRNA) and interaction between U2 snRNP and the 3' splice site. This involves a sequential recognition of the branch point sequence: first by the protein SF1/BBP (in the so called "E" complex - the first functional intermediate in spliceosome assembly) and then by the binding of U2 snRNP, through base-pairing with U2 snRNA (Figure 1.5). The later ATP-dependent association is mediated by both 35 and 65 kDa subunits of the heterodimeric U2 auxiliary factor (U2AF) which bind to 3' splice site (which last two nucleotides are usually AG) and the polypyrimidine tract (between branch point and 3' splice site) respectively [Banerjee et al., 2003; Guth and Valcarcel, 2000; Hastings and Krainer, 2001; Wu et al., 1999].

The "A" complex is then joined by the U4.U6/U5 tri-snRNP, in a ATP-dependent

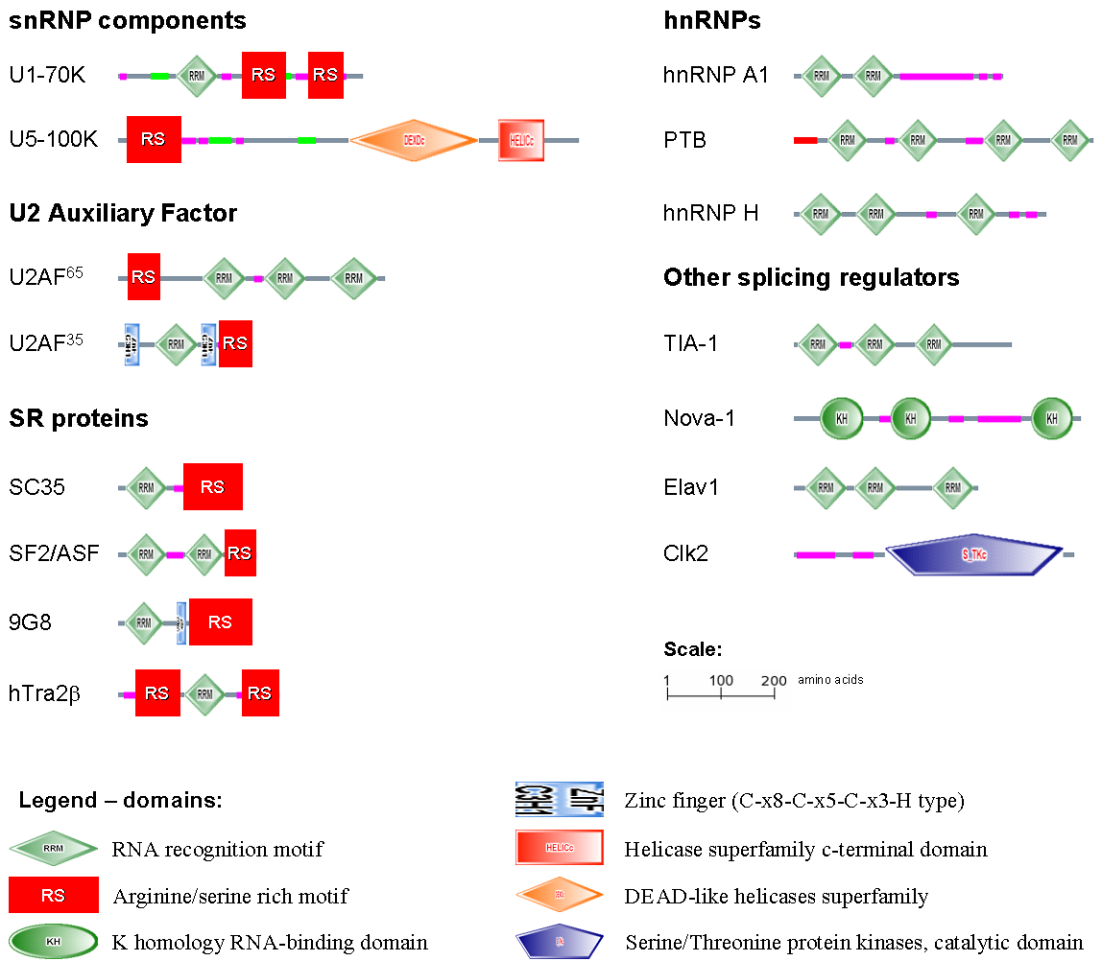

Figure 1.4: Domain structure of splicing factors

Diagrams are adapted from SMART [Letunic et al., 2004]. RS domain annotation is extracted from UniProt [Bairoch et al., 2005]. (Inspired in [Graveley, 2000] and [Black, 2003].)

process, and this "B1" complex is rearranged by the interaction of U6 with the 5' splice site and the lost of U1 an U4. The resulting "B2" (or "C") complex catalyses the two transesterification steps in which splicing takes place. The first consists of the cleavage of the 5' exon from the intron and the ligation of the 5'end of the intron to the branch point (producing a lariat structure). In the second, the lariat intron is released by the cleavage of the 3' splice site and the surrounding exons are ligated [Black, 2003; Sharp, 1994].

Additionally, metazoans contain a minor class of introns that share a distinct set of conserved elements: their 5' splice site consensus is different and longer and can start with GT or AT; the branch point consensus is stronger and longer; the 3' splice sites have AG or AC. Minor introns were recently identified for 183 known human genes [Levine and Durbin, 2001]. Splicing of minor introns is carried out by a distinct spliceosome that contains U11, U12, U4atac and U6atac snRNPs in the place of U1, U2, U4 and U6 respectively (U5 is common to both spliceosomes). Minor snRNPs and their major correspondents are shown to differ mainly on the srRNAs, their protein compositions are similar [Patel and Steitz, 2003; Will et al., 1999].

1.1.5 Alternative Splicing

As mentioned before, consensus sequence signals for splice sites are degenerate. Conserved sequence elements are not enough for unequivocal definition of intron/exon junctions and the splicing machinery is allowed to select between alternative splice sites. However, there is no randomness in this selection. The mRNA must be processed with extreme precision and specificity and alternative splicing is tightly regulated.

The majority of genes in metazoa are multi-exonic and the estimated proportion of human genes that undergo alternative splicing goes from 40% to more than 80% [Modrek and Lee, 2002; Johnson et al., 2003; Kampa et al., 2004]. For those genes, different combinations of exons can be spliced together and there are several types of alternative splicing events (Figure 1.6).

Alternative splicing affecting coding regions leads to the synthesis not only of different functional proteins but also of truncated and non-functional proteins (to

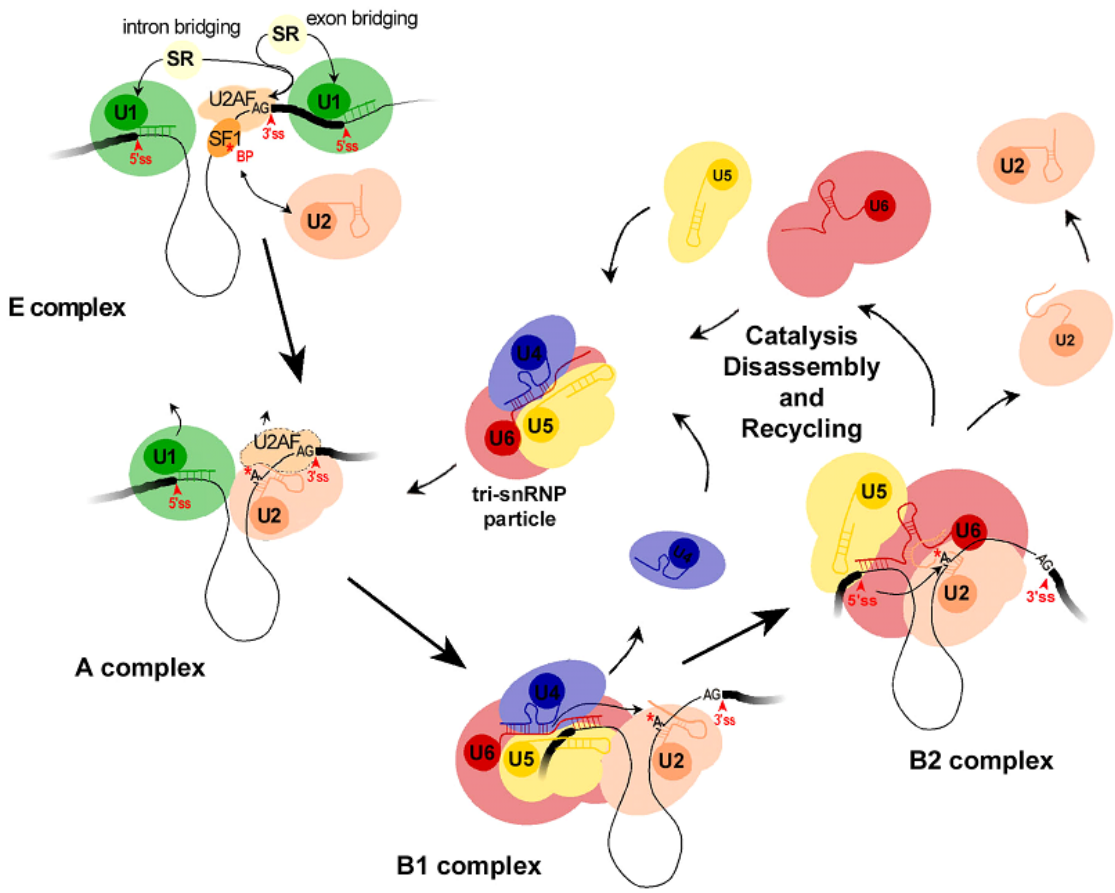

Figure 1.5: Spliceosome assembly.

Exons are represented by thick and introns by thin lines; protein particles (U snRNPs, U2AF and SF1) are represented by round shapes; snRNAs are depicted by the lines accompanying snRNPs; splice sites and branch point (A) are indicated (see text for details). (Adapted from [Gama-Carvalho, 2002].)

be degraded). Alternative splicing in untranslated regions (UTR) is known to be involved in the regulation of translation. Thus in both cases alternative splicing can act as a switch for the production of a protein [Graveley, 2001; Smith and Valcarcel, 2000].

In mammals, introns are outstandingly long when compared with exons. Therefore recognition of splice sites must rely primarily on exon-bridging interactions. Indeed most of natural mutations affecting the 5' splice site at the end internal exons originate exon skipping rather than intron inclusion [Robberson et al., 1990], which would be predicted assuming "cross-intron" splicing. An exon definition model, where the binding of U1snRNP to the 5' splice site stabilizes the interaction between U2AF and the upstream 3' splice site, has been proposed [Robberson et al., 1990; Maniatis and Tasic, 2002]. As described in 1.1.3, SR proteins mediate this 'cross-talk' between the upstream 3'ss and the downstream 5'ss by binding to ESEs (Figure 1.7).

Splicing is regulated by *cis* and *trans* elements: pre-mRNA sequences and so-called splicing factors, respectively. Although the relative 'strength' of splice sites has an important influence on the frequency of selection of an exon, regulation of both constitutive and alternative splicing relies, specially in vertebrates, on a complex system of other sequence elements. These can be both intronic and exonic and have both an enhancing and a silencing action - according to these features they are named with the acronyms ISE, ISS, ESE and ESS. Thus splice site selection depends on the

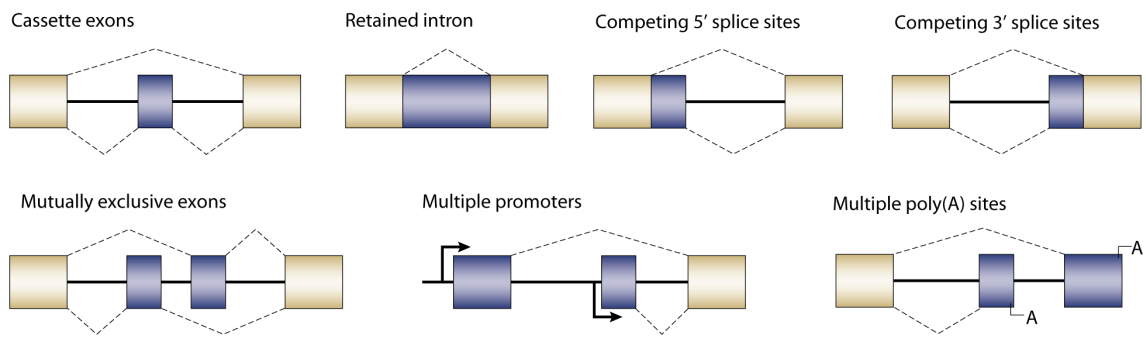

Figure 1.6: Types of alternative splicing events.
Introns are shown as solid lines, constitutive exons as beige boxes, alternative exons as blue boxes, splicing patterns above and below as dashed lines. (Adapted from [Matlin et al., 2005].)

balance between antagonistic activities (Figure 1.8A) [Matlin et al., 2005; Gama-Carvalho, 2002].

Several mechanisms of splicing activation and repression have been described. Most of ESEs comprise binding sites for SR proteins. As already mentioned, SR proteins are involved in the recruitment of the splicing machinery to the splice sites (namely of U2AF^{65} to weak poly-Y tracts) and play a decisive role in exon definition. They can also promote splicing just by antagonizing the silencing effect of an inhibitory protein that is susceptible of binding to an overlapping ESS. Typical ESE sequences can act as silencers when they fall near constitutive splice signals. Many silencers (both ISS and ESSs) include binding sites for hnRNPs, namely hnRNP A1, hnRNP F/H, hnRNP L and PTB/hnRNP I. Splicing repressors can exert their splicing inhibiting effect by directly competing with SR proteins for a binding site. They can also loop out and 'mask' an exon by binding to ISSs in the surrounding introns and dimerizing. Additionally, nucleation and cooperative binding of inhibitory factors can 'shield' ESEs and other binding sites for the splicing machinery. Apart from SR

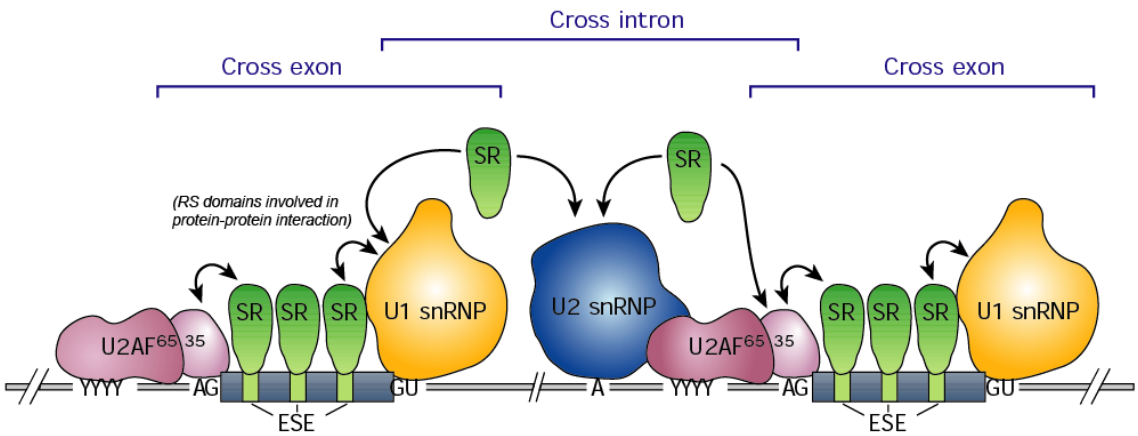

Figure 1.7: Exon definition

The splicing machinery recognizes 5' (GU) and 3' (AG) splice sites (ss) as exon flanking sequences. SR proteins, binding to exonic splicing enhancers (ESE), recruit U1 snRNP to the downstream 5'ss and U2AF to the upstream polypyrimidine (YYYY) tract (65 kDa subunit) and 3'ss (35 kDa subunit). U2AF recruits the U2 snRNP to the branch point (A). SR proteins function in both "cross-exon" and "cross-intron" recognition complexes. (Adapted from [Maniatis and Tasic, 2002].)

proteins and hnRNPs, there are other proteins known to act as tissue-specific splicing regulators: TIA1 and TIAR, Nova-1, NAPOR or members of the CELF/CUG-BP family. Examples of mechanisms for splicing silencing and enhancing can be found in Figure 1.8B. They illustrate how modulating the concentration of splicing regulatory factors can affect splice site choice and be determinant in alternative splicing [Cartegni et al., 2002; Matlin et al., 2005; Gama-Carvalho, 2002].

It has been suggested that the presence of more alternative splices in a species could account for further complexity but the notion of increased alternative splicing in higher eukaryotes (namely vertebrates) is still somewhat contentious. The relatively low number of human genes [Lander et al., 2001; Venter et al., 2001], when compared with simpler species, led, among many other hypothesis (greater gene modularity in human, post-translational modifications [Banks et al., 2000]), to the idea that alternative splicing may be responsible for more transcripts per gene and therefore a much larger proteome in human than in other species [Ewing and Green, 2000].

However, different large scale EST (expressed sequence tag) studies lead to different results. A recent estimate indicates greater amount of alternative splicing in mammals than in invertebrates [Kim et al., 2004] but those results were immediately disputed by the authors of a previous analysis which suggests that the total amount of alternative splicing is comparable among animals (mammals, insects and worms) [Brett et al., 2002].

Nevertheless, tissue and gene-specific alternative splicing patterns or subtle sophistication on the splicing regulatory pathways may contribute to an organism's complexity. For instance, alternative splicing is extensive in the brain of higher organisms and is known to be involved in the regulation of channel and receptor activities and synaptic function [Lipscombe, 2005].

1.2 Genome dymanics in Vertebrates

1.2.1 Genomic Expansion

It is believed that at least two rounds of whole-genome duplication (polyploidization), estimated to have occurred around 600 million years ago, are coincident with

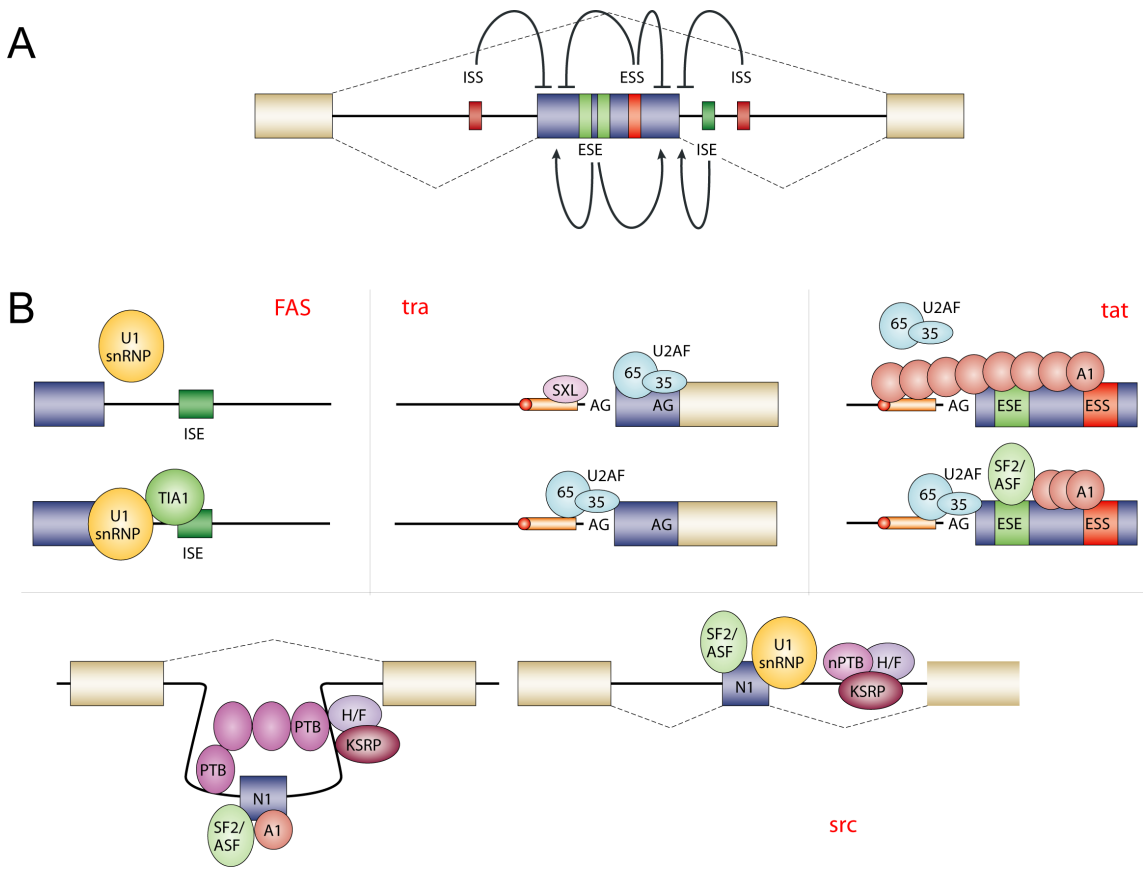

Figure 1.8: Regulation mechanisms of alternative splicing

A - Alternative splicing regulatory elements: exonic/intronic splicing enhancers/silencers (ESE, ESS, ISE, ISS). Enhancers can activate splice sites or repress silencers, silencers can repress splice sites or enhancers. The splicing pattern is determined by the balance between the two competing activities. **B** - Examples of mechanisms for splicing silencing and enhancing: for the *FAS* gene, a weak 5'ss is enhanced by TIA1 binding to a downstream ISE and promoting the interaction of U1 snRNP with the splice site; in *Drosophila*, for the *tra* gene, repression of the non-sex-specific 3'ss and selection of the downstream female-specific 3'ss involves the interaction between SXL and an ISS located in the poly-Y tract; in HIV1, the inclusion of exon 3 of *tat* depends on the relative abundance of hnRNP A1 and SF2/ASF, as the multimerization of the first from an ESS can be blocked by the binding of the second to an upstream ESE; for the *src* gene, splicing of exon N1 is regulated by a combination of antagonistic and cooperative interactions involving both enhancing and inhibiting factors (SF2/ASF, hnRNP A1, PTB, nPTB - neuronal PTB, hnRNP H/F, KSRP - KH-type splicing regulatory protein). (Adapted from [Matlin et al., 2005].)

the appearance of vertebrates, shaping their genome, and precede the emergence of mammals. Indeed, it has been widely observed that mammals benefit from four copies of genes that appear as an unique copy in invertebrates. Mammals possess, in general, four or less paralogous [9] gene clusters [Aparicio, 2000; Vandepoele et al., 2004].

The analysis of *Hox* genes and *Hox* gene clusters (important in development, known to be involved in the patterning of the anterior-posterior axis of vertebrate and invertebrate embryos) illustrates these findings. It was observed that protostome invertebrates and Amphioxus (deuterostome cephalochordate, an extant sister group to vertebrates) have a single *Hox* cluster, whereas the lobe-finned fish, amphibians, reptiles, birds, and mammals posses four clusters. These findings support two rounds of entire-genome duplication early in vertebrate evolution but it has been suggested that a first duplication occurred after the divergence of the cephalochordates, and a second one occurred after the divergence of the jawless vertebrates (Figure 1.9) [Aparicio, 2000; Vandepoele et al., 2004; Garcia-Fernandez and Holland, 1996; Holland et al., 1994; Holland, 1997].

Mapping of *Hox* genes in the teleosts *Fugu rubripes* (Japanese pufferfish) [Aparicio et al., 2002], *Spheroides nephelus* (Southern pufferfish) and *Danio rerio* (zebrafish) revealed extra sets of genes and suggests a lineage-specific genome duplication [Aparicio, 2000; Amores et al., 1998; Amores et al., 2004] (Figure 1.9). Indeed, the three species exhibit seven *Hox* compexes. Taken the four tetrapod counterpart has references, the pufferfish possesses two copies of *Hox B* and *Hox D* clusters, a single *Hox C* cluster and at least two *Hox A* clusters. Zebrafish has two copies of *Hox A*, *Hox B* and *Hox C* clusters and a single *Hox D*. These findings support genome duplication before divergence of zebrafish and pufferfish lineages, followed by differential loss of a *Hox C* cluster in the pufferfish lineage of a *Hox D* cluster in the zebrafish lineage.

Moreover, recent phylogenetic studies on *Fugu* duplicates and paralogons support

[9] In this case the copies are called paralogous because they arise from a duplication in an ancestor, followed by speciation. In general, homologous sequences are orthologous if they were separated by a speciation event: if a gene exists in a species, and that species diverges into two species, then the copies of this gene in the resulting species are orthologous. Homologous sequences are paralogous if they were separated by a gene duplication event: if a gene in an organism is duplicated, then the two copies are paralogous.

a fish-specific whole-genome duplication, early during the radiation of modern ray-finned fishes, probably before the origin of teleosts (around 350 million years ago) [Christoffels et al., 2004; Vandepoele et al., 2004].

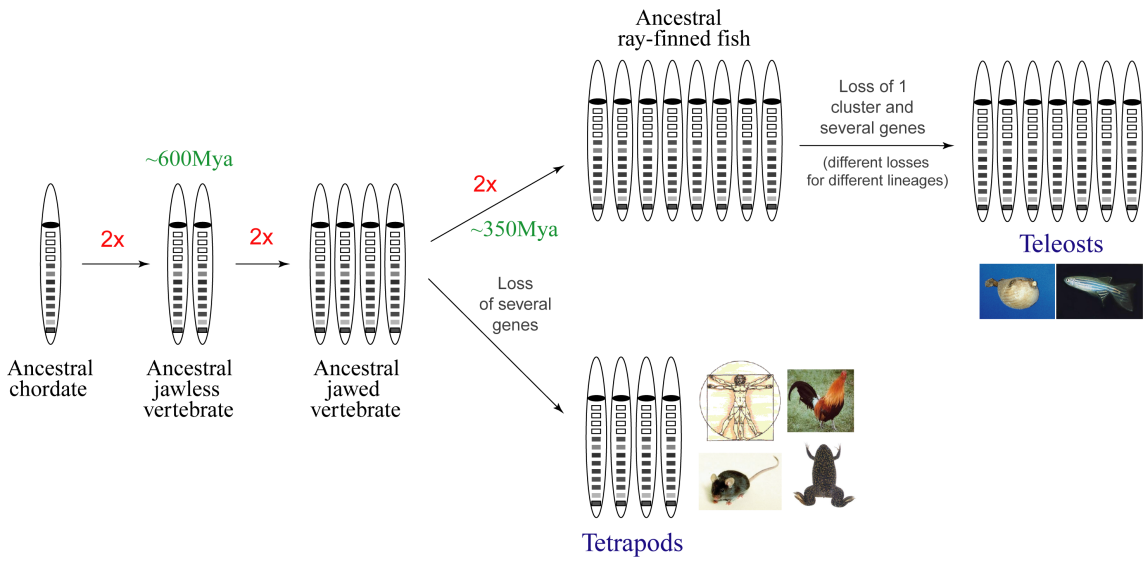

Figure 1.9: Vertebrate genome evolution

The ancestral chordate genome (represented by a schematic chromosome) underwent at least two rounds of duplications (2×), estimated to have occurred around 600 million years ago, to generate the ancestral jawed vertebrate genome. This produced a fourfold increase in the number of the *Hox* complex genes (represented by a series of boxes in the chromosome). Tetrapods then lost some genes during evolution but ray-finned fishes underwent an additional round of genome duplication (around 350 million years ago). *Fugu* and zebrafish lineages subsequently suffered different gene losses. (Adapted from [Aparicio, 2000].)

1.2.2 Fate of Gene Duplications

To be retained, duplicates must become selective advantageous before deleterious mutations disrupt their functionality. On one hand, coding sequences can undergo mutations that, by altering the resulting protein function, provide a selective advantage. On the other hand, expression patterns can be changed by mutations in regulatory sequences [Aparicio, 2000].

The classical model for the retention of duplicates states that initially the two

copies are redundant and, if dosage is not critical, one shields the other from natural selection [Ohno, 1970]. As degenerative mutations are more frequent, in most cases one of the copies should become non-functional and result in a pseudogene. The preservation of duplicates would then result from the fixation of rare advantageous mutations that would provide a new function to one of the copies, while the other copy retained the original function.

The rapid and frequent nonfunctionalization predicted by the classical model is contradicted by observations in species that underwent polyplidization. Some lineages of fish and frogs preserved significantly more genes than could be expected from the classical model and it is suggested that dosage effects are not responsible for this high retention. Additionally, nucleotide substitution patterns in *Xenopus laevis* indicate purifying selection of both copies. Finally, the number of null alleles segregating in extant species for loci that have avoided nonfunctionalization in both copies is relatively low. The classical model also underestimates the complexity of gene structure. Genes are very often multi-functional and their expression depends on different regulatory elements. These are often modular, independent and associated with distinct protein coding domains. Models for the evolution of gene duplicates must therefore consider the partitioning of a gene's function [Force et al., 1999].

The duplication-degeneration-complementation (DDC) model states that degenerative mutations facilitate the preservation of duplicate functional genes as the dominant mechanism of duplicate retention is the partitioning of the ancestor's original functions (subfunctionalization) rather than the evolution of new functions (neofunctionalization) [Force et al., 1999]. The DDC process is described and illustrated in Figure 1.10. The DDC model predicts a probability for subfunctionalization that is more consistent with observation, when compared with the classical model.

The DDC model accounts for increases in gene number and for the observed cases of subfunctionalization but it does not address the development of new biochemical capabilities, as it considers the gene duplication to be neutral. This gap led to the proposal of a new model that assumes a period of natural selection for the duplication itself and bursts of adaptive gene amplification as a response to selective environmental pressures [Francino, 2005].

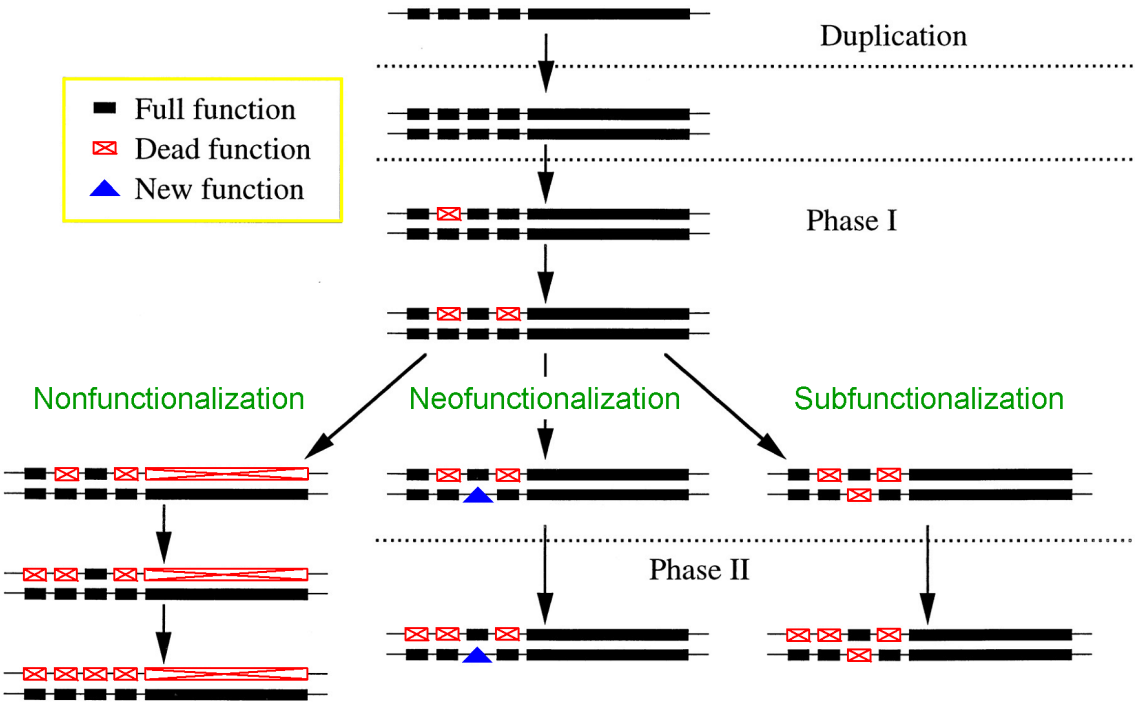

Figure 1.10: The fate of gene duplications

Three possible fates of duplicates with multiple regulatory genes are represented. Small boxes depict regulatory elements with unique function and large boxes represent transcribed regions. In the first two represented steps after duplication (Phase I), one of the copies (the upper one) acquire fixed null mutations in two regulatory regions. If the next mutation in the same copy abolishes the expression of a functional protein, that copy becomes a nonfunctional pseudogene, tending to accumulate random mutations (nonfunctionalization). Alternatively, if the third mutation affects a regulatory region in the lower copy that is preserved in the upper copy, both copies are needed for complete functional expression of the gene (subfunctionalization) and are therefore prevented from nonfunctionalization. The fourth regulatory element may still get mutated in any of the copies and the mutation may provide the gene with a new advantageous function (neofunctionalization). (Adapted from [Force et al., 1999].)

The new model finds support in the observed amplification mutagenesis phenomenon. The expansion of a defective gene can be selective advantageous when the higher levels of expression (due to existence of many copies) can compensate for the defect. An increase in the copy number directly leads to an increase in the mutation frequency. Many cases of adaptation by gene amplification have been reported in cell lines from bacteria, yeasts, insects and mammals.

The so-called adaptive radiation model postulates that neofunctionalization takes place in rapid and punctuated bursts with the emergence of new biochemical niches. Large and selected amplifications of the best preadapted genes are followed by competition involving the gene copies in the population for the filling of the new niche. Initially, if a new molecular function brings benefit to the organism's fitness, the niche can be occupied by a suboptimal protein. The amplification of the corresponding gene is selected as an increase in expression can compensate for the incomplete functionality. The adaptive radiation model is illustrated in Figure 1.11.

In summary, the adaptive radiation model addresses the main inconsistencies of the previous models for neofunctionalization. It postulates that gene amplification, followed by the evolution of new gene function, is beneficial *per se*. It assumes that the increase on the probability of advantageous mutations results from a greater number of targets and that this abundance of gene copies allows for the 'exploration' of the adaptive landscape and an easier convergence to fitness 'valleys'. Finally, the model considers that beneficial mutations can result from recombination among gene copies and, at every step, alternate with rounds of gene amplification.

The adaptive radiation model makes predictions that can be tested by the analysis of genomic data an indeed it is claimed that literature exhibits evidence for the predicted patterns [Francino, 2005]. The reported prevalence of duplications in functional and species-specific classes of genes seems to validate the assumption of punctuated bursts of amplification and fixation of duplicates as a response to selective pressures. Moreover, genome-wide analyses in several species show no evidence for long periods of neutral evolution and cases of positive selection after duplication are reported. These findings are consistent with the postulated early selection of paralogues. Genome-wide analyses also show an excess of pseudogene formation associated with the establish-

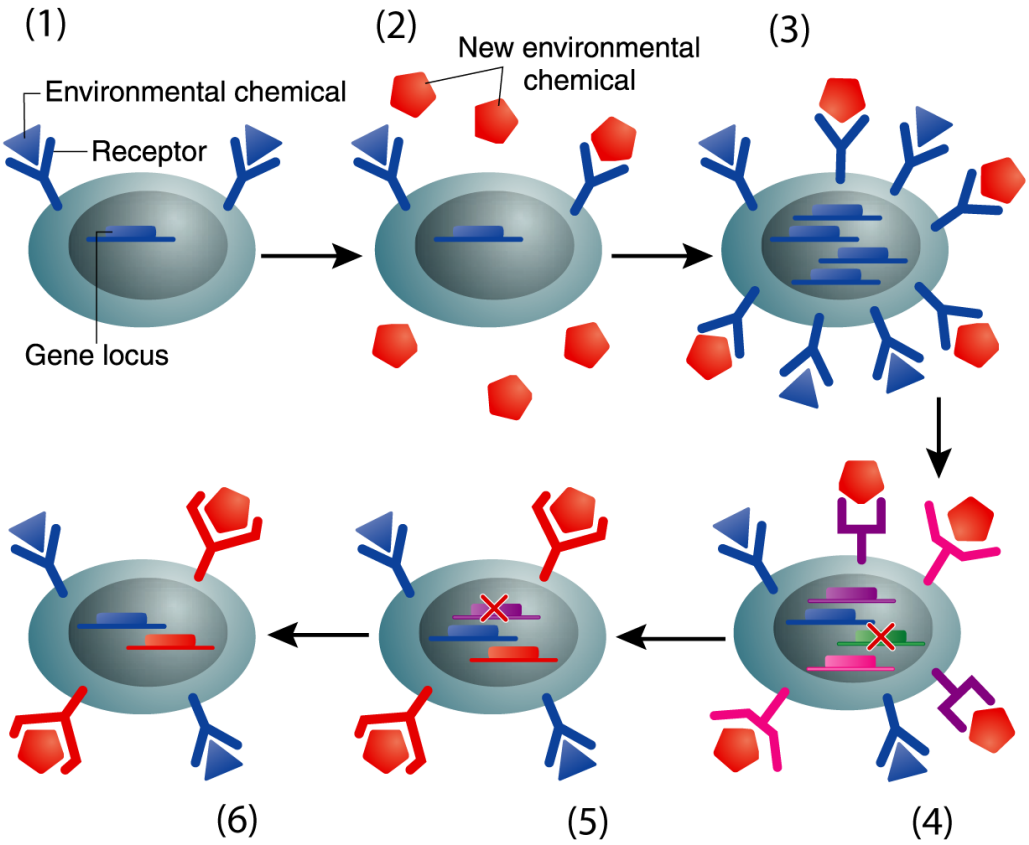

Figure 1.11: Adaptive radiation model

A gene (blue) encodes a receptor for an environmental chemical **(1)** when a new chemical (red) is introduced and the receptor is preadapted to bind it with very low affinity **(2)**. Amplification of the original gene brings a selective advantage because it allows relevant levels of binding of the new chemical, as long as it is not disadvantageous to bind the original chemical in excess **(3)**. In subsequent generations, different copies of the gene acquire different mutations that, for the fittest genotype, provide some paralogues (purple and pink) with intermediate binding affinity for the new chemical and prove deleterious for others (green), turning them into pseudogenes **(4)**. In future generations, old pseudogenes are lost, new pseudogenes (purple) emerge **(5)** and finally the optimal genotype is eventually reached with one gene encoding the original receptor, another gene encoding a receptor for the new chemical and the loss of all pseudogenes and copies encoding receptors with low binding affinity for the environmental chemicals **(6)**. (Adapted from [Francino, 2005].)

ment of new gene functions, as predicted by the model. In contrast with processed pseudogenes (described in 1.2.5), pseudogenes resulting from gene duplications tend to be associated with environmental response genes.

1.2.3 Gene Duplication and Alternative Splicing

Gene amplification provides complexity and specificity to a species or lineage by introducing extra diversity in a proteome and therefore refining and extending the range of functions performable by certain families of genes. The same can be stated on alternative splicing, responsible for the exponentiation of the coding potential of a genome. The two mechanisms provide the redundancy that is necessary for the development of new gene functions by the production of new isoforms.

Although gene duplication and alternative splicing are distinct evolutionary mechanisms, the analogy on the functional consequences of both phenomena raises the question: are the two mechanisms complementary and/or alternative in the 'fine graining' of a gene family? The comparison between examples of subfunctionalization resulting from teleost-specific duplications and the alternative splicing patterns of the corresponding human orthologues suggest interchangeability between both phenomena.

The gene encoding the paired box protein Pax6, present in both vertebrates and invertebrates and known to be expressed in the developing eye and in the central nervous system, has two copies in zebrafish [Nornes et al., 1998]. No evidence for a second *Pax6* gene was found in chicken, mouse or human and our phylogenetic analysis clearly suggests the two copies in zebrafish arose from a lineage-specific duplication. Interestingly the human *PAX6* gene is known to be regulated by alternative splicing and to encode two different functional proteins [Epstein et al., 1994]. We analyzed the gene structure of human *PAX6* and zebrafish *pax6a* and *pax6b* genes and observed that each of the zebrafish genes encodes a protein resembling one human isoform (Figure 1.12A). Moreover, experimental data described in the literature suggests homology between those two processes of subfunctionalization [Nornes et al., 1998; Epstein et al., 1994].

This type of resemblance has also been shown for the gene encoding the micro-

phthalmia-associated transcription factor (*Mitf*) [Lister et al., 2001]. The proteins encoded by the two zebrafish *mitf* genes look homologous to distinct isoforms generated by alternative splicing of the single mammalian *Mitf* gene (Figure 1.12B), suggesting specialization of the two zebrafish genes following a duplication event.

Another example involves the genes encoding synapsin (*Syn*) and the tissue inhibitor of metalloproteinase (*Timp*) [Yu et al., 2003]. They exhibit a nested organization that is conserved in *Drosophila* and vertebrates. Analysis of the human and *Fugu* genomes show that the evolution of *Syn-Timp* gene families is characterized by duplications, secondary loss and the partitioning of ancestral functions (subfunctionalization). There are two duplicate *Syn-Timp* loci in *Fugu* that have evolved in a way such that each *Syn* duplicate produces one of the two transcripts generated from the single ancestral gene, and one of the *Timp* genes is lost (Figure 1.12C).

We have shown that a similar mechanism of subfunctionalization is likely to have occurred in the two separate *Fugu U2AF*^{35} genes, leading to the degeneration of alternative exons [Pacheco et al., 2004]. In mammals there are two known functional isoforms (the constitutive and the so-called U2AF^{35}b), with same length, that differ from each other only in 7 amino-acids located at one RRM, associated to alternative exons 3. Both copies in *Fugu* resemble the human gene structure but the 3rd exon of one copy is homologue of the human constitutive exon 3 whereas the 3rd exon of the other is homologue to the human alternative exon 3, the so-called exon Ab (Figure 1.13). We were unable to detect either exon Ab sequence within the intron upstream of exon 3 in the *Fugu U2AF*^{35}*a* gene or exon 3 sequence within the intron downstream of exon Ab in the *Fugu U2AF*^{35}*b* gene. It is proposed that post-transcriptional regulation of *U2AF*^{35} gene expression may provide a mechanism by which the relative cellular concentration and availability of U2AF^{35} protein isoforms are modulated, thus contributing to the finely tuned control of splicing events in different tissues. The evolutionary selective pressure on both U2AF^{35}a and U2AF^{35}b isoforms and the observed tissue specificity of their expressions in mammals suggest that each protein has essential functions for vertebrates.

All these examples lead to the suggestion that this phenomenon could be generalized and that many teleost duplications may have been under the evolutionary forces

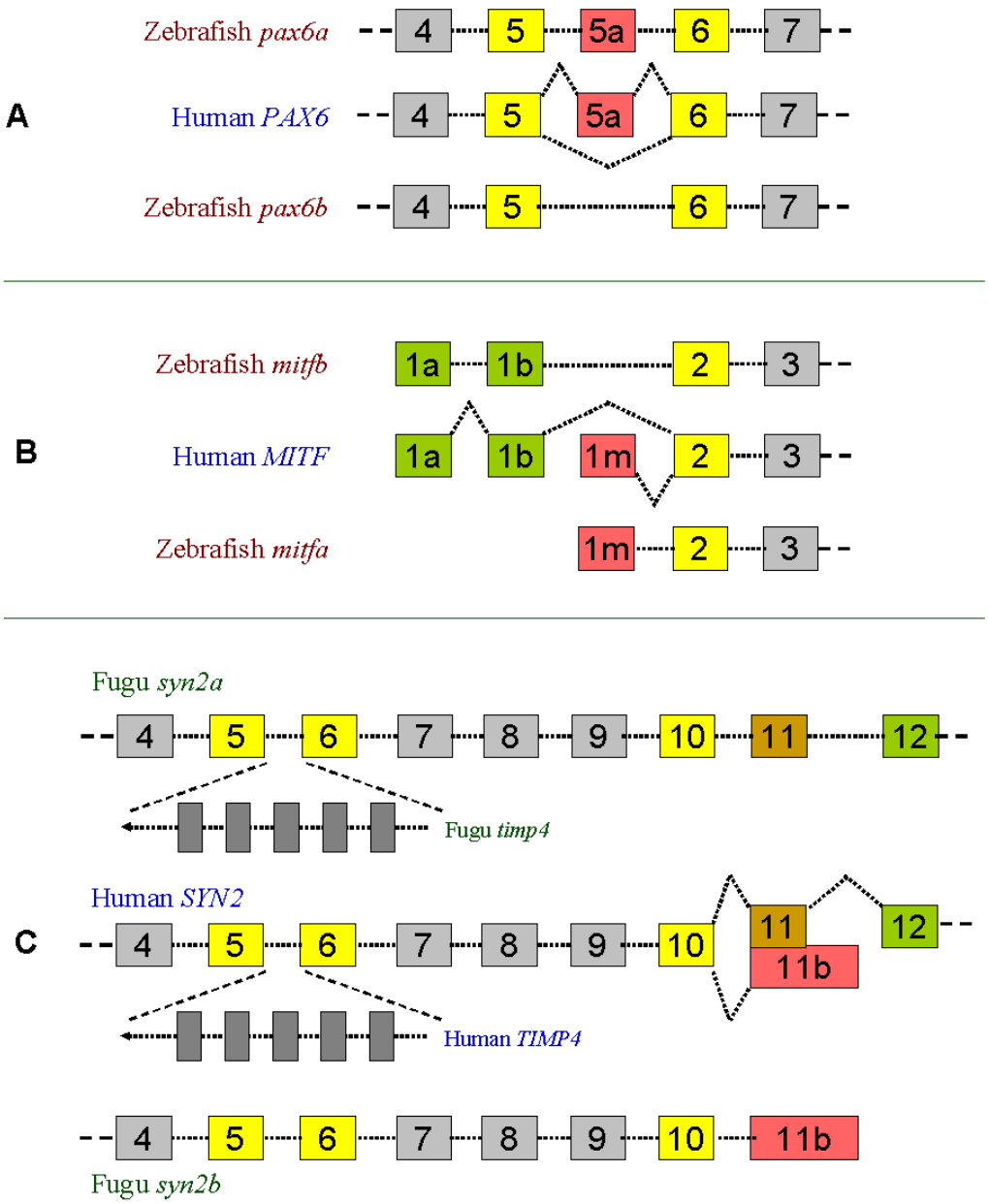

Figure 1.12: Examples of duplication / alternative splicing resemblance
Gene structure of (**A**) *PAX6*, (**B**) *MITF* and (**C**) *SYN2* orthologues.

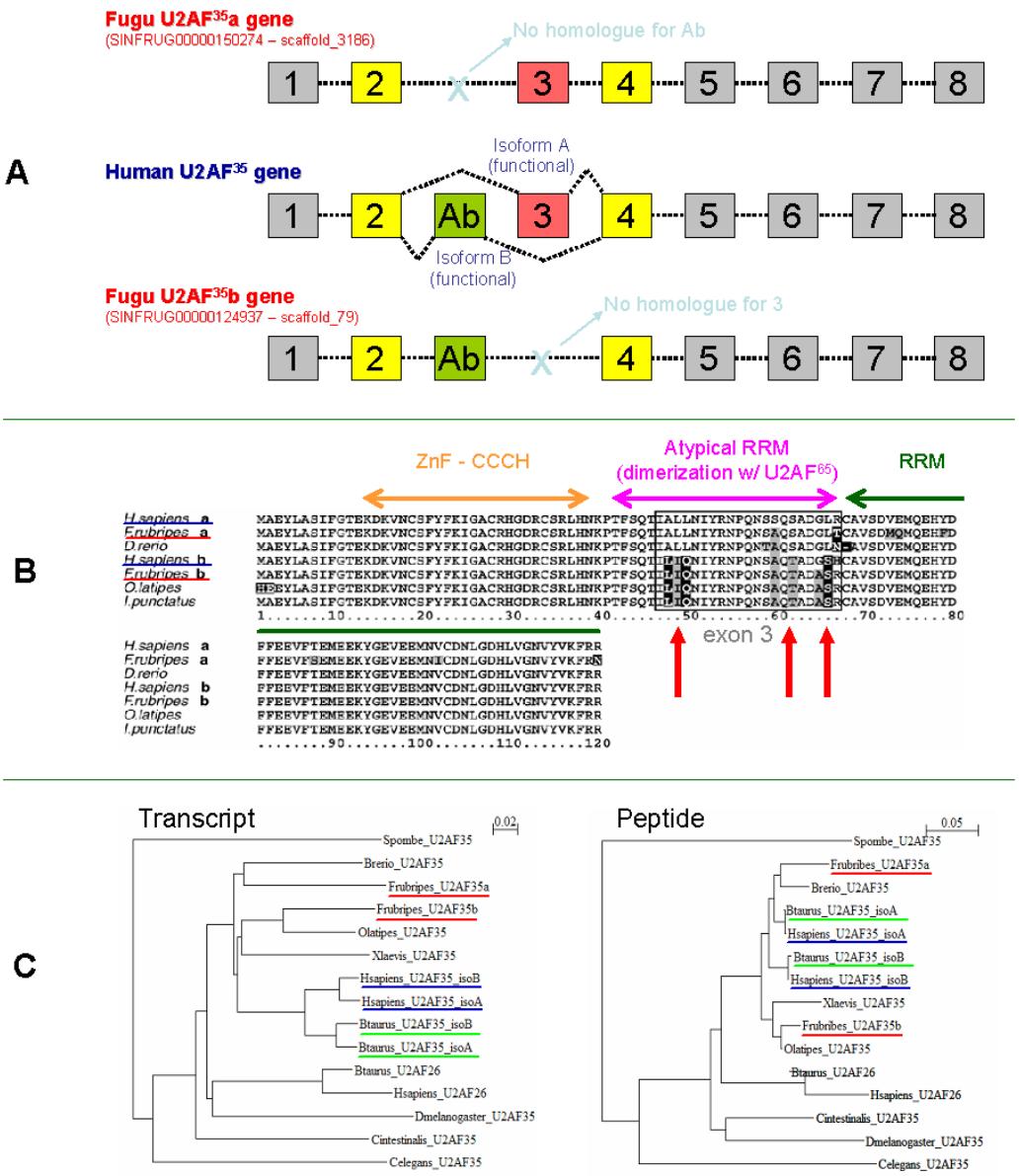

Figure 1.13: Human *U2AF*^{35} and orthologues

A - structure of Human and *Fugu U2AF*^{35} genes; **B** - alignment of Human and Fish *U2AF*^{35} orthologues (a.a. 1 - 120) (adapted from [Pacheco et al., 2004]); phylogeny of Metazoan *U2AF*^{35} orthologues.

that selected the correspondent alternative isoforms of the orthologous mammalian gene.

Interestingly, a recent study in human and mouse show an inverse correlation between the size of a gene's family and its use of alternatively spliced isoforms [Kopelman et al., 2005]. A cross-organism analysis suggests that selection for genome-wide genic proliferation might be interchangeably met by either evolutionary mechanism. This study also suggests that there is a trend for singletons to acquire splice variants rather than for duplicates to loose them. Nevertheless, duplicates generated by subfunctionalization are likely to partition original splice variants.

1.2.4 Segmental Duplications

Polyploidization is not the only duplication mechanism by which a genome can be expanded. Phenomena of tandem and interspersed duplications of chromosomic portions can also be responsible for additional genomic complexity.

Duplications can appear as tandem copies of genes. For instance, *Drosophila* and *C. elegans* possess hundreds of gene pairs are suggested to result from tandem duplications. The Hox complex is another example of a cluster generated by tandem duplications. The human genome contains clusters of odorant receptor genes (some clusters comprise more than a dozen genes) created in the same way [Patel and Prince, 2000].

The shape of mammalian genomes had a significant contribution from recent segmental duplications. The human genome is particularly abundant in blocks of genomic sequence, variable in size, that share a high degree of identity (>90%) [Eichler, 2001]. These blocks are characterized by both exonic and intronic sequences and can be interspersed both within a chromosome or throughout the genome.

Tandem duplications are theoretically expected to be a continuous process during evolution and the number of retained gene duplicates is supposed to undergo an exponential decay over time [Gu et al., 2002; Lynch and Conery, 2000]. This leads to the prediction that vertebrate genomes have a relatively high number of recently duplicated genes. A recent study shows that this trend can be observed in the human genome but it is absent in the *Fugu* genome [Vandepoele et al., 2004]. The number of

tandem duplications is shown to be 7-fold higher in human than in *Fugu*. The *Fugu* genome is known for its extreme tendency for compaction [Aparicio et al., 2002] and it has been shown that *Tetraodontideae* (including *Fugu*) have undergone a major genome contraction in the past 50-70 million years, probably due to a reduction of large insertions and a higher rate of deletions (that might have been responsible for the fast removal of redundant duplicates in *Fugu*) [Neafsey and Palumbi, 2003]. Moreover, an increased rate of segmental and tandem duplications in primate genomes has been reported [Eichler, 2001].

1.2.5 Retrotransposons

Transposable elements contribute for the evolution of a genome by providing both novel regulatory elements and coding sequences. They appear to be present in all eukaryotic genomes and are particularly abundant in mammals, accounting for at least 45% of the human genome [Jordan et al., 2003]. It is believed that mobile elements [10] might have played a very important role in early genome formation, as it is now widely accepted that the origins of life are in an "RNA world" followed by reverse transcription into DNA [Kazazian, 2004].

DNA transposons are mobile elements prevalent in bacteria (although found in metazoa) that are simply excised from a genomic site and integrated into another. In mammals, the dominant mobile elements are retrotransposons: RNA sequences that are reverse transcribed into DNA and reintegrated into the genome. LTR retrotransposons are characterized by long terminal repeats at both ends and are very similar to retroviruses. Among non-LTR retrotransposons, the most common in mammals are the LINE-1 (long interspersed nucleotide elements 1) or L1 elements [Kazazian, 2004]. In human, they constitute 17% of the genome [Ostertag and Kazazian, 2001]. The mechanisms of reverse transcription and generation of both LTR and non-LTR retrotransposons are illustrated and summarized in Figure 1.14.

Most retrotransposons are pseudogenes and retrotransposition accounts for the majority of the mammalian pseudogenes (retropseudogenes) [Zhang et al., 2004].

[10] Defined as DNA sequences that are able to integrate into the genome at a new site within their original cell.

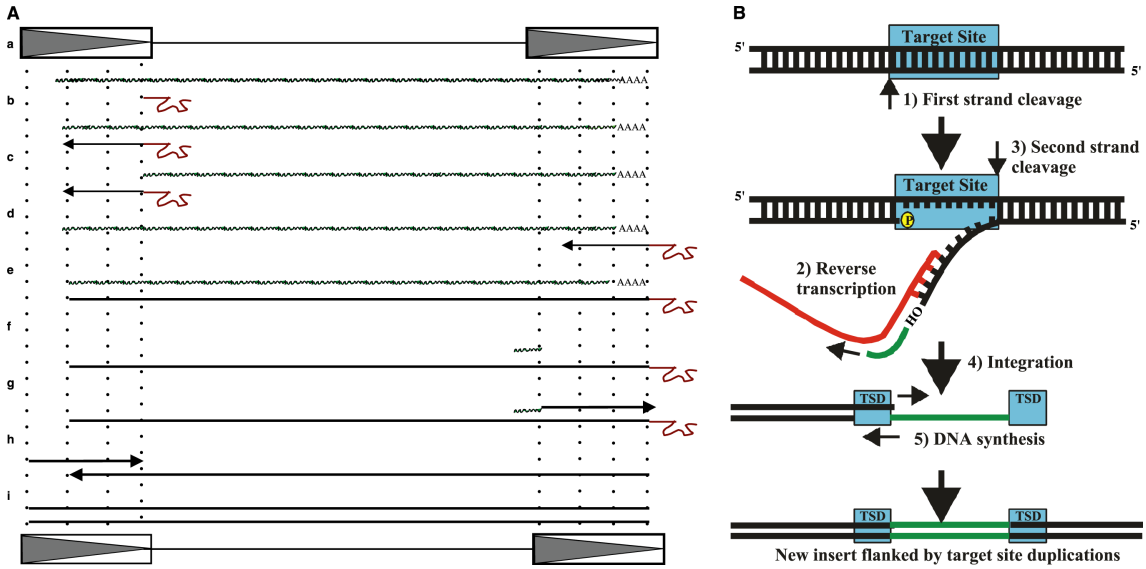

Figure 1.14: Mechanisms of reverse transcription

A - Reverse transcription of LTR retrotransposons and retroviruses: the region near the 5' end of the RNA is copied into DNA using a tRNA primer (**a** and **b**); the 5' region of the RNA is degraded (**c**), the newly synthesized DNA jumps to the 3' end of the RNA (**d**) and the synthesis of the first strand is completed (**e**); the element-encoded RNAse H degrades most of the RNA (**f**) and the short remaining RNA act as primer of the synthesis of the right end of the second DNA strand, using the first DNA strand as template (**g**); finally another jump of second-strand DNA to the left end of the DNA (**h**) is followed by completion of second-strand synthesis (**i**). **B** - Reverse transcription of non-LTR retrotransposons: the endonuclease nicks the bottom strand of DNA, leaving a 5'-PO_4 and a 3'-OH, which serves as a primer with the element RNA (R1, R2, L1, etc) as template for the RT; the second strand of DNA is cleaved during reverse transcription of the first strand and the 3'-OH of the second strand becomes a second primer for reverse transcription internally on L1 RNA; resolution of this second cDNA produces the inversion. (Adapted from [Kazazian, 2004].)

However, sometimes the retrotransposition of the mRNA of functional genes generates putative functional genes if a functional promoter and other regulatory elements are present and make expression possible. Due to its 'trancriptomic' nature, a retrotransposed gene is intronless and does not undergo alternative splicing but it can still represent an extra isoform for the respective gene family and introduce further functional complexity. The majority of retrotransposed pseudogenes (and putative genes) in human and mouse appear to be lineage specific and therefore very recent [Zhang et al., 2004]. As a consequence, some pseudogenes might be considered putative genes as there was not enough time for the accumulation of disruptive mutations and the open reading frames are preserved. Moreover, the expression of retrotransposed genes might be hard to detect as they can be confounded with their paralogues due to the high sequence similarity (and sometimes total identity). Examples of retrotransposed genes and the effects of their expression are discussed in 2.3.4.

1.3 Bioinformatics tools on the study of Gene Expression and Evolution

Bioinformatics can be defined as the application of computational tools and techniques to the management and analysis of biological data [Tisdall, 2001]. Computational analysis has become an integral part of research in biology. Several software tools perform different kinds of data analysis but it is challenging to automatically integrate data and results from multiple sources. Bioinformatics aims to achieve this integration by writing program logic to read and write data specific to the biological domain.

1.3.1 Sequence annotation

Computational genomics aims to understand and interpret information encoded and expressed from a genetic complement of organisms. A genome sequence provides a natural framework for the organization of biological data. The volume and diversity of genomic sequence in the public databases has been rapidly expanding. Biologists need tools that can help them search, view, organize and retrieve that public data.

Annotation has therefore become an important element in data analysis and interpretation.

Genome databases, such as Ensembl [11] [Hubbard et al., 2002] and the UCSC Genome Browser Database [12] [Karolchik et al., 2003], are invaluable resources to scientists and have been improving dramatically. They provide access to genome assemblies of several organisms, which are accompanied by large collections of annotation data: mRNA and EST alignments, gene predictions, cross-species homologies, single nucleotide polymorphisms (SNPs), etc. Moreover, they integrate and cross-link the annotation of other reliable high-quality and comprehensive transcriptomic and proteomic databases, such as GenBank [Wheeler et al., 2003], SwissProt [Boeckmann et al., 2003] or EMBL [Kanz et al., 2005].

Large scale computational analysis of sequence data requires not only access to databases (which should provide the data in a simple standard 'computer-friendly' format) but also software tools for automated sequence extraction, manipulation and further annotation.

Perl

Perl is one of the most widely used programming languages for biological data integration, performing analysis and combining from multiple sources. Perl is proved to be very useful for connecting software applications together into sequence analysis pipelines, converting file formats and extracting information from the output of analysis programs and other text files [Stajich et al., 2002].

Perl has particular features that make common bioinformatics tasks easier. It deals well with ASCII [13] text files or flat files, in which much important biological data appears (GenBank [Wheeler et al., 2003] and PDB [Deshpande et al., 2005] databases, for example). Processing and manipulation of long sequences, such as DNA

[11] `http://www.ensembl.org`

[12] `http://genome.ucsc.edu`

[13] ASCII stands for American Standard Code for Information Interchange and is a character encoding based on the English alphabet. ASCII codes represent text in computers, communications equipment and other devices that work with text. Most modern character encodings have a historical basis in ASCII.

and proteins, is made easy with Perl. It provides efficient support for text processing and pattern matching tasks. Perl also makes it convenient to write a program that controls one or more other programs and is very useful in the generation of dynamic web sites.

Rapid prototyping (i.e. the speed with which a programmer can write a typical program) is another benefit of using Perl. Many problems can be solved in far fewer lines of Perl code than in C or Java. Perl can be considered a portable language, as it runs on most operating systems (Windows, Mac, Linux). The speed with which Perl programs run is good, although speed of execution is not the main attribute of Perl (for instance, C is faster) [Tisdall, 2001].

Bioperl

Much of the Perl software in bioinformatics used to be written for immediate 'domestic' utility rather than reusability and redundant software was inefficiently rewritten several times. To avoid that, the Bioperl (`http://www.bioperl.org`) toolkit has been written to bring together reusable Perl modules containing generalized routines specific to life-science information [Stajich et al., 2002]. The code has been made freely available, under an open-source license, so that anybody in the scientific community can contribute to Bioperl.

Bioperl is built in an object-oriented [14] manner so that many modules depend on each other to achieve a task. This is because it was realized that, first, even though file formats of distinct analysis programs are different, the information they represent is the same. Second, the number of data structures needed to represent information flow is limited (and common to most applications such as sequences, annotation, features and alignments), which allows for a small set of modules to be reused for a variety of purposes. Third, a set of operations (like reading and writing information to a file,

[14] Object-oriented programming is the practice of grouping related tasks together into logical and broadly applicable components. It is a type of programming in which programmers define not only the data type of a data structure, but also the types of operations (functions) that can be applied to the data structure. In this way, the data structure becomes an object that includes both data and functions. In addition, programmers can create relationships between one object and another. For example, objects can inherit characteristics from other objects.

querying a sequence for its features and translating a coding sequence into protein) is commonly performed on these data structures.

Bioperl is primarily focused on sequence manipulation. It supports access to remote databases (such as GenBank [Wheeler et al., 2003], SwissProt [Boeckmann et al., 2003] or EMBL [Kanz et al., 2005]) for sequence data retrieval and comprises modules for transforming/converting formats of sequence files. Bioperl provides various helper objects to obtain basic sequence statistics, to identify restriction enzyme and amino acid cleavage sites, to manipulate sequence alignments, etc. It also offers numerous tools for the development of machine readable sequence annotations.

Bioperl is branched out into sequence-related fields of study, such as protein structure, phylogenetic trees and genetic maps. It also includes objects conceived to query bibliographic databases and to represent sequence (and respective features) objects graphically.

Besides the pure Perl solutions, Bioperl can take advantage of external data analysis packages. It is capable of parsing the output from a variety of programs including BLAST [Altschul et al., 1990], HMMer [Eddy, 1998], ClustalW [Thompson et al., 1994], T-Coffee [Notredame et al., 2000], Phylip [Felsenstein, 1989], many EMBOSS [Rice et al., 2000] programs, Genscan [Burge and Karlin, 1997] and many others. Moreover it can launch remote analyses using the EMBOSS suite, BLAST and the multiple sequence alignment programs ClustalW and T-Coffee.

Ensembl Perl API

The most advanced use of the Bioperl toolkit has come through the Ensembl project [Hubbard et al., 2002]. The basic sequence handling, file format parsing, and sequence features for annotation model have been used in the automatic annotating of genomes.

Ensembl stores these data in several MySQL [15] databases. A comprehensive Perl Application Programme Interface (API) was developed to provide efficient access to tables within the Ensembl databases. By encapsulating the underlying database structure, the libraries present end users with a simple, abstract interface to a complex

[15] MySQL is an open source relational database management system that uses Structured Query Language (SQL), the most popular language for adding, accessing, and processing data in a database.

data model [Stabenau et al., 2004].

1.3.2 Sequence search and pairwise alignment

One of the simplest tasks in sequence analysis is to assess the homology between sequences. This can be achieved by aligning the sequences and then evaluating if the alignment is obtained because sequences are related or just by chance. Performing sequence alignments involves decisions on selecting the sort of alignment to consider, the scoring system, the algorithm and the statistical methods to assess significance [Durbin et al., 1998].

In a sequence alignment, two or more sequences are arranged in a way that highlights their similarity (Figure 1.15 shows an example of a pairwise alignment). Sequences can be padded with gaps so that, where possible, columns contain identical or similar characters from the aligned sequences. When used to study the evolution of the sequences from a common ancestor, mismatches in the alignment correspond to mutations and gaps correspond to insertions or deletions.

```
 Score =  376 bits (966),  Expect = 5e-103
 Identities = 186/243 (76%), Positives = 201/243 (82%), Gaps = 9/243 (3%)

Query  1    MAEYLASIFGTEKDKVNCSFYFKIGACRHGDRCSRLHNKPTFSQTIALLNIYRNPQNSSQ  60
            MAEYLASIFGTEKDKVNCSFYFKIGACRHGDRCSR+HNKPTFSQT+ L N+Y NPQNS++
Sbjct  1    MAEYLASIFGTEKDKVNCSFYFKIGACRHGDRCSRIHNKPTFSQTVLLQNLYVNPQNSAK  60

Query  61   SADG--LRCAVSDVEMQEHYDEFFEEVFTEMEEKYGEVEEMNVCDNLGDHLVGNVYVKFR  118
            SADG  L   VSD EMQEHYD FFE+VF E E+KYGE+EEMNVCDNLGDHLVGNVY+KFR
Sbjct  61   SADGSHLVANVSDEEMQEHYDNFFEDVFVECEDKYGEIEEMNVCDNLGDHLVGNVYIKFR  120

Query  119  REEDAEKAVIDLNNRWFNGQPIHAELSPVTDFREACCRQYEMGECTRGGFCNFMHLKPIS  178
             E DAEKA  DLNNRWF G+P+++ELSPVTDFREACCRQYEMGECTR GFCNFMHLKPIS
Sbjct  121  NEADAEKAANDLNNRWFGGRPVYSELSPVTDFREACCRQYEMGECTRSGFCNFMHLKPIS  180

Query  179  RELRRELYGRRRK-KHRSRSRSRER--RSRSRDRGRGGGGGGGGGGGG----RERDRRRS  231
            RELRR LY RRR+ + RSRS  R R  RSRSR  GR GGG G G GGG     ERD  R
Sbjct  181  RELRRYLYSRRRRARSRSRSPGRRRGSRSRSRSPGRRGGGRGDGVGGGNYLNNERDNMRG  240

Query  232  RDR  234
             DR
Sbjct  241  NDR  243
```

Figure 1.15: Example of BLAST output

Protein sequences of human U2AF[35] (Query) and its orthologue in Drosophila U2AF[38] (Sbjct) are aligned using BLAST [Altschul et al., 1990]. The central lines indicate identical positions with letters and similar (based on their chemical/structural properties) residues with "+".

Sequence alignment can involve the construction of an alignment of given sequences or finding significant alignments in a database of potentially unrelated sequences.

Pairwise sequence alignment methods were developed to find best-matching local or global alignments of two amino acid or nucleotide sequences, aiming to identify homologues of a gene or its product in a database. This is very useful in evolutionary studies and in the identification of sequences of unknown structure or function, for example.

Global alignments involve all the characters in both sequences and are used in finding closely-related sequences. Local alignment methods are more flexible, as they find related regions within sequences and therefore related regions which appear in a different order in the two proteins (domain shuffling) can be related.

BLAST

BLAST (Basic Local Alignment Search Tool) [Altschul et al., 1990] is the best known and most widely used heuristic algorithm for local sequence alignment. It is provided with programs for finding high scoring local alignments between a query sequence and a target database, both of which can be either DNA or protein. BLAST is based on the principle that true match alignments are likely to contain a short stretch of identities (or, at least, very high scoring matches). Such short stretches are used as 'seeds' from which the search of longer alignment is extended. Short seed segments allow the pre-processing of the query sequence and the subsequent generation of a table of all the possible seeds with their corresponding start points.

BLAST lists all the 'neighborhood words' of a fixed length [16] that would match the query sequence somewhere with a score higher than a given threshold. It then scans for words through the database. Whenever it finds one, it starts a 'hit extension' process to extend the possible match as an ungapped alignment in both directions, stopping at the maximum scoring extension [17].

The most common implementation of BLAST finds only ungapped alignments but

[16] Default BLAST word length: 3 for protein sequences, 11 for nucleic acids.

[17] There is actually a small chance that it will stop short of the true maximal extension

it misses only a small proportion of significant matches: the expected best score of unrelated sequences drops, so partial ungapped scores can still be significant; BLAST can find and report more than one high scoring match per sequence pair and can give significance values for combined scores. Nevertheless, there are versions of BLAST that provide gapped alignments [Altschul et al., 1997; Durbin et al., 1998].

Figure 1.15 illustrates the output of a pairwise BLAST alignment of two proteins.

HMMs

Biological sequences can be classified into functional and/or structural families. Multiple alignments (see subsection 1.3.3) of a family reveal the pattern of conservation of sequences and show that different residues undergo different selective pressures. "Profile" methods have been developed to include position-specific information from multiple alignments in database search for homologues. Hidden Markov Models (HMMs) provide a consistent theoretical background for such methods [Eddy, 1998].

A Hidden Markov Model can be defined as a finite set of states, each associated with a probability distribution. Transitions among the states are determined by a set of so-called transition probabilities. An observation can be generated for a given state following the corresponding probability distribution. Only the outcome and not the state is visible to an external observer - the states are 'hidden' [18]. Thus an HMM is a statistical model where the system being modeled is assumed to be a Markov process [19] with unknown parameters and one aims to determine the hidden parameters from the observable parameters. Figure 1.16 illustrates a simple HMM application.

[18] In a regular Markov model the state is visible to the observer, and the transition probabilities are therefore the only parameters. In HMMs each state has a probability distribution over the possible outputs. A sequence of tokens generated by an HMM does not directly indicate the sequence of states. Different state sequences can generate the same symbol sequence, with different total probability.

[19] In a stochastic Markov process, a state c_k at time k is one of a finite number in the range $\{1,\ldots,M\}$. Assuming that the process runs only from time 0 to time N and that the initial and final states are known, the state sequence can be represented by a finite vector $C = (c_0,\ldots,c_N)$. Let $P(c_k|c_0, c_1, \ldots, c_{k-1})$ be the probability of occurrence of state c_k at time k, conditioned on all states up to time k-1. A process such that c_k depends only on the previous state c_{k-1} and is independent of all other previous states is called a first-order Markov process: $P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-1})$. For an nth-order Markov process, $P(c_k|c_0, c_1, \ldots, c_{k-1}) = P(c_k|c_{k-n}, \ldots, c_{k-1})$.

A profile HMM can be trained from unaligned sequences, like running a multiple alignment program (see subsection 1.3.3) before actually building the model. Alternatively, an HMM can be built from pre-aligned sequences. The profile HMM building process is fed with an existing multiple alignment, which provides the state paths, so the parameters are estimated by converting observed state transitions into probabilities.

In the HMM architecture for representing profiles of multiple sequence alignments, a "match" state models the distribution of residues in each consensus column of the alignment. The architecture includes "insert" and "delete" states at each column that allow for insertion of residues between that column and the next or for deleting the consensus residue, respectively [Eddy, 1998].

HMMER [20] is a widely used software package that implements profile HMMs for biological sequence analysis. It is primarily used to build database search models from pre-existing alignments (like those in Pfam [Bateman et al., 2002]). Most protein families have a number of strongly conserved key residues, separated by a characteristic spacing and these features are used by HMMER to build the profile HMMs and search sequence databases for members of given families. This approach outperforms BLASTing an individual family member against the database, specially for evolutionarily diverse families. HMMER can also be used in the automated annotation of the

[20] http://hmmer.wustl.edu/

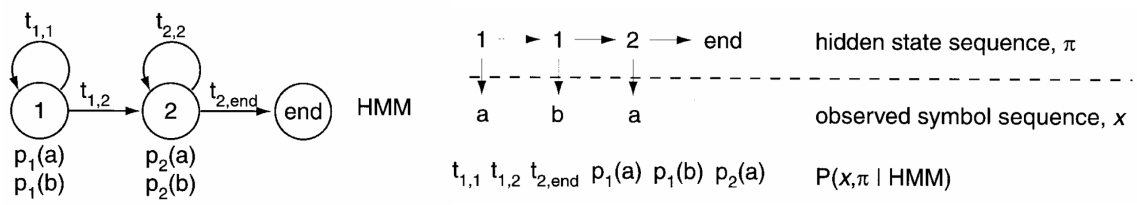

Figure 1.16: Example of simple HMM for sequence modelling

Sequences of as and bs are modelled as two regions of potentially different residue composition. In the model graphical illustration (left) circles represent states and arrows transitions. Generated possible state and symbol sequences are shown (right). $P(x, \pi|HMM)$, the joint probability os symbol and state sequences, is a product of all the transition and emission probabilities. (Adapted from [Eddy, 1998].)

domain structure of proteins, as databases of curated alignments and HMMER models of known domains are available, including Pfam.

1.3.3 Multiple sequence alignment

The simultaneous alignment of many nucleotide or amino acid sequences is one of the commonest tasks in bioinformatics and an essential tool in molecular biology. Multiple alignments are an essential pre-requisite to many further analyses of protein sequences: finding of diagnostic patterns to characterize protein families (illustrating conserved and variable sites within a family), detection and demonstration of homology between new sequences and existing families of sequences and homology modelling in general. They can be used to help predict the secondary and tertiary structures of new sequences or to suggest oligonucleotide primers for PCR. They are also essential in molecular evolutionary analysis, being the basis of phylogenetic reconstruction (section 1.3.4) [Thompson et al., 1994; Notredame et al., 2000].

In a multiple sequence alignment, homologous (in structural and evolutionary senses) residues among a set of sequences are aligned together in columns. A column of aligned residues should occupy similar three-dimensional structural positions and diverge from a common ancestral residue. However it is not possible to unambiguously identify homologous positions and create a single correct multiple alignment (except for highly identical sequences). Protein structures also evolve and two protein structures with different sequences are not entirely superposable. Even when structures diverge, there is a correct evolutionary alignment but such alignment can be more difficult to infer than a structural alignment. A structural alignment has an independent point of reference (superposition of crystal or NMR structures) but the evolutionary history of the residues of a sequence family is not independently known from any source, it must itself be inferred from sequence alignment. The subset of columns corresponding to key residues and core structural elements that can be aligned with more confidence should therefore become fiducial in the alignment procedure.

Automatic multiple sequence alignment methods must turn the biological criteria into a numerical scoring scheme that allows the program to recognize a good align-

ment. The scoring system must take into account the fact that some positions are more conserved than others and the fact that the sequences are not independent, being related by a phylogenetic tree. Ideally a multiple alignment could be scored by specifying a complete probabilistic model of molecular sequence evolution: given the correct phylogenetic tree for the sequences, the probability of a multiple alignment would be the product of the probabilities of all the evolutionary events necessary to produce that alignment via ancestral intermediate sequences with the prior probability of the root ancestral sequence. However there is not enough data to parameterize such a complex evolutionary model [21] and simplifying assumptions (namely approximations that partly or entirely ignore the phylogenetic tree, while doing some sort of position-specific scoring of aligning structurally compatible residues) must be made [Durbin et al., 1998].

Progressive alignments and ClustalW

The most commonly used heuristic approach to multiple sequence alignment is progressive alignment. The idea is to construct a succession of pairwise alignments. First, two sequences are chosen and aligned by standard pairwise alignment and this alignment is fixed. Then, a third sequence is chosen and aligned to the first alignment. This process is iterated until all sequences have been aligned.

The most important heuristic of this methods is to align the most similar pairs of sequences first, as these are the most reliable alignments. Most algorithms build a "guide tree", a binary tree whose leaves represent sequences and whose interior nodes represent alignments. The root node represents a complete multiple alignment and the nodes furthest from the root represent the most similar pairs. Thus the principle is to take an initial, approximate, phylogenetic tree between the sequences and to gradually build up the alignment, following the order in the tree.

Several progressive alignment methods have been implemented and their differ-

[21] In the idealized evolutionary method, the probabilities of evolutionary change would depend on the evolutionary times along each tree branch, on the position-specific structural and functional constraints imposed by natural selection, so that key residues and structural elements would be conserved. High-probability alignments would be the good structural and evolutionary alignments.

ences can lie in the way the order to do the alignment is chosen, in wether the progession involves only alignment of sequences to a growing alignment or wether subfamilies can be built and alignments can be aligned to alignments, or in the scoring procedure [Durbin et al., 1998].

The most widely used implementation of the progressive alignment is ClustalW [Thompson et al., 1994]. The algorithm is similar to the Feng-Doolittle method [Feng and Doolittle, 1987] (one of the first and most relevant progressive alignment algorithms) but it relies on the carefully tuned use of profile alignment methods [Durbin et al., 1998]. It consists in three main steps.

First, all pairs of sequences are aligned separately in order to calculate a distance matrix giving the divergence of each pair of sequences ($N(N$-1)/2 pairs for N sequences). The scores were canonically calculated as the number of k-tuple matches (runs of k identical residues) in the best alignment between two sequences minus a fixed penalty for every gap. ClustalW allows to choose between this method and the slower but more accurate scores from full dynamic programming alignments using two gap penalties (for opening or extending gaps) and a full amino acid weight matrix. These scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded). Both of the scores are initially calculated as percent identity scores and converted to distances by dividing by 100 and subtracting from 1.0, to give number of differences per site. No correction for multiple substitutions is made at this stage.

Second, the trees used to guide the final multiple alignment process are calculated from the distance matrix using the Neighbour-Joining clustering algorithm [Saitou and Nei, 1987], producing unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed at a position where the means of the branch lengths on either side of the root are equal. These trees are used to derive a weight for each sequence. The weights are dependent upon the distance from the root of the tree but sequences which have a common branch with other sequences share the weight derived from the shared branch (in the normal progressive alignment algorithm, all sequences would be equally weighted).

Third, the sequences are progressively aligned according to the branching order

in the guide tree (using sequence-sequence, sequence-profile, and profile-profile alignment). A series of pairwise alignments are used to align larger and larger groups of sequences, following the branching order in the guide tree (proceeding from the tips of the rooted tree towards the root). At each stage a full dynamic programming algorithm is used with a residue weight matrix and penalties for opening and extending gaps. Each step consists of aligning two existing alignments or sequences. Gaps that are present in older alignments remain fixed. In the basic algorithm, new gaps that are introduced at each stage get full gap opening and extension penalties, even if they are introduced inside old gap positions. In order to calculate the score between a position from one sequence or alignment and one from another, the average of all the pairwise weight matrix scores from the amino acids in the two sets of sequences is used. If either set of sequences contains one or more gaps in one of the positions being considered, each gap versus a residue is scored as zero. The default amino acid weight matrices are re-scored to have only positive values. When weighting the sequences, each weight matrix value is multiplied by the weights from the 2 sequences.

The additional heuristics of ClustalW introduced an improvement in the accuracy and sensitivity of the progressive multiple sequence alignment method, namely for the alignment of divergent protein sequences: individual weights are assigned to each sequence in a partial alignment in order to downweight near-duplicate sequences and up-weight the most divergent ones; amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned; residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure; positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions [Thompson et al., 1994].

T-Coffee

T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) is a more accurate (although slightly slower) method for multiple sequence alignment [Notredame et al., 2000]. It is based on the progressive approach and involves pre-

processing a data set of all pairwise alignments between the sequences, providing a library of alignment information that can be used to guide the progressive alignment. T-Coffee generates intermediate alignments that are based not only on the sequences to be aligned next but also on how all of the sequences align with each other and this information can derive from heterogeneous sources (like a mixture of alignment programs and/or structure superposition).

The primary library contains a set of pair-wise alignments between all of the sequences to be aligned. Two alignment sources, global and local, are used for each pair of sequences: global alignments (figure 1.17) are constructed using ClustalW [Thompson et al., 1994] on the sequences, two at a time, to give one full-length alignment between each pair of sequences; local alignments are the ten top-scoring non-intersecting local alignments, between each pair of sequences, gathered using Lalign [Huang and Miller, 1991]. Libraries are then lists of weighted pair-wise constraints. Each constraint receives a weight equal to percent identity within the pair-wise alignment it comes from (figure 1.17B). For each set of sequences, two primary libraries (global and local) are computed along with their weights.

The two primary libraries are pooled in a simple process of addition: if any pair is duplicated between the two libraries, it is merged into a single entry that has a weight equal to the sum of the two weights; otherwise, a new entry is created for the pair being considered. Pairs of residues that did not occur are not represented (weight zero).

T-Coffee increases the value of the information in the library by examining the consistency of each pair of residues with residue pairs from all of the other alignments. For each pair of aligned residues in the library, the program assigns a weight that is the sum of all the weights gathered through the examination of all the triplets involving that pair, reflecting the degree to which those residues align consistently with residues from all the other sequences. The more intermediate sequences supporting the alignment of that pair, the higher its weight. This process is called library extension (figure 1.17C).

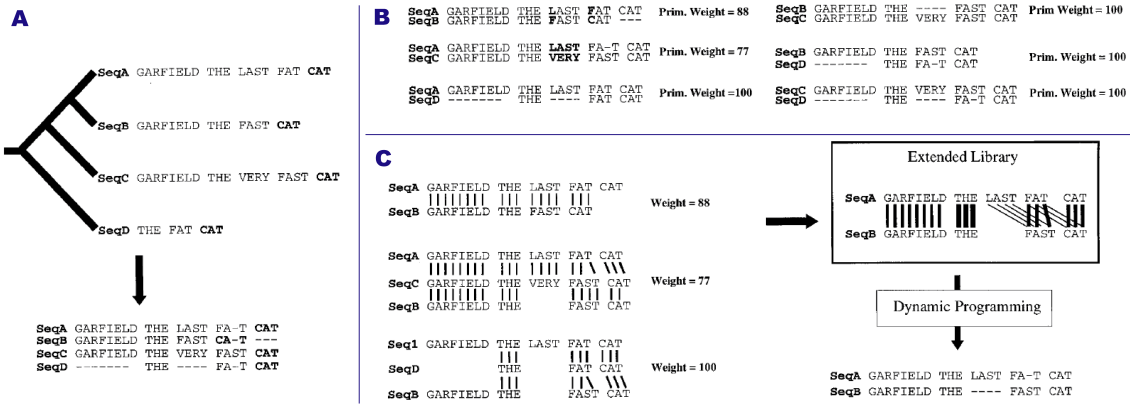

Figure 1.17: T-Coffee - the library extension

A - Progressive alignment: the tree indicates the order in which the four designed sequences are aligned when using a progressive method; in the resulting alignment, the word CAT is misaligned. **B** - Primary library: each pair of sequences is aligned using ClustalW [Thompson et al., 1994] and each pair of aligned residues is associated with a weight equal to the average identity among matched residues within the complete alignment (mismatches in bold). **C** - Library extension for a pair of sequences: the three possible alignments of sequence A and B are shown (A and B, A and B through C, A and B through D); these alignments are combined to produce the position-specific library, which is resolved by dynamic programming to give the correct alignment (thickness of lines indicates the strength of the weight). (Adapted from [Notredame et al., 2000].)

1.3.4 Evolution and phylogeny

The similarity of molecular processes of all studied living creatures led to the suggestion that there was a common ancestor for all organisms on Earth. Any set of species is therefore evolutionary related and this relationship, named phylogeny, is usually represented by a phylogenetic tree [Durbin et al., 1998].

Phylogenetic trees

Phylogenetic analysis of DNA or protein sequences has naturally become an important tool for studying the evolutionary history of organisms from bacteria to humans. The rate of sequence evolution varies extensively with gene or DNA segment and the evolutionary relationships of all levels of classification of organisms can be studied. Phylogenetic analysis is also important for assessing the evolutionary pattern of multigene families, for comprehending the adaptive evolution at the molecular level or for understanding the mechanism of maintenance of polymorphic alleles in populations.

Trees representing phylogenetic relationships of genes and organisms can be presented with a root (rooted tree - v. Figure 1.19A) or without any root (unrooted tree - v. Figure 1.19B). The branching pattern of a tree is called a topology and there are many possible rooted and unrooted topologies for a given number of taxa [22].

There are several statistical methods that can be used for building phylogenetic trees from molecular data. The reconstruction of a phylogenetic tree is a statistical inference of a true phylogenetic tree, which is unknown. Two processes are involved in this inference: estimation of the topology and estimation of branch lengths for a given tree topology. If the topology is known, statistical estimation of branch lengths is relatively simple. Estimating or reconstructing a topology can be problematic. The number of possible topologies rapidly increases with the increasing number of taxa and it is generally difficult to choose the correct topology. In phylogenetic inference a certain optimization principle (e.g. maximum likelihood, minimum evolution) is

[22] Groups of organisms may be formalized groupings recognized by biological classification systems (e.g. species, genus, family) or they may be different populations within a species. Recognized groups are called taxa (plural of taxon). Thus 'taxa' can refer to any kind of taxonomic unit, families, species, populations, DNA sequences, etc [Deonier et al., 2005; Nei and Kumar, 2000].

often used for choosing the most likely topology. Although their theoretical basis is not clearly understood, these principles generally perform well for long sequences [Nei and Kumar, 2000; Nei, 1996].

Distance methods and Neighbor Joining

In distance matrix methods for phylogenetic inference, evolutionary distances are calculated for all pairs of sequences. The phylogenetic tree is then built based on the relationship between the distance values. There are several ways of defining distance but it is usually considered an estimate of the number of nucleotide or amino acid substitutions per site [Nei and Kumar, 2000; Durbin et al., 1998; Nei, 1996].

The simplest distance method for tree construction is the UPGMA [23] [Sneath and Sokal, 1973]. It clusters the sequences, at each stage amalgamating two clusters, simultaneously creating a new node on the tree. The tree can be imagined as being assembled upwards, each node being added above the others, and the edge lengths being determined by the difference in the heights of the nodes at the top and bottom of an edge [Durbin et al., 1998]. When the rate of substitutions varies across lineages, UPGMA tends to generate an incorrect topology, as it assumes constant rate of evolution. Least squares (LS) methods allow different rates of substitutions for different branches. The principle is to compute the minimum sum of squared differences between observed pairwise distances and estimated pairwise distances for a given topology and to choose a topology that shows the smallest minimum sum of squared differences. LS methods often give negative branch lengths and, mainly for this reason, the accuracy of the topology obtained is not particularly high. If the topology of the tree is incorrect, branch lengths substantially loose biological meaning. One way to improve the method's efficiency is to conduct the least squares estimation with the restriction of no negative branch lengths and indeed it has been shown to give, in the case of four sequences, the same results as those obtained by the neighbor joining method (described below) [Nei and Kumar, 2000; Nei, 1996].

The minimum evolution (ME) method compute the total sum (S) of branch length estimates for each of the plausible topologies and. The most likely tree will be the

[23] Unweighted Pair-Group Method using arithmetic Averages

topology with the smallest S. ME requires a large amount of computational time to examine all different topologies if the number of sequences is greater than 10. Neighbor joining (NJ) [Saitou and Nei, 1987] is an efficient simplified ME-based tree building method that does not examine all possible topologies but uses a ME principle at each stage of taxon clustering.

Application of the NJ method is illustrated in Figure 1.18. Computation of S begins with a star phylogeny. All interior branches are assumed to be 0, which is clearly incorrect, and the S value (S_0) is much higher than the S for the true tree. The following step is to compute S_{ij} for a tree in which sequences i and j are paired and are separated from the rest of the sequences that still form a star tree. If i and j are the neighbors connected by only one node (like sequences 1 and 2 in Figure 1.18), then S_{ij} is smaller than S_0. Thus a pair of neighbors can be identified by computing S_{ij}'s for all pairs of sequences and choosing the smallest S_{ij}. Once this pair is identified, they are combined as a single unit and treated as a single sequence in the next step and this process is repeated until all multifurcating nodes are resolved into bifurcating ones. The NJ tree obtained may not necessarily be the true tree, as distance measures are subject to stochastic errors [Nei, 1996].

The NJ method usually produces the same topology as that of the ME tree when the extent of sequence differences is sufficiently large and the number of nucleotides examined is large (>500). However, if the latter condition is not satisfied the NJ tree can be considerably different from the ME tree but the difference in S between the NJ and ME trees is usually statistically nonsignificant [Nei, 1996].

Parsimony

Maximum parsimony (MP) is one of the most widely used tree building algorithms and it consists in finding the tree requiring a minimal number of substitutions to explain the observed sequences [24]. The nucleotides (or amino acids) of ancestral sequences for a hypothetical topology are inferred under the assumption that mutational changes occur in all directions among the four different nucleotides (or 20 amino acids). All

[24]This approach is an application of Ockham's Razor: "*Pluralitas non est ponenda sine neccesitate*" ("plurality should not be posited without necessity").

Distance Matrix

O.T.U.	O.T.U. 1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

Cycle 1: Neighbors =[1,2]

O.T.U.	O.T.U. 1	2	3	4	5	6	7
2	36.67						
3	38.33	38.33					
4	39.00	39.00	38.67				
5	40.33	40.33	40.00	39.67			
6	40.33	40.33	40.00	39.67	37.00		
7	40.17	40.17	39.83	39.50	38.83	38.83	
8	40.17	40.17	39.83	39.50	38.83	38.83	37.67

Cycle 2: Neighbors =[5,6]

O.T.U.	O.T.U. 1-2	3	4	5	6	7
3	31.50					
4	32.30	32.30				
5	33.90	33.90	33.70			
6	33.90	33.90	33.70	31.30		
7	33.70	33.70	33.50	33.10	33.10	
8	33.70	33.70	33.50	33.10	33.10	31.90

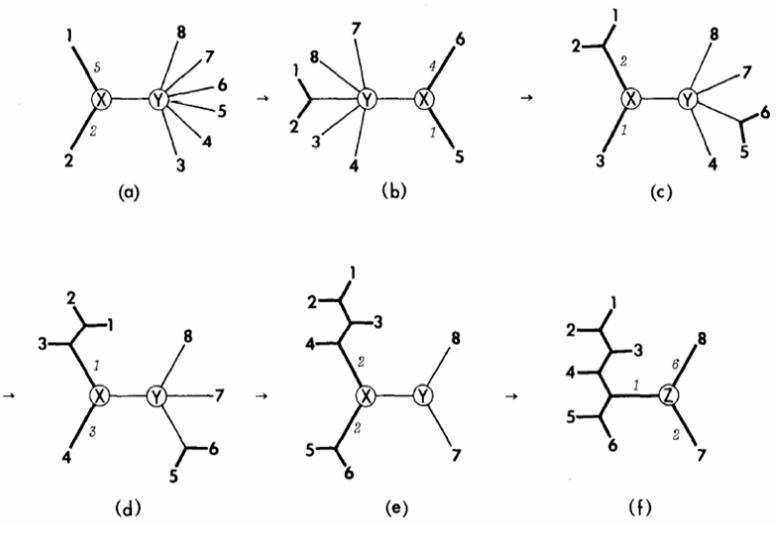

Figure 1.18: The Neighbor-Joining method

The NJ method is applied to a distance matrix (top left, in bold). S_{ij} matrices for two cycles of the method are also represented (center and bottom left). O.T.U. stands for operational taxonomic units (neighbors). Italic numbers are branch lengths and branches with thicker lines indicate that their lengths have been determined. See text for details. (Adapted from [Saitou and Nei, 1987].)

topologies are computed and assigned a cost (in number of mutations) and the tree with the lowest cost is chosen as the best one [Durbin et al., 1998; Deonier et al., 2005; Nei, 1996].

If there are no multiple substitutions at each site, MP is expected to generate the correct (realized) topology as long as enough parsimony-informative sites are examined. However, in practice nucleotide sequences are often subject to backward and parallel substitutions and the number of sites is rather small, introducing uncertainties in phylogenetic inference. Moreover MP may generate an incorrect topology even if an infinite number of nucleotides are examined, when the rate of nucleotide/amino acid varies with evolutionary lineage.

Nonetheless, MP methods are relatively free from some assumptions required for substitution in distance or likelihood methods and they can produce relatively more reliable trees when sequence divergence is low, the rate of substitution is approximately constant and the number of sites is large. MP is also the only method that can easily take care of insertions and deletions of nucleotides, which sometimes give important phylogenetic information [Nei and Kumar, 2000; Nei, 1996].

Maximum likelihood

In maximum likelihood (ML) methods, the likelihood [25] of observing a given set of sequence data for a specific substitution model is maximized for each topology. The 'best' tree is the topology giving the highest maximum likelihood. The considered parameters are not the topologies but the branch lengths for each topology. The likelihood is maximized to estimate branch lengths. ML is known to give the smallest variance of a parameter when sample size is large [Nei and Kumar, 2000].

To explain how likelihood values are computed, lets consider a tree of four DNA sequences illustrated in Figure 1.19A. It is assumed that sequences are n nucleotides long and are aligned with no insertions or deletions.

[25] Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. If the probability of an event X dependent on model parameters p is written $P(X|p)$ then we would talk about the likelihood $L(p|X)$ that is, the likelihood of the parameters given the data.

The observed nucleotides for sequences 1, 2, 3 and 4 at a given k-th site are denoted by x_1, x_2, x_3 and x_4, respectively. Likewise, unknown nucleotides at nodes 0, 5 and 6 are denoted by x_0, x_5 and x_6, respectively. x_i takes any of the four nucleotides A, T, C and G.

For a given site, let $P_{ij}(t)$ be the probability that nucleotide i at time 0 becomes nucleotide j at time t. Allowing the rate of substitutions r to vary from branch to branch, the expected number of substitutions for branch i can be denoted by $v_i \equiv r_i t_i$. In ML, v_i's, regarded as parameters, are estimated by maximizing the likelihood function for a given set of observed nucleotides. The likelihood function for the k-th nucleotide site is

$$l_k = g_{x_0} P_{x_0 x_5}(v_5) P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_0 x_6}(v_6) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \tag{1.1}$$

where g_{x_0} is the prior probability that node 0 has nucleotide x_0 [26].

A specific substitution model is needed to know $P_{ij}(v)$ explicitly. In the equal-input model [Felsenstein, 1981], $P_{ii}(v)$ and $P_{ij}(v)$ (i $\neq$ j) are

$$P_{ii}(v) = g_i + (1 - g_i)e^{-v} \tag{1.2}$$

[26] The relative frequency of nucleotide x_0 in the entire set of sequences is often used as g_{x_0}. However g_{x_0} can also be estimated by ML.

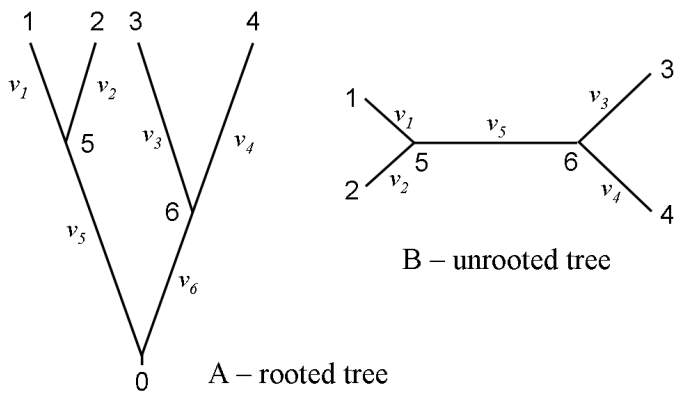

Figure 1.19: Trees to explain the Maximum-Likelihood method

Rooted (A) and unrooted (B) phylogenetic trees for four sequences. In tree B, v_5 is the sum of v_5 and v_6 in tree A. (Adapted from [Nei and Kumar, 2000].)

$$P_{ij}(v) = g_j(1 - e^{-v}) \tag{1.3}$$

where g_i is the relative frequency of the i-th nucleotide.

Using a reversible [27] method of nucleotide substitution for $P_{ij}(v)$, there is no need to consider a root and an unrooted tree can be used (Figure 1.19B) and the reversibility condition is given by

$$g_i P_{ij}(v) = g_j P_{ji}(v) \tag{1.4}$$

for all i and j. Equations 1.2 and 1.3 satisfy condition 1.4.

If the reversible model is used, the number of nucleotide substitutions between nodes 5 and 6 (v_5 + v_6) of tree A remains the same irrespectively of the location of root 0. Designating v_5 + v_6 in tree A by v_5 in tree B and assuming the evolutionary change starts from some point of the tree, the likelihood function can be given by

$$l_k = g_{x_5} P_{x_5x_1}(v_1) P_{x_5x_2}(v_2) P_{x_5x_6}(v_5) P_{x_6x_3}(v_3) P_{x_6x_4}(v_4) \tag{1.5}$$

As x_5 and x_6 are unknown, the likelihood will be the sum of l_k over all possible nucleotides at nodes 5 and 6:

$$\begin{aligned} L_k &= \sum_{x_5} \sum_{x_6} g_{x_5} P_{x_5x_1}(v_1) P_{x_5x_2}(v_2) P_{x_5x_6}(v_5) P_{x_6x_3}(v_3) P_{x_6x_4}(v_4) \\ &= \sum_{x_5} g_{x_5} [P_{x_5x_1}(v_1) P_{x_5x_2}(v_2)][\sum_{x_6} P_{x_5x_6}(v_5) P_{x_6x_3}(v_3) P_{x_6x_4}(v_4)] \end{aligned} \tag{1.6}$$

The likelihood L for the entire sequence is the product of L_k's for all sites, thus

$$lnL = \sum_{k=1}^{n} lnL_k \tag{1.7}$$

ln$L$ is to be maximized by changing the v_i's, using a numerical method. The maximization gives ML estimates of branch lengths v_i's for this topology. The ML values of the remaining topologies (two in the case of four sequences) must also be computed. The ML tree is the topology with the highest ML value and the respective

[27] A reversible model means that the nucleotide substitution between times 0 and t remains the same whether the process is considered to evolve forward or backward in time.

branch lengths are given by the ML estimates of v_i's for this topology [Nei and Kumar, 2000].

ML methods have solid statistics foundations and present some interesting advantages. They have often lower variance than other methods, being frequently the estimation methods least affected by sampling error. They also tend to be robust to many violations of the assumptions in the evolutionary model and, even with very short sequences, they tend to outperform alternative methods (such as parimony or distance methods). Moreover they evaluate different tree topologies and use all the sequence information.

However, in ML the result is dependent on the model of evolution used. Despite several efforts in developing faster algorithms, ML also requires an enormous amount of computational time if many topologies are examined or the extent of sequence divergence is high [Nei and Kumar, 2000].

Bootstrap

The described tree building algorithms give us no measure of how much the generated trees should be trusted. Probably the most common test of the reliability of an inferred tree and assessing the significance of some phylogenetic feature is the bootstrap [Felsenstein, 1985; Efron and Tibshirani, 1993].

Given a dataset consisting of an alignment of sequences, an artificial dataset of the same size is generated by picking columns from the alignment at random with replacement [28]. The tree building algorithm is then applied to this artificial dataset. The whole selection and tree generating procedure is repeated several (typically 1000) times and the frequency with which a chosen phylogenetic feature appears is a measure of its reliability [Durbin et al., 1998].

More specifically, the topology of each bootstrap tree is compared with that of the original tree. Any interior branch of the original tree generating the same partition of sequences as that in the bootstrap tree is assigned value 1, the other branches are given 0. After the process is repeated several hundred times, the percentage of times each original internal branch gets value 1 is taken as the bootstrap confidence value.

[28] A column in the original dataset can appear several times in the artificial dataset.

This procedure is actually not equal to the original phylogenetic bootstrap method [Felsenstein, 1985], which does not assess the reliability of a tree reconstructed from the original data but that of the consensus tree (generated by considering all the bootstrap trees) [Nei and Kumar, 2000].

Molecular clocks and linearized trees

The rate of nucleotide or amino acid substitution would never be constant over the entire evolutionary process as it depends on the evolutionary stability and functional changes of genes. Studying a sufficiently large number of nucleotides/amino acids, for which the extent of sequence divergence is also sufficiently large, should allow the detection of the heterogeneity of evolutionary rate. Nevertheless, the extent of rate heterogeneity is usually moderate when relatively closely related sequences are used and an approximate clock can be used to obtain estimates of times of divergence between sequences from molecular data. The molecular clock hypothesis states that the rate of substitution is approximately constant over evolutionary time, despite the stochastic error associated with the actual number of substitutions.

The use of a molecular clock for estimating divergence times requires a test of the applicability of a clock for the data set of interest. Sequences deviating significantly from the assumption of rate constancy must be identified and excluded. After elimination of these sequences, the branch lengths of the tree for the remaining sequences can be reestimated under the assumption of rate constancy. The resulting linearized tree can be used for estimating the divergence time of any pair of sequences (provided that the rate of substitution can be estimated from other sources such as fossil records or geological dates).

Amongst the molecular clock phylogenetic tests that can be used in the construction of linearized trees, there are two simple methods of testing rate constancy specifically designed to identify sequences evolving excessively fast or slow: the two-cluster and the branch-length tests [Takezaki et al., 1995]. These tests are intended to be applied to a tree built without the assumption of rate constancy and the root of the tree is first located by using outgroup sequences. The two-cluster test examines if the difference in average branch length between two clusters of sequences created

by an interior node is statistically significant. The standard error of the difference between the two average branch lengths is computed. If the subsequently applied Z test indicates that one branch length is significantly different from the other, the most different from the average root-to-tip distance for all sequences is eliminated. In the branch-length test, the root-to-tip branch length is computed for all sequences and the difference between that value for a particular sequence and the average for all sequences is determined. This difference is then subjected to a statistical test to identify the sequences that evolve significantly faster or slower [Nei and Kumar, 2000; Nei, 1996].

1.4 DNA Microarrays

DNA microarrays (also called Gene Chips) are tools for studying gene dynamics in a highly parallel fashion. They are ordered collections of DNA sequences (probes) deposited on solid surfaces or three-dimensional matrices. The basic principle by which DNA microarrays work is the process of hybridization: a single strand of DNA (the probe, immobilized on a surface) is capable of annealing to a complementary strand of DNA (the target), forming a highly stable duplex structure [Knudsen, 2002; Miller, 2004].

1.4.1 Expression arrays

Gene expression studies try to assess the amount of transcribed mRNA in a biological system. Most changes in a cell state are associated with variations in mRNA levels for some genes, despite post-translation modifications in many proteins. It is then extremely useful to systematically measure the transcriptome. Thus it is not surprising that the original and still most popular format of DNA array is the expression microarray, designed to measure the relative abundance of mRNA transcripts - the DNA probes (which can number in the hundreds of thousands on a single chip) are derived from the transcribed regions of genes. The probes may be long sequences several hundred to several thousand bases in length amplified from cloned mRNA transcripts (cDNA probes) or short DNA sequences (oligonucleotide probes) that are

20 to 80 bases in length and can be synthesized *in situ* [29]. This expression profiling has become the dominant use mode because it provides a wealth of important functional information about the biological sample being analyzed. Full or partial transcript sequences are now available for nearly all genes in the most commonly studied organisms. Thus the new high-density arrays can provide genome-wide response profiles for the changes in transcription rate associated with drug treatments, disease states, phenotypic differences and mutations [Miller, 2004; Stoughton, 2005; Relógio, 2002; Parmigiani et al., 2003].

A typical "two-color" microarray hybridization experiment starts by the labelling of the cDNA targets: cellular mRNA is extracted (figure 1.20A) from two cell populations (or tissues) for which relative gene expression levels are to be compared (the 'test' and 'reference' RNAs); the RNAs are then reverse transcribed into cDNA (figure 1.20B) and fluorescently labelled (figure 1.20C) with different color fluorophores (Cy3 and Cy5 dyes) - one cDNA target population will fluoresce green (Cy3) and the other will fluoresce red (Cy5). The targets are subsequently purified, mixed together and simultaneously hybridized to the same microarray (figure 1.20D). After the hybridization reaction, the microarrays are washed, dried, and scanned for detection of fluorescence on the DNA probes. A gene expressed in one or both of the RNA samples will have its mRNA converted into fluorescently labelled cDNA, which will subsequently bind to its corresponding probe during the hybridization reaction. This is detected by a scanner which focuses specific wavelengths of laser light on the probes in order to excite fluorescence of Cy3 and Cy5. The signal is then captured in a two-channel 16-bit TIFF image which encodes the emitted fluorescence of each fluorophore as relative units of pixel color saturation (signal intensity) in each channel (figure 1.20E). If the magnitude of the average signal intensity for a given gene probe is equivalent in the Cy3 and Cy5 channels, it is assumed that the transcript levels of that gene in the two RNA samples are equivalently expressed. If they are not equivalent, the gene is differentially expressed, the channel with the higher intensity corresponding to the sample in which the gene is more highly expressed [30].

[29] Probes are sinthesized directly on the microarray surface by phosphoramidite chemistry and light-sensitive enzymes.

[30] Differences in transcript levels can also occur as the result of variation in mRNA half-life and

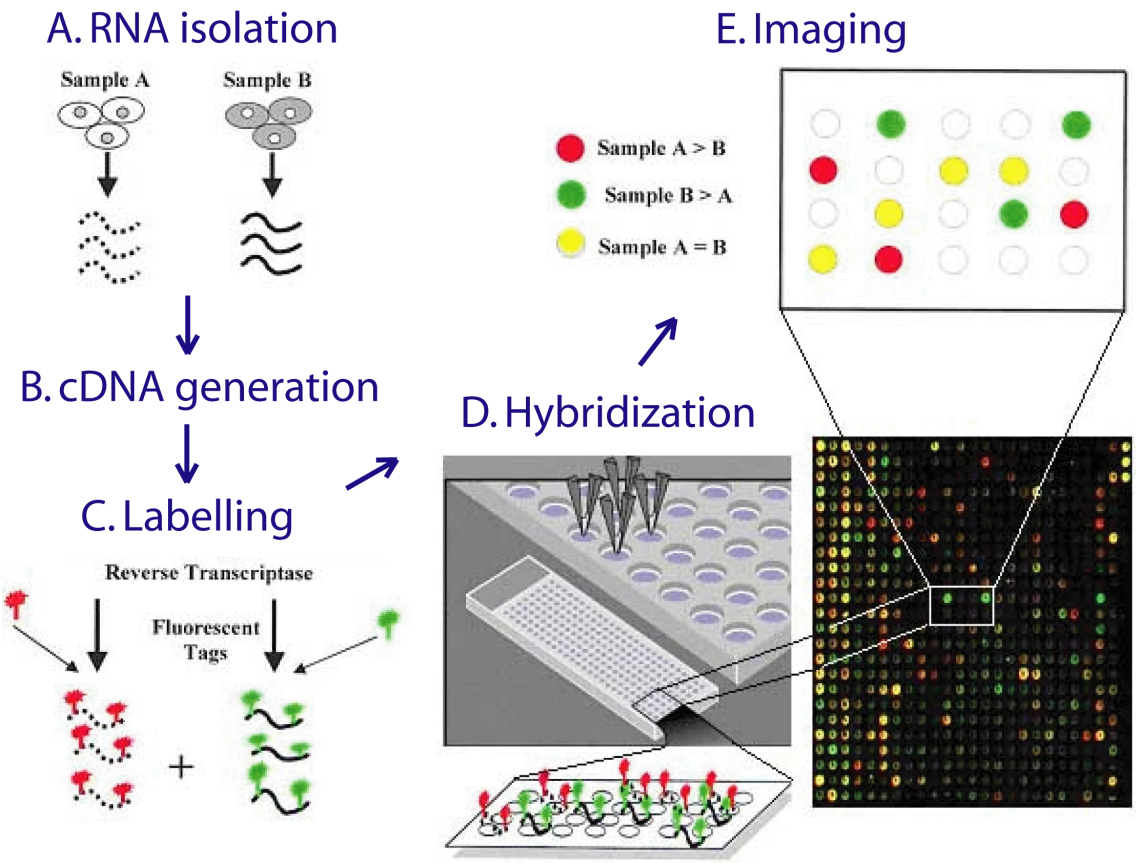

Figure 1.20: "Two-color" DNA microarray experiment
See text for details. (Adapted from the Food and Agriculture Organization of the United Nations website - `http://www.fao.org`.)

Affymetrix uses an alternative approach, named "single-color": only one sample is assessed per chip. The mRNA from a single sample is reverse transcribed into cDNA, which is then utilized to transcribe and amplify target cRNA. This biotinylated cRNA is then hybridized to the chip and subsequently labelled, being bound by the fluorescing molecule phycoerythrin. The array is scanned for fluorescent signals and the resultant single-channel image is analyzed for probe signal intensities that are supposed to approximate the absolute expression levels of the bound mRNA transcripts.

1.4.2 Array CGH

The flexibility of the DNA microarray technology allowed it to extend its range of applications. Comparative genomic hybridization (CGH) is a technique that detects and maps changes in copy number of DNA sequences [Miller, 2004; Albertson and Pinkel, 2003].

In array CGH genomic DNA is used to generate fluorescent targets. DNA from a test (e.g. tumor) and a reference genome (genomic DNA from a normal individual) are differentially labeled and hybridized to a representation of the genome (originally a metaphase chromosome spread) [31]. The fluorescence ratio of the test and reference hybridization signals is determined at different positions along the genome. This gives information on the relative copy number of sequences in the test genome compared with the normal diploid genome.

Array CGH has been widely used for the analysis of tumor genomes and constitutional chromosomal aberrations, as often in cancer and other genetic diseases the genome becomes unstable and certain chromosomal regions are amplified or deleted. The amplification of oncogenes (genes that promote tumorigenesis) and the deletion of tumor suppressor genes can result in populations of cells with a growth advantage

not necessarily due to differences in gene expression *per se*.

[31] Array-CGH microarrays use not only cDNA probes designed for expression analysis but also PCR products or long oligonucleotides representing intergenic sequences or alternatively very large contiguous fragments of chromosomal DNA contained within BACs (bacterial artificial chromosomes).

leading to eventual tumor outgrowth.

Genomic arrays have been used for other applications. For example, epigenetic changes in a genome can be measured on arrays prepared from a CpG island library by assessing their methylation status.

1.4.3 Data analysis

Microarrays are a powerful tool but there are several sources of variation in the measurement process can make it hard to extract the biological information of interest. There are technology specific manufacturing errors and, for instance, variability can be introduced in the amplification, purification and concentration of DNA clones for spotting. The protocol of preparation of mRNA from biological samples includes several procedures that can become sources of variability: labelling, extraction, amplification, etc. Ambient conditions (temperature, humidity, etc) introduce additional variability during hybridization. Natural fluorescence and binding of genetic material to the array in unspotted regions can also introduce background noise in the scanning step. Finally, the initialization of algorithms for imaging is human dependent and different algorithms can lead to different fluorescence quantifications. Most of these sources of variation are relatively small but the accumulation of all the errors can become important. Thus all those artifacts must be taken into account when performing the microarray data analysis and several statistical techniques have been developed for all stages of experimentation [Parmigiani et al., 2003].

The identification of the biological questions of interest (and their specificity) leads to the design of the experiment. Choosing the sample size is a key factor on the design and, at this stage, the experimental conditions must be properly assigned to the arrays. Furthermore microarray experiments must be replicated. It is important to have not only internal controls in the arrays but both "biological replicates" (RNA of the same type from different subjects) and "technical replicates" (multiple arrays using the same RNA).

The raw data of a microarray experiment is the set of pixel intensities stored in the image files generated by the scanner. Image analysis tools are then used for

segmentation [32] and summarization of pixel-level data. Data from every array must be visually inspected to diagnose the existence of possible artifacts. Several techniques for exploratory data analysis and quality control have been developed that allow, for example, the detection of print tip effects, the evaluation of spatial bias or the assessment of intensity effects (like saturation). Subtracting background noise is also part of the preprocessing. Furthermore, data from each probe set must be summarized into a single measure, to estimate the expression level of the gene of interest. Finally, before screening for differential expression, it is important to normalize the signals within each array (e.g. to account for differential response of the two channels) and across arrays (e.g. to account for differences in the environment, sample preparation or the processing of the arrays).

Analyzing preprocessed data involves the selection of genes that are differentially expressed across experimental conditions, as usually the main goal of the experiment is to identify those regulated by modifying conditions of interest. There are several methods for the evaluation of reliability of results and statistical validation of putative differentially expressed genes. Moreover clustering methods are used to classify biological samples or genes by dividing a set of objects (samples or genes) into groups so that gene expression patterns within a group are more alike than patterns across groups. Methods like PCA (principal component analysis) create a small number of variables that summarize most of the variability.

Microarray analysis is also used for the classification of samples, based on gene expression patterns, into known categories associated with biological/clinical features (class prediction). Specific statistical modelling and pattern recognition tools have been developed for this purpose.

More robust validation and interpretation of microarray data results can be obtained through comparisons across platforms and the use of multiple independent datasets [Parmigiani et al., 2003].

[32] Segmentation is the definition of the areas in the image that represent expression information.

1.5 Objectives

The main goal of this work has been to identify and characterize the mechanisms associated with complexity in eukaryotic gene expression, following bioinformatics approaches. My studies have been focused on mRNA splicing and its regulation. I have analyzed both *trans* elements (spliceosomal protein components - the so called splicing factors) and *cis* regulators (namely RNA sequences recognized and bound by splicing factors).

I have aimed to shed some light on the evolutionary history of the splicing machinery by annotating splicing factors in different eukaryotic species. I have tried to address questions such as if splicing in vertebrates benefited from novel lineage-specific mechanisms or just evolved upon the refinement of the ancestral contrivance. I have also aimed to evaluate and discriminate specificities in the evolution of the different elements that comprise the splicing apparatus.

My work also centered on the distinction and identification of RNA sequence-level splicing regulatory elements. I have applied bioinformatics techniques to the recognition of different RNA binding motifs for splicing factors. Moreover, I have tried to establish the functional consequences of variations in the abundance and sequence of those signals. This analysis was applied to cellular processes as important as mRNA metabolism and apoptosis and involved studying the patterns of alternative splicing for associated key genes.

My bioinformatics tools have been applied in the annotation and analysis of sequences for microarray projects. The microarrays technology is a powerful tool in complexity studies and can be very useful in addressing all the described questions, as it allows the evaluation of gene expression profiles on a genomewide scale. For example, we are interested in profiling CpG islands and their methylation patterns, as they are critical in gene expression regulation, cell differentiation and tumor suppression.

Chapter 2

Selective expansion of splicing regulatory factors

(The original work described in this chapter has been integrally published [Barbosa-Morais et al., 2006].)

Keywords: spliceosome; splicing regulation; alternative splicing; retrotransposition; evolution.

Abstract: Although more than 200 human spliceosomal and splicing-associated proteins are known, the evolution of the splicing machinery has not been studied extensively. The recent near-complete sequencing and annotation of distant vertebrate and chordate genomes provides the opportunity for an exhaustive comparative analysis of splicing factors across eukaryotes. We describe here our semi-automated computational pipeline to identify and annotate splicing factors in representative species of eukaryotes. We focussed on protein families whose role in splicing is confirmed by experimental evidence. We visually inspected 1894 proteins and manually curated 224 of them. Our analysis shows a general conservation of the core splicosomal proteins across the eukaryotic lineage, contrasting with selective expansions of protein families known to play a role in the regulation of splicing, most notably of SR proteins in metazoans and of heterogeneous nuclear ribonucleoproteins (hnRNP) in vertebrates. We also observed vertebrate-specific expansion of the CLK and SRPK kinases (which

phosphorylate SR proteins), and the CUG-BP/CELF family of splicing regulators. Furthermore we report several intronless genes amongst splicing proteins in mammals, suggesting that retrotransposition contributed to the complexity of the mammalian splicing apparatus.

2.1 Introduction

In most eukaryotes, functional messenger RNAs (mRNAs) are produced by accurately removing noncoding sequences (introns) from precursors (pre-mRNAs) in a process termed 'RNA splicing'. The spliceosome, a large multicomponent ribonucleoprotein complex, carries out this intron excision [Jurica and Moore, 2003; Burge et al., 1999]. Extensive genetic and biochemical studies in a variety of systems have revealed that the spliceosome contains five essential small RNAs (snRNAs), each of which functions as an RNA-protein complex called a small nuclear ribonucleoprotein (snRNP). Each snRNP comprises one of these five snRNAs bound stably to two classes of proteins: Sm proteins, which are present in all snRNPs, and specific proteins that are uniquely associated with only one snRNP [Luhrmann et al., 1990]. Higher eukaryotes have two distinct types of spliceosomes. The major or U2-type spliceosome, which catalyses the removal of most introns, is composed of U1, U2, U4, U5 and U6 snRNPs. The minor or U12-type spliceosome, which recognises <1% of all human introns, comprises U11, U12, U4atac, U5 and U6atac snRNPs [Patel and Steitz, 2003]. In addition to snRNPs, splicing requires many non-snRNP protein factors. Recent improved methods to purify spliceosomes coupled with advances in mass spectrometry have revealed that the spliceosome may be composed of as many as 300 distinct proteins [Jurica and Moore, 2003; Nilsen, 2003].

The initial events of spliceosome assembly require recognition of specific sequences located at the 5' and 3' splice sites, which define the intron boundaries. In metazoans, however, the splice site sequences are only weakly conserved and although introns are excised with a high degree of precision, at least 74% of human genes encode alternatively spliced mRNAs [Johnson et al., 2003]. Alternative splicing is the process by which multiple mRNAs can be generated from the same pre-mRNA by the differential

joining of 5' and 3' splice sites. Alternative splicing produces multiple mRNAs encoding distinct proteins, thus expanding the coding capacity of genes and contributing to the proteomic complexity of higher organisms [Black, 2003; Brett et al., 2002; Maniatis and Tasic, 2002].

In general, alternative splicing is regulated by protein factors that recognise and associate with specific RNA sequence elements either to enhance or to repress the ability of the spliceosome to recognise and select nearby splice sites [Maniatis and Tasic, 2002; Smith and Valcarcel, 2000]. The multiplicity of protein-protein and protein-RNA interactions that modulate the association of the spliceosome with the pre-mRNA is thought to control alternative splicing [Black, 2003; Caceres and Kornblihtt, 2002; Graveley, 2002].

The evolutionary history of the splicing machinery has not been fully elucidated, in part because appropriate near-complete genome sequences have only recently become available. The recent sequencing and annotation of the genomes of the Japanese puffer fish, *Fugu rubripes* [Aparicio et al., 2002] and the sea squirt, *Ciona intestinalis* [Dehal et al., 2002] allows us now to fill that gap with fiducial branches of distant vertebrates and chordates respectively, providing an opportunity to exhaustively look at splicing factors in those species and extend our knowledge about their evolution. In this study we report a semi-automated computational pipeline designed to identify and annotate splicing factors in representative species of eukaryotes.

2.2 Methods

The key steps in our pipeline are illustrated in Figure 2.1.

All the human splicing factors (Table A.1) and homologues annotated for other species were listed and their protein sequences were retrieved. Grouping into families was performed based on full-length homology, functional domains composition and the Ensembl Protein Family classification [Hubbard et al., 2002] [1] (v30). For each family, spurious and truncated proteins were identified and removed manually and all the remaining members were aligned with T-Coffee [Notredame et al., 2000] (de-

[1] `http://www.ensembl.org`

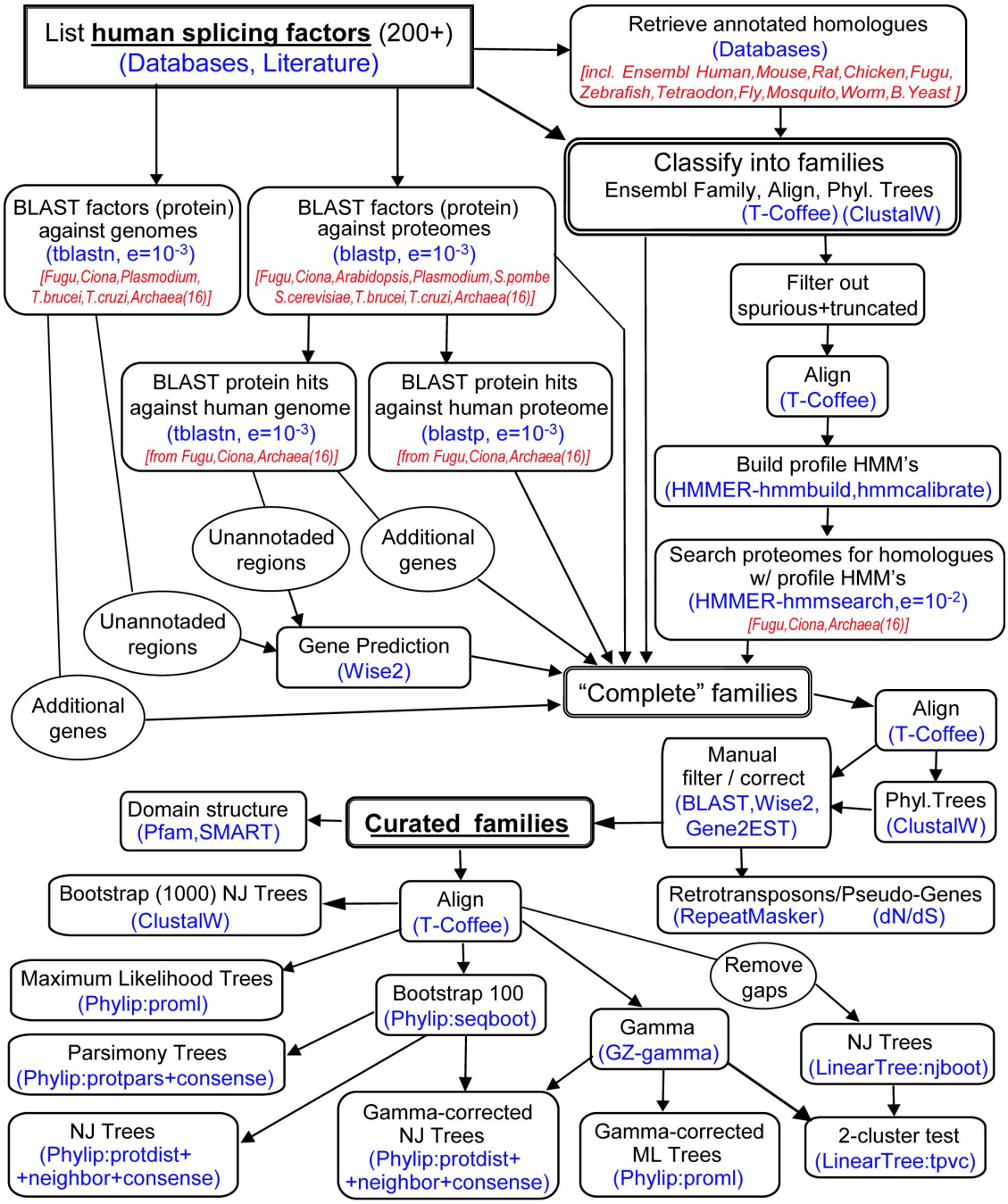

Figure 2.1: Schematics of the computational pipeline flow

Sources, software and parameters are represented in blue and species in red.

fault parameters). The alignment was used to build a profile HMM (Hidden Markov Model), using HMMER [Eddy, 1998] (hmmbuild, hmmcalibrate), with which the proteomes of *Fugu*, *Ciona* and 16 species of Archaea were searched (hmmsearch, e-value = 10^{-2}). In parallel, all the human splicing factors were BLASTed [Altschul et al., 1990] (tblastn, BLOSUM62 matrix, SEG filter on, e-value = 10^{-3}) against the genomes of *Fugu*, *Ciona*, Archaea, *Plasmodium*, *Trypanosomas* and proteomes of the previous species plus *A. thaliana*, *S. pombe* and *S. cerevisiae*. Gene prediction was carried out in hit unannotated genomic regions, using Wise2 [2]. A reciprocal BLAST between the protein hits and the human genome and proteome (blastp, BLOSUM62 matrix, SEG filter on, e-value = 10^{-3}) was performed. Gene predictions were again made for hit unannotaded genomic regions in Human.

The obtained members of each 'complete' family were aligned and a phylogenetic tree was built with ClustalW [Thompson et al., 1994]. The families of factors with relevant annotated function benefited from further curation: removal of false homologues and redundancies, correction of truncated and missannotated proteins, assessment of the likelihood of splice sites. This curation was assisted by BLAST, Wise 2 and EST searches, carried out in the Gene2EST BLAST Server [Gemund et al., 2001] (EMBL) [3] and the NCBI BLAST website [4] (blastn, low complexity filter on), relying on the GenBank/dbEST database (v147.0) [Benson et al., 2004; Boguski et al., 1993].

The same approach was used to identify putative retrotransposons and discriminate pseudo-genes (based on the appearance of frame disruptions like cryptic stop codons and frameshifts introduced by missing or extra nucleotides in the conserved coding region). This procedure was complemented with the estimation of the ratio ds/dn of synonymous / non-synonymous substitutions (using SNAP [5]) and the identification of LINEs and LTR elements by searching the involving genomic sequences (1.2kb upstream and downstream of the putative transcribed sequence) with RepeatMasker [6] (default parameters).

[2] http://www.ebi.ac.uk/Wise2
[3] http://woody.embl-heidelberg.de/gene2est
[4] http://www.ncbi.nlm.nih.gov/BLAST
[5] http://www.es.embnet.org/Doc/SNAP/
[6] http://www.repeatmasker.org

New alignments were built for the resulting curated families and, for each family, the functional domain composition of its members was compared. The domain organisation of proteins relied on the Pfam database [Bateman et al., 2002] [7] (version 16.0), the HMMER program hmmpfam and the SMART tool [Letunic et al., 2004] [8] (version 4.0).

We then performed the phylogenetic analysis of all the families by generating bootstrapped Neighbor-Joining (NJ) trees with ClustalW (1000 bootstraps). Alternatively, we bootstrapped our alignments using the Phylip [Felsenstein, 1989] program Seqboot (100 bootstraps). Then rooted and bootstrapped NJ and Parsimony trees were built using the Phylip programs Neighbor (preceded by Protdist) and Protpars, respectively. In both cases we generated the consensus trees with Phylip program Consense. We also created rooted Maximum-Likelihood (ML) trees using the Phylip program Proml. For details on tree rooting see Table A.2. We did the molecular clock analysis following a procedure similar to that adopted by [Christoffels et al., 2004] (Table A.3).

Table A.4 summarises the sources for the whole genomes and predicted proteomes used in our search. The automated searches relied on BioPerl [Stajich et al., 2002] [9](v1.30) and Ensembl Perl modules on a Linux platform. All the phylogenetic trees and alignments can be found in Supplementary Materials (A.1).

2.3 Results

2.3.1 Pipeline-assisted annotation of splicing factors

Although recent reports have identified up to 300 distinct proteins associated with the spliceosome [Zhou et al., 2002; Rappsilber et al., 2002], many of these new proteins have not yet been shown to function in splicing and, therefore, they cannot be considered as bona fide splicing factors [Jurica and Moore, 2003]. In this study, we limited our analysis to proteins for which there is experimental evidence of their

[7] http://www.sanger.ac.uk/Software/Pfam
[8] http://smart.embl-heidelberg.de
[9] http://www.bioperl.org

involvement in splicing. Our first goal was to enumerate and annotate the genes encoding spliceosomal proteins in the genomes of human, pufferfish, *Ciona*, the budding yeast *Saccharomyces cerevisiae*, the fission yeast *Schizosaccharomyces pombe*, the plant *Arabidopsis thaliana*, and several species of archaebacteria and protozoa (see 2.2, Figure 2.1 and Table A.4).

Although the availability of 'raw' and 'first pass annotated' genomes (for example the ones in Ensembl [Hubbard et al., 2002]) is proving indispensable for genome-wide studies, detailed analyses are still hampered by the fact that most databases are 'contaminated' with erroneous annotation. In many cases, the current algorithms used in completely automated gene-building pipelines unreliably predict features such as short exons. The algorithms are particularly ineffective with repetitive protein motifs, such as those in RS (arginine-serine-rich) domains, responsible for the protein-protein interactions of SR (Ser-Arg) proteins (important splicing factors - see below). The goal of our semi-automated pipeline was to search *ab-initio* the raw genomic sequence of representative eukaryotes and thus to complement pre-existing annotations, even though these acted as a seed for the pipeline. This approach demanded manual inspection and validation of the results. We therefore visually inspected a total of 1894 putative spliceosomal proteins across eukaryotic genomes and we manually curated 224 sequences (12%). The results are listed in Supplementary Material (A.1). Despite the effort made to manually correct sequences, errors and uncertainties remain, especially for genes poorly supported with EST evidence, and this reduces the precision of the phylogenetic analysis (namely for Parsimony methods) and the consistency of tree topology between different methods of phylogenetic inference (all the trees can be found in Supplementary Materials). We were unable to correct completely 388 proteins (20%) of ambiguous sequence. We identified only five putative splicing factors (all from *Fugu*) that had no previously annotated gene locus. We also report 3 factors (from Zebrafish) that were annotated in older versions of Ensembl but do not appear in version 30. In the process of manual curation we have identified 83 putative pseudo-genes that Ensembl annotates as active genes in human and mouse (Table A.5; see below), indicating that automated annotation is oversensitive.

2.3.2 Selective expansion of splicing regulatory protein families

Having enumerated all currently known splicing proteins, we asked whether major patterns of protein family expansion were evident between different animal phyla. We looked at the genes encoding the seven Sm protein families that associate with all the snRNAs, the Lsm protein families that associate with the U6 snRNA, and several snRNP-specific protein families. Most of these spliceosomal components show apparent one:one orthology mapping (or numerical concordance in the occurrence of paralogs) between vertebrates, invertebrates and unicellular eukaryotes, consistent with previous reports (see [Will and Luhrmann, 2001]). In contrast, we observed a different evolutionary pattern of the minor spliceosome U11/U12-snRNP proteins; they are absent from protozoa, trypanosomes, yeasts and the nematode worm *Caenorhabditis elegans* Table 2.1, but present in *Arabidopsis*, consistent with the identification of U12-dependent introns in this plant [Zhu and Brendel, 2003].

In addition to snRNPs, the spliceosome comprises many non-snRNP protein factors, including DExD/H-box proteins, SR proteins and hnRNP proteins. DExD/H-box proteins constitute a prominent family of core splicing factors. Genetic studies in *S. cerevisiae* have implicated eight DExD/H-box proteins in splicing [Staley and Guthrie, 1998]. Each of these conserved proteins (Prp2p, Prp16p, Prp22p, Prp43p, Brr2, Prp5p, Prp28p, Sub2p) is required for pre-mRNA splicing. Seven additional DExD/H-box proteins were recently found associated with mammalian spliceosomes [Jurica and Moore, 2003]. As shown in Table 2.1, no major expansion of the DExD/H-box gene family occurred during evolution.

The SR proteins, characterised by their typical RS domain containing repeated Arg/Ser dipeptides, are essential factors required for both constitutive and alternative splicing [Maniatis and Tasic, 2002]. Our results show that metazoans contain nine families of SR proteins, six of which have two or more members in mammals, whereas in unicellular eukaryotes there are only one or two SR protein genes (Table 2.2). Thus, the diversity of SR proteins seems to have emerged with multicellularity. Consistent with previous reports, we found no SR proteins in budding yeast but two proteins in fission yeast [Kaufer and Potashkin, 2000; Tacke and Manley, 1999], and we confirmed

Family	Human	Mouse	Rat	Chicken	Fugu	Zebrafish	Tetraodon	Ciona	Fly	Mosquito	C.elegans	Arabidopsis	S.pombe	S.cerevisiae	Plasmodium	T.cruzi
U11/U12-20	UB20	UB20	UB20	UB20	UB20	UB20a UB20b	UB20		UB20							
U11/U12-25	UB25	UB25	UB25		UB25	UB25	UB25	UB25				UB25a UB25b				
U11/U12-31	UB31	UB31	UB31	UB31	UB31	UB31	UB31	UB31		UB31		UB31				
U11/U12-35	UB35	UB35	UB35	UB35	UB35		UB35	UB35		UB35		UB35				
U11/U12-48	UB48	UB48	UB48	UB48	UB48	UB48	UB48	UB48								
U11/U12-65	UB65	UB65	UB65	UB65	UB65	UB65	UB65	UB65	UB65	UB65						
ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABS	ABSa ABSb			ABS	
DDX26	DDX26 DD26B	DDX26 DD26B	DDX26	DDX26 DD26B	DDX26 DD26B	DDX26 DX26b DD26B	DDX26 DD26B	DDX26	DDX26		DDX26					
DDX39	DDX39 BAT1	DDX39 BAT1	DDX39 BAT1	BAT1	DDX39	DDX39 BAT1		DDX39	WM6	DDX39	DDX39	DDX39	UAP56	SUB2	DDX39	DDX39
DDX3XY	DDX3X DDX3Y	DDX3X DDX3Y	DDX3X	DDX3	DDX3a DDX3b	DDX3	DDX3	DDX3	DDX3	DDX3	DDX3a DDX3b	DDX3a DDX3b DDX3c	DED1	DED1 DBP1		DDX3a DDX3b DDX3c DDX3d
DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DDX46	DD46a DD46b	DD46a DD46b	PRP11	PRP5	DDX46	DDX46
DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DDX48	DD48a DD48b	IF4Aa IF4Ab	EIF4A	FAL1	EIF	DDX48
DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DHX15	DH15a DH15b	DHX15	PRP43	DHX15	DH15a DH15b DH15c
DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16	DHX16		DHX16	DHX16	DH16a DH16b	CDC28			
DHX35	DHX35	DHX35	DHX35	DHX35	DHX35		DHX35	DHX35	DHX35	DHX35	DHX35	DHX35				
DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	DHX38	PRP16	PRP16	DHX38	DHX38
DHX8	DHX8	DHX8	DHX8	DHX8a DHX8b	DHX8a DHX8b	DHX8a DHX8b	DHX8	DHX8	DHX8	DHX8	DHX8	DHX8	DHX8	PRP22	DHX8	DHX8
DHX9	DHX9	DHX9	DHX9		DHX9	DHX9	DHX9a DHX9b	DHX9	MLE	DHX9	DHX9					
KIAA0052	K052	K052	K052	K052	K052	K052	K052	K052	K052	K052	K052	K052a K052b	K052a K052b	MTR4	K052	K052a K052b
P68p72	DDX5 DDX17	DDX5 DDX17	DDX5 DDX17	DDX5 DDX17	DDX5 DD17a DD17b	DDX5 DDX17	DDX5 DDX17	p68	DDXP	DDXP	DDXP	RH20 RH30	DBP2	DBP2	DDXP	DDXPa DDXPb DDXPc
U5-100*	DDX23	DDX23			DDX23	DDX23	DDX23	DDX23	DDX23	DDX23	DDX23	DDX23	PRP28	PRP28	DDX23	
U5-200*	U5200	U5200	U5200	U5200	U5200	U5200	U5200	U5200	U5200	U5200	U5200 U5Hyp	U520a U520b	BRR2	BRR2	U5200	U520a U520b

Table 2.1: Compilation of U11/U12 snRNP and DExD/H-box (DEAD) proteins identified in the analyzed genomes

Detailed identification of each gene is provided in Supplementary Material (A.1). Small termination characters identify species/phylum specific duplications.

*Families annotated as snRNP specific

the existence of 19 SR protein genes in *Arabidopsis* [Kalyna and Barta, 2004; Reddy, 2004].

Family	Human	Mouse	Rat	Chicken	Fugu	Zebrafish	Tetraodon	Ciona	Fly	Mosquito	C.elegans	Arabidopsis	S.pombe	Plasmodium	T.cruzi
9G8-SRp20	9G8 SR20	9G8 SR20	9G8 SR20	9G8a 9G8b SR20	9G8 SR20a SR20b	9G8a 9G8b 9G8c SR20a SR20b	9G8 SR20	9G8a 9G8b SR20	9G8 RBP1 RBP1L RSF1	9G8 RBP1 RSF1	RSP6 RSPY	RS21 RS22 RS22A RS32* RS33*			
p54	p54 SR86	p54 SR86	p54 SR86	p54 SR86	p54a p54b SR86	p54 SR86	p54a p54b p54c	SR86a SR86b	p54	p54	p54				
RY1	RY1	RY1	RY1	RY1	RY1		RY1	RY1	RY1	RY1	RY1	RY1			
SC35	SC35 SR46	SC35	SC35	SC35	SC35a SC35b	SC35a SC35b		SC35	SC35	SC35	SC35	SC28* SC30* SC30A* SC33* SC35	SRP1†		
SRm300	SR300	SR300	SR300		SR300	SR300	SR300	SR300	SR300	SR300	SRRM2	SR45			
SRp30c-ASF	ASF SR30C	ASF SR30C	ASF SR30C	ASF	ASFa ASFb SR30C	ASFa ASFb SR30C	ASF	ASFa ASFb	SF2	SF2	SF2	RS31A* SR34 SR34A SR34B SRp30		SF†	
SRp40-55-75	SR40 SR55 SR75	SR40 SR55 SR75	SR40 SR55 SR75	SR40a SR40b SR55 SR75	SR40a SR40b SR55 SR75	SR40a SR40b SR55a SR55b SR75	SR40a SR40b	SR40a SR40b SR40c SR55	SR55	SR40	RSP1 RSP2 RSP5	RSp31* RSp40* RSp41*	SRP2†		SR1†
TopoI-B	T1B	T1B	T1B	T1B	T1Ba T1Bb	T1Ba T1Bb	T1Ba T1Bb		T1B	T1B					
Tra2	Tra2A Tra2B	Tra2A Tra2B	Tra2A Tra2B	Tra2A Tra2B	Tra2B	Tra2A Tra2B	Tra2B	Tra2	Tra2	Tra2a Tra2b	Tra2				

Table 2.2: Compilation of SR proteins identified in the analyzed genomes

Detailed identification of each gene is provided in Supplementary Material (A.1). Small termination characters identify species/phylum specific duplications. None of the analysed SR protein genes was found for *Saccharomyces cerevisiae*.

* *Arabidopsis*-specific SR proteins, technically considered orthologues of the human proteins in the same family (reciprocal BLAST hit) but exhibiting a considerably lower degree of identity with the human factor than their *Arabidopsis* paralogues.

†SR proteins in unicellular eukaryotes can be considered common homologues of all the SR proteins in metazoans; here we include them in the same families of their technical human orthologues (reciprocal BLAST hit).

The hnRNP proteins are a large group of molecules identified by their association with unspliced mRNA precursors (hnRNAs). The hnRNP proteins A, C, F, G, H, I (also termed PTB) and M have been implicated in the regulation of splicing [Black, 2003]. We find that a single *S. pombe* protein shows significant sequence homology to hnRNPs, whereas 13 gene families are found in metazoans (Table 2.3). For each invertebrate hnRNP in *Ciona*, insects or worms, there are, on average, three co-orthologues in the vertebrates human, mouse and *Fugu*. *Ciona* has 16 hnRNP genes, whereas human has 37. Thus, a striking expansion of hnRNP protein gene families occurred in vertebrates.

Interestingly, gene families encoding additional splicing regulators have also ex-

Family	Human	Mouse	Rat	Chicken	Fugu	Zebrafish	Tetraodon	Ciona	Fly	Mosquito	C.elegans	Arabidopsis	S.pombe	T.cruzi*
hnRNP-A	ROA0 ROA1 ROA2 ROA3	ROA0 ROA1 ROA2 ROA3	ROA1 ROA2	ROA0 ROA1 ROA3	ROA0 ROA1 ROA3	ROA0a ROA0b ROA0c ROA3	ROA0 ROA1 ROA3	ROA1a ROA1b ROA3	RO87F RO97D ROA1	RO87F	ROAa ROAb	ROAa ROAb ROAc		
hnRNP-C	RLY RLYL ROC ROCL	RLY RLYL ROC	RLY RLYL ROC	RLY RLYL	RLYLa RLYLb ROCa ROCb	RLYa RLYb RLYLa RLYLb ROCa ROCb	RLYLa RLYLb ROC	ROC						
hnRNP-D-U2	ROAB ROD0 RODL	ROAB ROD ROD0 RODL	ROAB ROD0 RODL	ROABa ROABb RODL	ROABa ROABb ROD0	ROAB ROD0 RODL	ROABa ROABb ROD0	ROAB	RO40 ROD	RO40	RODU2			RODa? RODb?
hnRNP-E	PCB1 PCB2 PCB3 PCB4	PCB1 PCB2 PCB3 PCB4	PCB1 PCB2 PCB3 PCB4	PCB3	PCB2a PCB2b PCB3a PCB3b PCB4a PCB4b	PCB2 PCB3	PCB2 PCB3a PCB3b	PCB	PCB	PCB	PCB			
hnRNP-F-H	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROF ROH1 ROH2 ROH3	GRSF1 ROH1 ROH3	GRSF1 ROFH ROH3	GRSF1 ROFHa ROFHb ROH3	GRSF1 ROFH ROH3	ROFHa ROFHb	ROFH	ROFH	ROFHa ROFHb	ROFHa ROFHb		ROFHa? ROFHb?
hnRNP-G	ROG ROGT	ROG ROGT	ROG ROGT	ROG	ROG	ROG	ROG							
hnRNP-I	PTB1 PTB2 ROD1	PTB1 PTB2 ROD1 smPTB	PTB1 PTB2 ROD1 smPTB	PTB1 PTB2 ROD1	PTB1a PTB1b PTB2 ROD1	PTB1a PTB1b PTB2a PTB2b ROD1	PTB1 PTB2 ROD1	PTBa PTBb	PTB	PTB	PTB	PTBa PTBb PTBc		PTBa? PTBb? PTBc?
hnRNP-K	ROK	ROK	ROK		ROKa ROKb	ROKa ROKb	ROKa ROKb	ROK	ROK	ROK	ROK			
hnRNP-L	ROL ROLH	ROL ROLH	ROL ROLH		ROL ROLHa ROLHb	ROLa ROLb	ROL ROLH	ROL	ROL	ROL	ROL			
hnRNP-M	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	Myel ROM	ROM	ROM	ROM	ROM			ROM?
hnRNP-R	ROQ ROR	ROQ ROR	ROQ ROR	ROQ ROR	ROQa ROQb ROR	ROQ ROR	ROQa ROQb ROR	ROQR	ROQR	ROQR	ROQR	ROQRa ROQRb ROQRc		
hnRNP-U	E1BA ROU ROUHY	E1BA ROU ROUHY	E1BA ROU ROUHY		E1BA ROU ROUH	E1BAa E1BAb ROU0 ROUa ROUb	E1BA ROU ROUHY	ROUa ROUb	ROU	ROU	ROU			
Musashi	MUS1 MUS2	MUS1 MUS2	MUS1 MUS2	MUS1 MUS2	MUS1	MUS1 MUS2a MUS2b	MUS1		MUSa MUSb	MUS	MUS		MUS	

Table 2.3: Compilation of hnRNP proteins identified in the analyzed genomes

Detailed identification of each gene is provided in Supplementary Material (A.1). Small termination characters identify species/phylum specific duplications. None of the analysed hnRNP genes was found for *Saccharomyces cerevisiae* and *Plasmodium falciparum*.

*Proteins signed with '?' are technically orthologues (reciprocal BLAST hit) but the large evolutionary distance (and low sequence similarity) and the absence of experimental data does not allow us to classify them as functional homologues.

panded during the evolution of primitive metazoans into vertebrates (Table 2.4). These include the CLK (CDC-like) and SRPK (SR-protein-specific) kinases that phosphorylate SR proteins, modulating their function in splicing; the CUGBP (CUG-binding) and ETR-like proteins (CELF) implicated in tissue-specific and developmentally regulated alternative splicing; and the alternative splicing regulators FUSE (far upstream element binding) and Elav (embryonic lethal abnormal visual) proteins (for a recent review see [Black, 2003]).

Family	Human	Mouse	Rat	Chicken	Fugu	Zebrafish	Tetraodon	Ciona	Fly	Mosquito	C.elegans	Arabidopsis	S.pombe	S.cerevisiae	Plasmodium	T.cruzi
CLK	CLK1 CLK2 CLK3 CLK4	CLK1 CLK2 CLK3 CLK4	CLK1 CLK2 CLK3 CLK4	CLK2 CLK3 CLK4	CLK2a CLK2b CLK4	CLK2 CLK4	CLK2a CLK2b CLK4	CLK	CLK	CLK	CLK	CLK AFC1 AFC2 AFC3	CLK	CLK	CLK	CLKa CLKb
CUG	CUG1 CUG2 CUG3 CUG4 CUG5 CUG6	CUG1 CUG2 CUG3 CUG4 CUG5 CUG6	CUG1 CUG2 CUG3 CUG4 CUG5 CUG6	CUG1 CUG2 CUG3 CUG4 CUG5 CUG6	CUG1 CUG2a CUG2b CUG3a CUG3b CUG4 CUG5	CUG1 CUG2 CUG3 CUG4 CUG5	CUG2a CUG2b CUG3a CUG3b CUG4 CUG5	CUG1 ETR3	CUGa CUGb	CUGa CUGb	CUG	RBPa RBPb				
ELAV	ELAV1 ELAV2 ELAV3 ELAV4	ELAV1 ELAV2 ELAV3 ELAV4	ELAV1 ELAV2 ELAV3 ELAV4	ELAV1 ELAV3 ELAV4	ELV1a ELV1b ELAV3 ELAV4 ELAV	ELV1a ELV1b ELAV2 ELAV3 ELAV4 ELAV	ELAV1 ELAV2 ELAV3 ELAV4	ELAVa ELAVb	ELAVa ELAVb ELAVc	ELAVa ELAVb ELAVc	ELAV					
FUSE	FUSE1 FUSE2 FUSE3	FUSE1 FUSE2 FUSE3	FUSE2 FUSE3	FUSE1 FUSE2 FUSE3	FUSE1 FUSE2 FUSE3	FUSE1 FUS2a FUS2b FUSE3	FUSE1 FUSE2 FUS3a FUS3b	FUSE	PSI	PSI	FUSEa FUSEb FUSEc FUSEd FUSEe	KHa KHb KHc				
SRPK	MSSK1 SRPK1 SRPK2	MSSK1 SRPK1 SRPK2	MSSK1 SRPK1 SRPK2	SRPK1 SRPK2 SRPK	MSSK1 SRPK1 SRPK2 SRPKa SRPKb	MSSK1 SRK1a SRK1b SRPK2 SRPKa SRPKb SRPKc	SRK1a SRK1b SRK1c SRPK2 SRPKa SRPKb SRPKc	SRPK	SRPK	SRPK	SRPK	MSSKa MSSKb MSSKc SRPKa SRPKb	SRPK	SRPK	SRPK	SRPK

Table 2.4: Evolution of miscellaneous splicing regulatory proteins
Detailed identification of each gene is provided in Supplementary Material (A.1). Small termination characters identify species/phylum specific duplications.

Since genome duplication is known to have occurred at the vertebrate stem [Mazet and Shimeld, 2002; McLysaght et al., 2002], we performed a phylogenetic analysis, using rate-linearised trees (see 2.2) to determine whether the splicing factor family expansions are co-incident with that duplication. Despite some topological inconsistencies between the different methods of phylogenetic inference, the evolutionary trees we generated are most consistent with the model that hnRNP genes underwent one or two rounds of duplication just after the divergence of vertebrates (Figure 2.2) and urochordates.

Furthermore, analysis of the teleost radiation, and of *Arabidopsis* revealed several

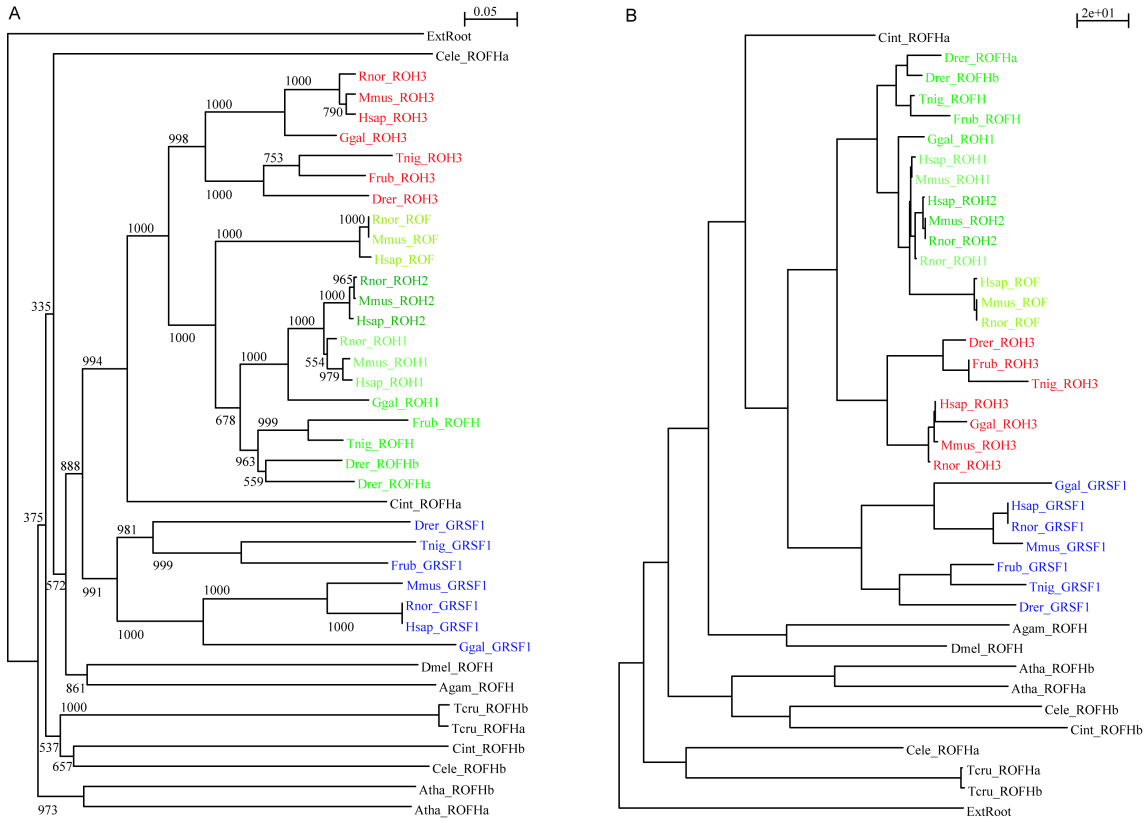

Figure 2.2: Evolutionary relationship among the protein members of hnRNP F/H family in several eukaryotes

Species: human (Hsap), mouse (Mmus), rat (Rnor), chicken (Ggal), *Fugu* (Frub), zebrafish (Drer), *Tetraodon* (Tnig), *Ciona intestinalis* (Cint), fruit fly (Dmel), mosquito (Agam), *C.elegans* (Cele), *Arabidopsis* (Atha) and *Trypanosoma* (Tcru). Vertebrate factors are highlighted in blue, red and shades of green.

A - Rooted Neighbour-Joining phylogenetic tree generated using ClustalW (1000 bootstraps), based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units.

B - Rooted Gamma-corrected Maximum-Likelihood phylogenetic tree generated using GAMMA and the Phylip program Proml, based on amino-acid alignment generated by T-Coffee. Branch lengths are scaled in arbitrary units.

localised gene duplications in *Fugu*, the zebrafish *Danio rerio*, *Tetraodon* (all teleosts) and *Arabidopsis*. These results are consistent with the currently accepted models proposing additional rounds of whole genome duplication in ray-finned and lobe-finned fish, before teleost radiation [Amores et al., 1998; Aparicio et al., 2002; Christoffels et al., 2004] , and the propensity of angiosperms to become polyploid [Bowers et al., 2003; Simillion et al., 2002]. Thus, teleost fish and plants tend to have more copies of splicing genes than do mammals (Tables 2.1, 2.2, 2.3, 2.4). However, there is no evidence for additional selective expansion of any particular family of splicing proteins in these organisms, beyond that which had occurred in the stem organism.

2.3.3 The domain evolution of splicing factors

Our data show conservation of the protein domain structure of splicing factors across species and we found no evidence for domain shuffling. We observed no trend for gain or loss of domains in families of splicing factors, as has occurred in other nuclear protein families (for example, in the Polycomb and Trithorax protein families [Ringrose and Paro, 2004]). We checked, for example, whether the expansion of SR protein families coincided with the appending of RS domains onto general RNA-binding splicing factors. In species without SR proteins, we found no relevant homology with SR protein RNA recognition motifs (RRMs). Each factor seems to have evolved as a whole and its domains have evolved together (Figure 2.3). Similarly, for the hnRNP families that are expanded in vertebrates the motif structures are generally conserved (Figure 2.4). One exception is hnRNP H3, which in mammals and chicken appears to have lost the first of the three RRM's that are common to its paralogues.

2.3.4 Retrotransposition and identification of putative novel splicing factors and pseudogenes in mammals

The absence of introns from mammalian genes is often indicative of retrotransposition, where a spliced mRNA is reverse-transcribed into DNA and integrates back into the genome. Retrotransposition appears to have contributed as a general mechanism of gene duplication amongst mammals. We found that, with the exception of U2AF[26]

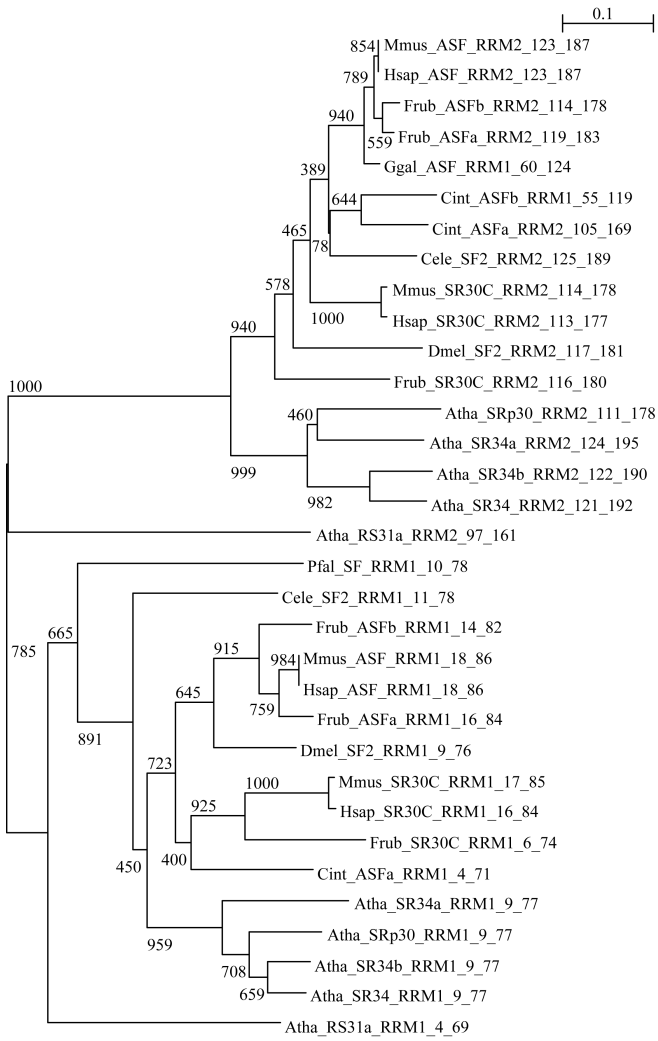

Figure 2.3: Evolutionary relationship among the RNA-recognition motifs (RRM) of members of the family SRp30c-ASF for several eukaryotes

Species: human (Hsap), mouse (Mmus), chicken (Ggal), *Fugu* (Frub), *Ciona* (Cint), fruit fly (Dmel), *C.elegans* (Cele), *Arabidopsis* (Atha) and *Plasmodium* (Pfal) (for simplicity only one rodent, one teleost and one insect are shown).

Amino-acid positions of each domain within the protein are also indicated in the domain identification. The unrooted Neighbour-Joining phylogenetic tree was generated using ClustalW (1000 bootstraps) based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units. RRM1 in Ggal_ASF and Cint_ASFb corresponds to RRM2 in the other proteins as their sequences are truncated in the N-terminal. Pfal_SF is found to have only one RRM. Atha_RS31A can be technically considered an orthologue of the Hsap_SR30C (reciprocal BLAST hit) but exhibits a considerably lower degree of identity (36%) with the human factor than its *Arabidopsis* paralogues (e.g. 53% for Atha_SRp30).

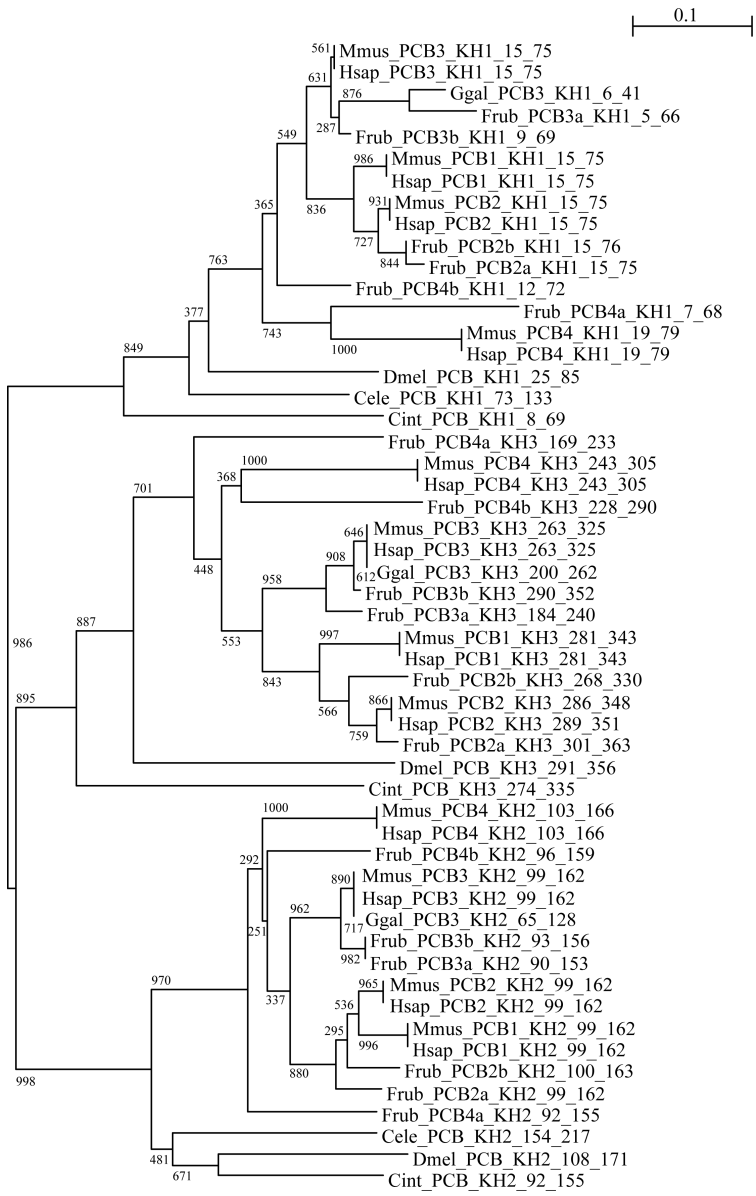

Figure 2.4: Evolutionary relationship among the RNA-binding K-Homology (KH) domains of members of the family hnRNP-E/PCB for several metazoans

Species: human (Hsap), mouse (Mmus), chicken (Ggal), *Fugu* (Frub), *Ciona* (Cint), fruit fly (Dmel) and *C.elegans* (Cele) (for simplicity only one rodent, one teleost and one insect are shown). Amino-acid positions of each domain within the protein are also indicated in the domain identification. The unrooted Neighbour-Joining phylogenetic tree was generated using ClustalW (1000 bootstraps) based on amino-acid alignment generated by T-Coffee. Bootstrap values are shown. Branch lengths are scaled in arbitrary units.

(a mammalian splicing factor [Shepard et al., 2002] that diverged from U2AF^{35} before vertebrates radiation and is likely to have been lost by defunctionalization in teleosts) and Sm N, all of the mammalian specific factors SRp46, U2AF1-RS1 and hnRNPs C-like, E1, smPTB and G-T are intronless whereas their closer paralogues are multiexonic. SRp46, U2AF1-RS1, hnRNP E1 and hnRNP G-T have previously been reported to be retrotransposons [Elliott et al., 2000; Makeyev et al., 1999; Soret et al., 1998; Wang et al., 2004a], which is consistent with our data. We therefore propose that retrotransposition contributed to generate the diversity of the splicing machinery observed in mammals.

We found evidence for additional seven mouse putative intronless genes that appear to have no frame disruption in their coding sequences and for which we find evidence for transcription (Table A.6) and/or have an outstandingly high ratio of synonymous / non-synonymous substitutions when compared with the closest active paralogue. Six of these putative intronless genes are annotated in Ensembl but one of the genes is located in an unannotated genomic region. Two putative intronless genes exhibit transcript sequences equal to their closest paralogues'. Whether these are novel functional splicing genes in mouse or very recent pseudogenes remains an open question.

In addition, we identified 107 human and 90 mouse putative pseudogenes (Tables A.5 and A.7), none being found in other phyla. Of these, 30 human and 53 mouse pseudogenes are annotated as putative functional genes in Ensembl (Table A.5). The majority (∼80%) of all the analysed intronless genes/pseudogenes contain evidence for surrounding LINE1 or LTR (long terminal repeat) sequences (repeats associated with transposable elements [Kazazian, 2004]) and are therefore likely to be retrotransposons. Some families of Sm proteins and the hnRNP-A family contain particularly large numbers of retrotransposons (Tables A.5 and A.7).

2.4 Discussion

Here we report a systematic comparison of the genes encoding the splicing machinery across diverse phyla. We designed a semi-automated computational pipeline to iden-

tify and annotate spliceosomal proteins that will also assist in the rapid re-annotation of new splicing proteins as genomic sequences are updated. Our analysis shows differential gene family expansions across the eukaryotic lineage, with a disproportionate expansion of hnRNP proteins in vertebrates.

Although the origin of introns remains unknown, current data strongly indicate that introns and a spliceosome sufficient for their excision was present in the last common ancestor of eukaryotes [Johnson, 2002; Collins and Penny, 2005]. Introns have been discovered in eukaryotes as primitive as the single-celled parasite *Giardia lamblia* [Nixon et al., 2002] and its close relative *Carpendiemona membranifera* [Simpson et al., 2002], and a core spliceosomal protein gene (*Prp8*) is remarkably conserved between metazoans and the deep-branching protist *Trichomonas vaginalis* [Fast and Doolittle, 1999]. Our finding that genes encoding snRNP proteins are generally conserved in animals, *Arabidopsis*, yeasts, trypanosomes and *Plasmodium* is consistent with previous reports (reviewed by [Will and Luhrmann, 2001]). Our observation that *Plasmodium*, trypanosomes, yeasts and *C. elegans* lack U11/U12 protein homologues is also in agreement with the hypothesis that the minor (U12-dependent) spliceosome was absent from the "first eukaryote" [Collins and Penny, 2005].

In contrast with the conservation of snRNP protein genes, our analysis reveals that metazoans have many more genes implicated in the regulation of splicing than unicellular eukaryotes. Most probably, splicing regulatory proteins evolved as a consequence of whole-genome duplications that occurred at the vertebrate stem [Mazet and Shimeld, 2002; McLysaght et al., 2002]. According to the 'classical' model for selective retention of gene family duplicates [Force et al., 1999; Nei and Rooney, 2005; Ohno, 1970], one of the duplicate genes retained the original function while the other accumulated mutations that eventually conferred an advantageous new function (neofunctionalisation).

We provide surprising evidence that retrotransposition introduced an additional level of diversity to the mammalian splicing machinery. Despite the fact that the majority of retrotransposons are non-functional [Goncalves et al., 2000], and that intronless genes may be transcribed less efficiently than their intron-containing homologues [Le Hir et al., 2003], we identified several retrotransposed genes, specific to mammals,

encoding multifunctional RNA-binding proteins. These include SRp46 [Soret et al., 1998], hnRNP E1 [Antony et al., 2004; Bandiera et al., 2003; de Hoog et al., 2004; Krecic and Swanson, 1999; Leffers et al., 1995; Morris et al., 2004; Ostareck-Lederer et al., 1998; Persson et al., 2003; Reimann et al., 2002], hnRNP G-T [Nasim et al., 2003; Elliott et al., 2000], smPTB [Gooding et al., 2003] and U2AF1-RS1 [Wang et al., 2004a]. We also identified seven mouse putative novel active retrotransposed genes, paralogues of *NHP2-like*, *U1C*, *LSm6*, *LSm7*, *SmD2*, *SmG* and *U2AF*[35].

Although splicing of introns from pre-mRNAs occurs in practically all eukaryotes, alternative splicing is important and widespread only in multicellular organisms. The yeast *S. cerevisiae* has introns in only 3% of its genes and only six genes with more than one intron [Barrass and Beggs, 2003]. Although in the fission yeast *S. pombe*, 43% of the genes are spliced, with many of them containing multiple introns [Wood et al., 2002], no regulated alternative splicing has been detected in this organism or in any other unicellular eukaryote [Ast, 2004; Barrass and Beggs, 2003].

There are two current models to explain the evolution of alternative splicing, which are not mutually exclusive [Ast, 2004]. One is based on the accumulation of mutations that make splice sites sub-optimal (or 'weaker'), providing an opportunity for the splicing machinery to skip that site. In the second model, the evolution of splicing regulatory factors that either enhance or inhibit the binding of the splicing machinery to constitutive splice sites, it argues, releases the selective pressure from that sequence resulting in mutations that weaken the splice sites. Our results clearly support this second model, which so far has not received much experimental attention. The choice of splice site is thought to be regulated by altering the binding of the spliceosome to the pre-mRNA. This is achieved by RNA-binding proteins that associate with non-splice site sequences, located either in exons or introns. The best-characterised families of splicing regulators are SR proteins and hnRNP proteins (reviewed in [Black, 2003]). In vitro selection experiments have identified optimal binding sequences for different SR and hnRNP proteins, but the binding sites for a given family member can be fairly degenerate. Moreover, regulatory proteins can act as either splicing activators or repressors, depending on where in the pre-mRNA they bind. We propose, therefore, that the evolution of novel members of splicing regulatory protein families permitted

the diversification of their canonical binding sites in pre-mRNAs giving the cell the potential to produce new transcripts by altering splice choices. This hypothesis may be testable by correlating functional specificity of individual factors for their splice isoforms, with the cognate recognition sequences in different species.

Chapter 3

Diversity of human U2AF splicing factors

(This chapter is written as review article [Mollet et al., 2006].)

Keywords: U2AF; PUF60; CAPER; RNA splicing.

Abstract: U2 snRNP auxiliary factor (U2AF) is an essential heterodimeric splicing factor composed of two subunits, $U2AF^{65}$ and $U2AF^{35}$. During the past years, a number of proteins related to both $U2AF^{65}$ and $U2AF^{35}$ have been discovered. Here, we review the conserved structural features that characterize the U2AF protein families and their evolutionary emergence, we perform a comprehensive database search designed to identify U2AF protein isoforms produced by alternative splicing, and we discuss the potential implications of U2AF protein diversity for splicing regulation.

3.1 Introduction

In eukaryotes, protein-coding regions (exons) within precursor messenger RNAs (pre-mRNAs) are separated by intervening sequences (introns) that must be removed to produce a functional mRNA. Pre-mRNA splicing is an essential step for gene expression and the vast majority of human genes comprise multiple exons that are

alternatively spliced [Johnson et al., 2003]. Alternative splicing is used to generate multiple proteins from a single gene thus contributing to increase proteome diversity. Alternative splicing can also regulate gene expression by generating mRNAs targeted for degradation [Lareau et al., 2004]. Proteins produced by alternative splicing control many physiological processes and defects in splicing have been linked to an increasing number of human diseases [Nissim-Rafinia and Kerem, 2005].

Pre-mRNA splicing occurs in a large, dynamic complex called the spliceosome, which is composed of four small nuclear ribonucleoprotein particles (the U1, U2, U5 and U4/U6 snRNPs) and more than 100 non-snRNP proteins [Jurica and Moore, 2003]. Spliceosome assembly follows an ordered sequence of events that begins with recognition of the 5' splice site by U1snRNP and binding of U2AF (U2 small nuclear ribonucleoprotein auxiliary factor) to the polypyrimidine (Py)-tract and 3' splice site [Burge et al., 1999]. Human U2AF is an heterodimer composed of a 65-kDa subunit ($U2AF^{65}$) that contacts the Py-tract [Ruskin et al., 1988; Zamore and Green, 1989], and a 35-kDa subunit ($U2AF^{35}$) that interacts with the AG dinucleotide at the 3' splice site [Merendino et al., 1999; Zorio and Blumenthal, 1999a; Wu et al., 1999]. Binding of U2AF is essential for subsequent recruitment of U2snRNP to the spliceosome and splicing of the pre-mRNA.

U2AF has been highly conserved during evolution. In addition, a number of U2AF-related genes are present in the human genome, and some are alternatively spliced. Here, we review currently available information on the diversity of U2AF proteins and we discuss resulting implications for splicing regulation.

3.2 Structural features of U2AF and U2AF-related proteins

The $U2AF^{65}$ protein contains three RNA recognition motifs or RRMs [Zamore et al., 1992] (Table 3.1). The two central motifs (RRM1 and RRM2) are canonical RRM domains responsible for recognition of the polypyrimidine tract (Py-tract) in the pre-mRNA, while the third RRM has unusual features and is specialized in protein-protein interaction. This unusual RRM-like domain, called UHM for U2AF homology

motif, is present in many other splicing proteins [Kielkopf et al., 2004]. The UHM in U2AF^{65} recognizes Spicing Factor 1 (SF1) and this cooperative protein-protein interaction strengthens the binding to the Py-tract (Figure 3.1). At the opposite end, the N-terminal part of U2AF^{65} interacts with U2AF^{35} and this association further strengthens the binding to the Py-tract [Kielkopf et al., 2004]. Although it is not a member of the serine-arginine (SR) family of splicing factors, the U2AF^{65} protein contains an arginine and serine rich (RS) domain that is required for spliceosome assembly [Valcarcel et al., 1996; Shen and Green, 2004].

Gene	Protein	Domain Organization		References
U2AF2	U2AF^{65}	RS – RRM1 – RRM2 – UHM	475aa	(Zamore et al. 1992)
SIAHBP1	PUF60	RRM1 – RRM2 – UHM	559aa	(Page-McCaw et al. 1999)
RNPC2	CAPERα	RS – RRM1 – RRM2 – UHM	530aa	(Jung et al. 2002; Dowhan et al. 2005)
RBM23	CAPERβ	RS – RRM1 – RRM2	424aa	(Dowhan et al. 2005)

Table 3.1: Domain organization of U2AF^{65} and U2AF^{65}-related proteins
Domains are annotated according to [Kielkopf et al., 2004]. RS: Arg-Ser rich; RRM: RNA recognition motif; UHM: U2AF homology motif. The gene names approved by the HUGO Gene Nomenclature Committee, `http://www.gene.ucl.ac.uk/nomenclature/` have been included.

PUF60 (Poly U binding Factor-60kDa) was first isolated as a protein closely related to U2AF^{65} that was required for efficient reconstitution of RNA splicing *in vitro* [Page-McCaw et al., 1999]. The homology between PUF60 and U2AF^{65} extends

across their entire length with the exception of the N-terminal where PUF60 lacks a recognizable RS domain (Table 3.1 and Figure 3.2A). CAPERα and CAPERβ are the most recently characterized proteins related to U2AF^{65} [Jung et al., 2002; Dowhan et al., 2005]. Both have a domain organization similar to U2AF^{65}, except for the C-terminus of CAPERβ that lacks the UHM domain (Table 3.1 and Figure 3.2A).

The U2AF^{35} protein contains a central UHM domain (previously called Ψ-RRM) involved in the interaction with U2AF^{65}, flanked by two Zn^{2+} binding motifs and a C-terminal RS domain [Zhang et al., 1992] (Table 3.2 and Figures 3.1 and 3.2B). Three-dimensional structural information revealed that despite low primary sequence identity (23%), ligand recognition by the U2AF^{65}-UHM and U2AF^{35}-UHM domains is very similar [Kielkopf et al., 2004]. Both the U2AF^{35}/ U2AF^{65} and U2AF^{65}/SF1 interactions involve a critical Trp residue in the ligand sequence that inserts into a tight hydrophobic pocket created by the UHM (Figure 3.3).

In the human genome there are at least three genes that encode proteins with a high degree of homology to U2AF^{35} (Table 3.2 and Figure 3.2B). U2AF^{26} is a 26-

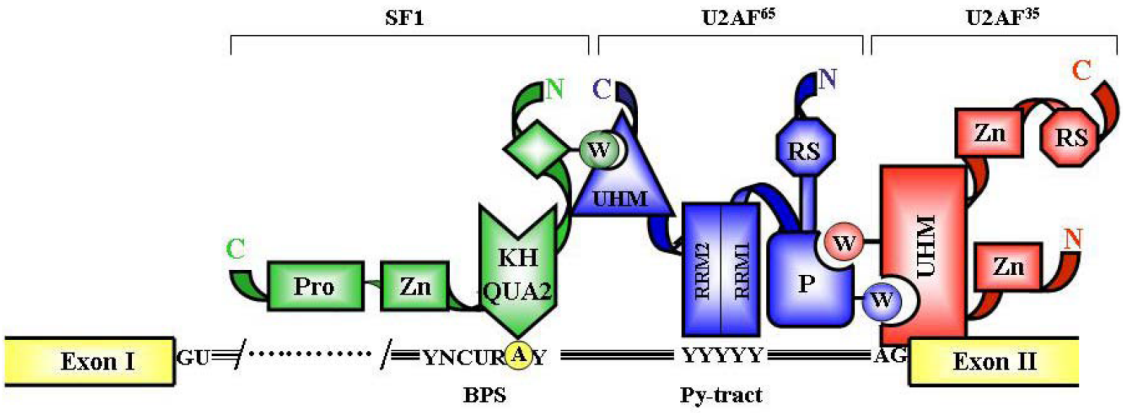

Figure 3.1: Schematic representation of protein-protein and protein-RNA interactions mediated by the U2AF heterodimer during the early steps of spliceosome assembly
Binding of the U2AF heterodimer to the poly-pyrimidine tract (Py-tract) and 3'-splice site AG is strengthened by the cooperative interaction between U2AF^{65} and SF1 at the branchpoint sequence (BPS). The ligand Trp residues (W) in SF1 and U2AF^{65} insert into the UHM pockets in U2AF^{65} and U2AF^{35}, respectively. An additionally exposed Trp residue on the U2AF^{35} UHM domain inserts between a series of unique Pro residues at the C-terminus of the U2AF^{65} ligand (P).

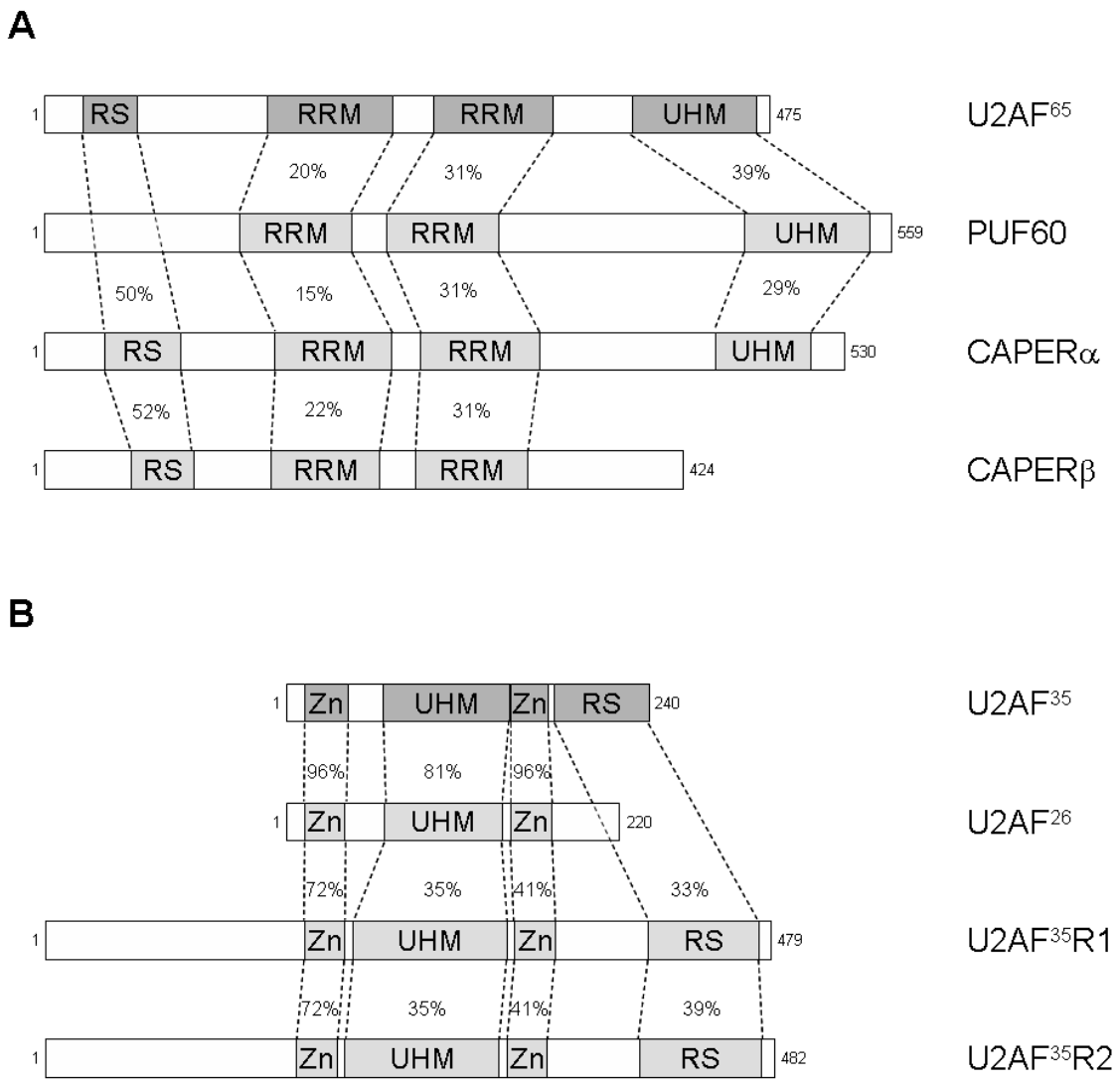

Figure 3.2: A schematic alignment of human protein families related to U2AF^{65} (A) and U2AF^{35} (B)

A - The putative functional domains in each protein are aligned with U2AF^{65} and the similarity (% identity) of these domains in relation to U2AF^{65} is indicated. **B** - The putative functional domains in each protein are aligned with U2AF^{35} and the similarity (% identity) of these domains in relation to U2AF^{35} is indicated.

Gene	Protein	Domain Organization	References
U2AF1	U2AF^{35}	Zn UHM Zn RS **240aa**	(Zhang et al. 1992)
U2AF1L4	U2AF^{26}	Zn UHM Zn **220aa**	(Shepard et al. 2002)
U2AF1L1	U2AF^{35}R1	Zn UHM Zn RS **479aa**	(Kitagawa et al. 1995)
U2AF1L2	U2AF^{35}R2	Zn UHM Zn RS **482aa**	(Kitagawa et al. 1995; Tronchere et al. 1997)

Table 3.2: Domain organization of U2AF^{35} and U2AF^{35}-related proteins
Domains are annotated according to [Kielkopf et al., 2004]. Zn: zinc binding; UHM: U2AF homology motif; RS: Arg-Ser rich. The gene names approved by the HUGO Gene Nomenclature Committee, `http://www.gene.ucl.ac.uk/nomenclature/` have been included.

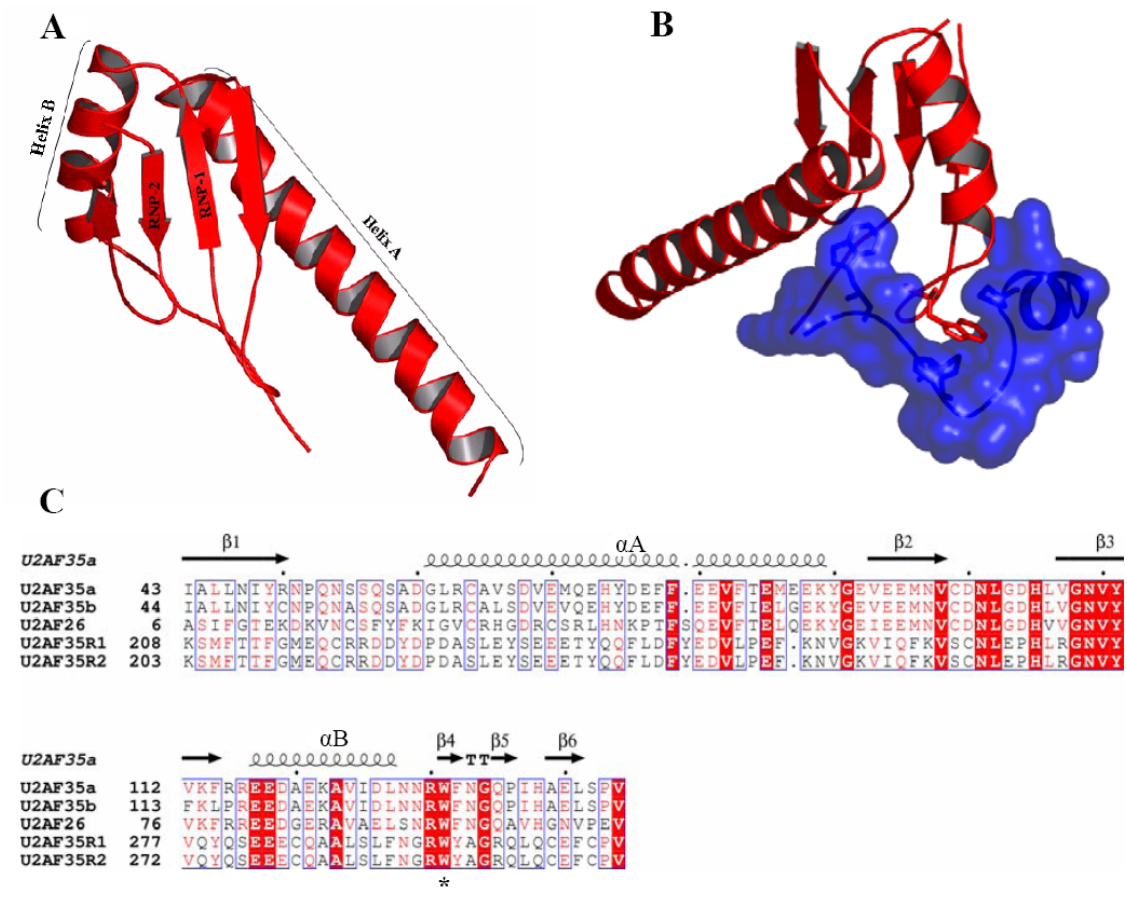

Figure 3.3: The U2AF^{35}-UHM/U2AF^{65}-ligand complex

A - Ribbon representation of the U2AF^{35} UHM. Residues 43-146; pdb code: 1jmt. **B** - Structure of the U2AF^{35}-UHM/U2AF^{65}-ligand (blue) complex [Kielkopf et al., 2001]. A critical W residue (Trp92 in U2AF65) inserts into a tight hydrophobic pocket between the α-helices and the RNP1- and RNP2-like motifs in U2AF^{35} [Kielkopf et al., 2001]. An Arg residue (Arg 133 in U2AF^{35}) on the loop connecting the last α-helix and β-strand of the UHM contributes to the Trp-binding pocket. A neighboring W residue (Trp 134 in U2AF^{35}) inserts between a series of unique Pro residues at the C-teminus of U2AF^{65} (residues 85-112). Additionally, a series of acidic residues in Helix A of the UHM interacts with basic residues at the N-terminus of U2AF^{65}. The molecular representations were generated using PyMol (`http://www.pymol.org`). **C** - Sequence alignment of the UHM region in the alternatively spliced U2AF^{35} isoforms (U2AF^{35}a and U2AF^{35}b) and in the genes that encode U2AF^{35}-related proteins. The conserved Trp residues are identified by an *. The alignment was generated by the program MULTALIN [Corpet, 1988] and the figure was prepared using ESPript [Gouet et al., 1999]. The secondary structure of U2AF^{35}, derived from three-dimensional data [Kielkopf et al., 2001], is represented in the upper part of the alignment.

kDa protein bearing strong sequence similarity to U2AF^{35}; the N-terminal 187 amino acids are 89% identical, but the C-terminus of U2AF^{26} lacks the RS domain present in U2AF^{35} [Shepard et al., 2002]. U2AF^{35}-R1 and U2AF^{35}-R2/Urp are 94% identical to one another and contain stretches that are approximately 50% identical to corresponding regions of U2AF^{35} [Kitagawa et al., 1995; Tronchere et al., 1997]. Additional sequences encoding putative new proteins related to U2AF^{35} were identified in the draft of the human genome [Tupler et al., 2001; Barbosa-Morais et al., 2006], but these have not yet been validated experimentally.

3.3 The evolution of U2AF genes

Phylogenetic analysis indicates that the origin of U2AF gene families falls in the roots of eukaryotes, more than 1500 million years ago [Barbosa-Morais et al., 2006]. Orthologs of both U2AF^{65} and U2AF^{35} are found in *Drosophila melanogaster* [Kanaar et al., 1993; Rudner et al., 1996], *Caenorhabditis elegans* [Zorio and Blumenthal, 1999a; Zorio and Blumenthal, 1999b], *Schizosaccharomyces pombe* [Potashkin et al., 1993; Wentz-Hunter and Potashkin, 1996], *Arabidopsis thaliana* [Domon et al., 1998], and *Plasmodium falciparum* [Barbosa-Morais et al., 2006]. In contrast, the genome of *Saccharomyces cerevisiae* contains a poorly conserved ortholog of the U2AF large subunit, Mud2p, and no open reading frame that resembles the small subunit [Abovich et al., 1994]. Orthologs of human PUF60 are present across metazoans, while CAPER proteins are found all across the eukaryotic lineage. Orthologs of U2AF^{35}R2/Urp exist in insects, chordates and vertebrates (Figure 3.4).

Phylogenetic studies show that both the U2AF^{35} and CAPER genes were duplicated most likely during the wave of whole-genome duplications that occurred at the early emergence of vertebrates 650-450 million years ago, giving rise to U2AF^{26} and CAPERβ, respectively. Orthologs of either U2AF^{26} or CAPERβ are not detected in lower eukaryotes like *Drosophila*, *C. elegans* or plants. Intriguingly, these two genes were apparently lost in some vertebrate lineages and remained in others (Figure 3.4). Orthologs of U2AF^{26} are present in the human and mouse genomes and ESTs more similar to U2AF^{26} than U2AF^{35} are found in rat, pig, and cow. However, there is

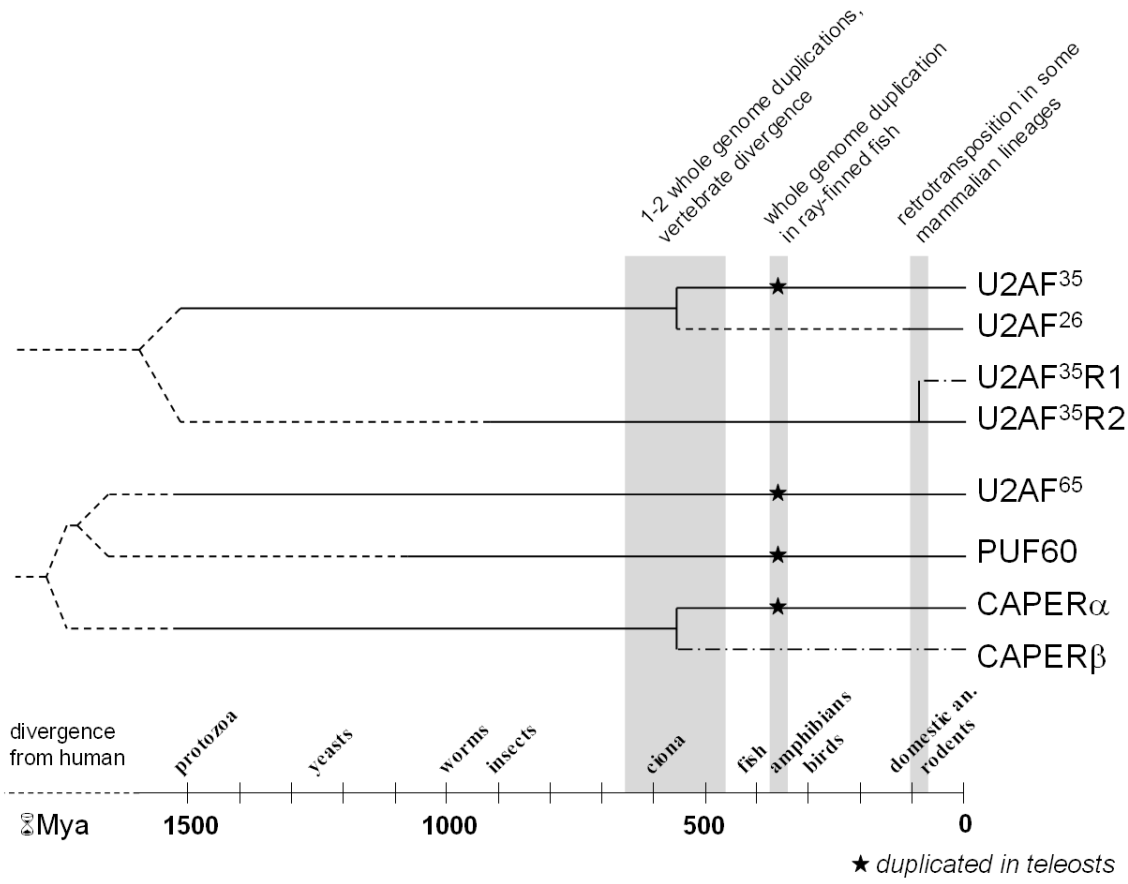

Figure 3.4: Evolution of U2AF-related proteins

The possible origins of U2AF proteins are shown in relation to key metazoan evolutionary events. Solid lines represent presence of the indicated protein in all species that diverged from human within the corresponding period of time. Dashed lines represent loss of the indicated proteins in all extant species that diverged from human within the corresponding period of time. Dashed-dotted lines represent lineage specific loss/preservation or appearance/absence of the indicated protein amongst species that diverged from human within the corresponding period of time (e.g. CAPERβ apparently disappeared from fish, birds and rodents but remained in *Xenopus* and some mammals, U2AF^{35}R1 results from independent retrotransposition events affecting only primates and rodents). A star indicates that U2AF^{35}, U2AF^{65}, PUF60 and CAPERα genes are duplicated in teleosts, most probably as a consequence of the whole-genome duplication that occured in ray-finned fish ~350 million years ago (Mya).

no evidence for the existence of the gene encoding U2AF^{26} in the genomes of birds, amphibians or fish. A comparison of the mouse and human *U2AF1L4* genes revealed that the exon-intron boundaries are located in the same positions as in the human *U2AF1* gene, although the introns are much smaller in the *U2AF1L4* gene. In addition, the exon sequences of the human and mouse *U2AF1L4* genes are 90% identical at the nucleotide level; the majority of the differences are neutral, third position changes [Shepard et al., 2002]. The evolutionary pattern for CAPERβ is more peculiar. Amongst mammals, orthologs can be found for primates (chimp and rhesus) and domestic animals (dog and cow) but not for rodents. CAPERβ can also be found in *Xenopus tropicalis* but there is no evidence for its existence in chicken or fish. A comparison of *CAPERβ* genes from different mammals revealed that most of the exon-intron boundaries are located in the same positions as in the human *CAPERα* gene and the introns are found to be smaller in the CAPERβ gene. Given this analogy of evolutionary behaviour between the *U2AF^{26}* and *CAPERβ* genes, it is likely that these new splicing proteins perform unique and lineage-specific functions.

Retrotransposition rather then gene duplication appears to have originated gene *U2AF1L1* less than 100 million years ago. The mouse *U2AF1L1* gene located on chromosome 11 was formed by retrotransposition of *U2AF1L2* on the X chromosome [Nabetani et al., 1997]. *U2AF1L1* is regulated by genomic imprinting [Hayashizaki et al., 1994], and the whole gene is located in an intron of another gene, *Murr1*, that is not imprinted [Nabetani et al., 1997]. The transposition that originated the mouse *U2AF1L1* gene must have occurred after mice and humans diverged, because the human ortholog of *Murr1* is located on chromosome 2 and there are no *U2AF1*-related genes on human chromosome 2. Indeed, the phylogenetic analysis of this family of genes indicates independent events of retrotransposition in rodents (mouse and rat) and primates (human and chimp). Similarly to the mouse gene, the human *U2AF1L1* gene located on chromosome 5 is intronless while human *U2AF1L2* is multiexonic, suggesting that it has also originated by retrotransposition [Barbosa-Morais et al., 2006]. However, by contrast to the mouse gene, human *U2af1-rs1* is not imprinted [Pearsall et al., 1996].

3.4 Alternative splicing and diversity of human U2AF proteins

Our laboratory has recently reported that human transcripts encoding U2AF^{35} can be alternatively spliced giving rise to three different mRNA isoforms called U2AF^{35}a, U2AF^{35}b, and U2AF^{35}c [Pacheco et al., 2004]. This discovery raised the question of whether additional U2AF genes produce alternatively spliced mRNAs. Very few examples of U2AF mRNA isoforms have been described in the literature. Namely, two CAPERβ mRNAs and four CAPERα mRNAs were detected in several human tissues by Northern blotting [Dowhan et al., 2005], and a splicing variant of PUF60/FIR was identified in colorectal cancers [Matsushita et al., 2006]. This scarcity of published data prompted us to use bioinformatics search strategies to review alternative splicing of U2AF and U2AF-related genes in existing databases. The revision was carried out with the aid of the UCSC Genome Browser [1] [Kent et al., 2002] for the human genome assembly hg17, May 2004, NCBI Build 35. The gene region was defined by the BLAT mapping [Kent, 2002] of the available RefSeq [2]transcript (RNA) sequences [Pruitt et al., 2005], , for a particular gene. Using the UCSC Table Browser [Karolchik et al., 2004], the tables for the BLAT mappings of cDNAs (from RefSeq) and expression sequence tags (ESTs) were obtained for this gene region. Making allowance only for GT_AG, GC_AG or AT_AC splice site consensus and excluding isoforms with extensive intron retentions, the non redundant set of longest isoforms and corresponding accessions was determined. The splicing patterns obtained were crosschecked with two alternative splicing databases: the ASAP [3]; and the Hollywood RNA Alternative Splicing Database [4].

Our analysis revealed that with the single exception the *U2AF1L1* gene, which is devoid of introns, all genes coding for U2AF and U2AF-related proteins can be alternatively spliced (Table 3.3). Many alternatively spliced mRNA isoforms are predicted to contain premature stop codons and are therefore expected to be targeted

[1] `http://genome.ucsc.edu/`
[2] `http://www.ncbi.nlm.nih.gov/projects/RefSeq/`
[3] `http://bioinfo.mbi.ucla.edu/ASAP/`
[4] `http://hollywood.mit.edu`

for degradation by non-sense mediated decay, as already demonstrated for U2AF^{35}c (corresponding to RefSeq mRNA NM_001025204 in Table 3.3). Additionally, we found evidence for several transcripts that could generate functional protein isoforms containing the conserved RRM motifs characteristic of each protein family (Table 3.3). However, further studies are needed to experimentally validate the existence and specific roles of these putative new human proteins.

Protein (Gene Symbol)	Confirmed mRNA isoforms (Accessions)	Predicted splicing patterns producing a premature stop codon (Accessions)	Predicted splicing patterns of candidates for putative novel protein (Accessions)
U2AF^{65} (U2AF2)	**2** (NM_007279.2, NM_001012478.1)	**2** (CD624005.1, CR982513.1, CA488904.1)	**2** (CR609498.1, BI909492.1)
PUF60 (SIAHBP1)	**4** (NM_014281.3, NM_078480.1, BC009734.1, BC011265.1)	**0**	**10** (BI915396.1, AL522753.3, AL514886.3, BX384203.2, AK055941.1, BQ421738.1, BQ956878.1, BG115238.1, BE393389.1, BU170641.1)
CAPERα (RNPC2)	**5** (NM_184234.1, NM_004902.2, NM_184241.1, NM_184244.1, NM_184237.1)	**5** (NM_184241.1, NM_184244.1, NM_184237.1, BC107886.1, BM468718.1, BE816688.1, DA115481.1, AL711019.1, CA419145.1, DA372839.1, BP352717.1, DB027200.1, DB150523.1, BG764840.1, DA922841.1, AW993266.1, AL513896.3)	**10** (BC107886.1, AL833168.1, BP352717.1, BX483043.1, BQ893325.1, CR995560.1, BQ954122.1, BE933146.1, BM983358.1, BU075848.1, DB023865.1)
CAPERβ (RBM23)	**4** (NM_018107.3, CR595426.1, BX161440.1, AL834198.1)	**10** (DA821789.1, DB164369.1, BM464794.1, DA145418.1, BI823680.1, DB166416.1, AA633094.1, BI915247.1, DA299707.1, DA026292.1, CN483101.1, CX165727.1, BC106012.1)	**8** (DA675412.1, BG033916.1, DA117163.1, DA311282.1, BQ707907.1, BQ071908.1, BX388764.2, BI915247.1, DA299707.1, DA026292.1, CN483101.1, CX165727.1,BC106012.1)
U2AF^{35} (U2AF1)	**3** (NM_006758.2, NM_001025203.1, NM_001025204.1)	**2** (NM_001025204.1, BE736536.1)	**1** (BG612658.1)
U2AF^{26} (U2AF1L4)	**2** (NM_144987.2, NM_001040425.1)	**6** (BM696851.1, BM970675.1, AW274826.1, DB127360.1, BU628789.1, AA455588.1, BI770029.1, BC010865.1, BG481735.1, W51842.1)	**6** (BE856544.1, BM696851.1, BM970675.1, AW274826.1, DB127360.1, BU628789.1, AA455588.1, BU608847.1, DB338076.1, BF821614.1)
U2AF^{35}R2 (U2AF1L2)	**1** (NM_005089.2)	**6** (BC065719.1, DA173194.1, DA383795.1, CN289520.1, BE619312.1, DA261525.1, CA425173.1)	**0**

Table 3.3: Alternative splicing of U2AF and U2AF-related transcripts

An alternatively spliced mRNA isoform was considered confirmed if its corresponding protein sequence is annotated in RefSeq or SwissProt databases. A splicing pattern observed in an mRNA or EST was predicted to produce a premature termination codon if it contained an inframe stop signal within an internal exon. For the predicted patterns of splicing there is redundancy in the number of accessions shown due to the fragmented nature of ESTs and some mRNAs.

3.5 Perspectives: evolution of U2AF functions

Following the discovery that U2AF^{65} is required to reconstitute mammalian splicing *in vitro* [Ruskin et al., 1988; Zamore and Green, 1989], the protein was shown to

be highly conserved and its homologues are essential in *Schizosaccharomyces pombe* [Potashkin et al., 1993], *Drosophila melanogaster* [Kanaar et al., 1993] and *Caenorhabditis elegans* [Zorio and Blumenthal, 1999a]. While it remains an open question whether U2AF^{65} performs other functions in the cell in addition to its fundamental role in pre-mRNA splicing, the U2AF^{65}-related proteins are clearly implicated in both splicing and transcription. In particular, CAPER (also known as CC1.3) was independently identified as a protein that interacts with the estrogen receptor and stimulates its transcriptional activity [Jung et al., 2002], and purified as a spliceosome component capable of affecting the splicing reaction [Rappsilber et al., 2002; Hartmuth et al., 2002; Auboeuf et al., 2004]. More recently an additional related protein was identified, CAPERβ, and both CAPER (renamed CAPERα) and CAPERβ were shown to regulate transcription and alternative splicing in a steroid hormone-dependent manner [Dowhan et al., 2005]. Importantly, both CAPERα and CAPERβ are expressed at higher levels in the placenta and liver, two tissues with active steroid hormone signalling. According to one possible model, the CAPER proteins interact first with transcription factors to stimulate transcription in response to steroid hormones; by interacting with promoter bound transcription factors the CAPER proteins can be incorporated into the pre-initiation complex and thereby have direct access to the nascent RNA transcript; the CAPER proteins may then interact with splicing factors required for early recognition of the 3' splice site and thereby influence the commitment for splicing [Dowhan et al., 2005].

PUF60 was originally identified as a pyrimidine-tract binding protein that is required, together with U2AF, for efficient reconstitution of RNA splicing *in vitro* [Page-McCaw et al., 1999]. In the meantime, the human protein was identified as a modulator of TFIIH activity and named FIR [Liu et al., 2000b]. An interaction between PUF60/FIR (FUSE-binding protein-interacting repressor) and the TFIIH/p89/XPB helicase was found to repress *c-myc* transcription, and enforced expression of FIR induced apoptosis. Interestingly, a splicing variant of FIR was detected in human primary colorectal cancers and recent data suggests that this variant may promote tumor development by disabling FIR repression of *c-myc* and opposing apoptosis [Matsushita et al., 2006]. Unlike the CAPER proteins, PUF60/FIR (similarly to U2AF^{65}) is ex-

pressed in most tissues [Dowhan et al., 2005], as predicted for a constitutive splicing factor. Yet, the *Drosophila* ortholog of human PUF60, *Half Pint*, was found to function in both constitutive and alternative splicing *in vivo* [Van Buskirk and Schupbach, 2002], raising the question of whether human PUF60 regulates alternative splicing. It is also unknown whether the dual function of PUF60 on transcription and splicing is coupled as in the case of the CAPER proteins or whether PUF60 affects independently the transcription and splicing of distinct genes. Although answers to these and other questions are likely to provide new clues to understanding the functional diversity of U2AF^{65}-related proteins, we may speculate that these proteins evolved in response to the needs of coordinating the multiple steps of gene expression in complex organisms. As mRNA biogenesis became progressively more targeted for regulation, new sequence characteristics developed to allow the same molecule to engage in sequential transcriptional and splicing events, acting as coupling proteins in regulated gene expression.

Contrasting to U2AF^{65}-related proteins, there is no evidence implicating the U2AF^{35}-like proteins in any process other than splicing. Unlike U2AF^{65}, which is essential for splicing, U2AF^{35} is dispensable for *in vitro* splicing of some model pre-mRNAs containing strong Py tracts (i.e., a stretch of pyrimidines beginning at position -5 relative to the 3' splice site and extending 10 or more nucleotides upstream into the intron [Burge et al., 1999]). The presence of U2AF^{35} and its interaction with U2AF^{65} was however found essential for *in vitro* splicing of a pre-mRNA substrate with a Py tract that deviates from the consensus [Guth et al., 1999]. Introns with nonconsensual or weak Py tracts were previously called 'AG-dependent' [Reed, 1989]. Biochemical complementation experiments performed with extracts depleted of endogenous U2AF demonstrated that splicing of AG-dependent introns was rescued only when both U2AF subunits were added and not with U2AF^{65} alone [Zuo and Maniatis, 1996; Guth et al., 1999; Wu et al., 1999].

The importance of the small subunit of U2AF *in vivo* was first shown by the finding that the fruit fly *Drosophila melanogaster* ortholog of human U2AF^{35} (dU2AF^{38}) is essential for viability [Rudner et al., 1996]. Orthologs of U2AF^{35} are also essential for the viability of the fission yeast *Schizosaccharomyces pombe* [Wentz-Hunter and

Potashkin, 1996] and the nematode *Caenorhabditis elegans* [Zorio and Blumenthal, 1999b] and for the early development of zebrafish [Golling et al., 2002]. Additional studies in both *Drosophila* and human cells further provided hints of a role for U2AF^{35} in splicing regulation. First, loss-of-function mutations in dU2AF^{38} affected splicing of the pre-mRNA encoding the female-specific RNA binding protein Sex-lethal [Nagengast et al., 2003]. Second, depletion of dU2AF^{38} by RNA interference (RNAi) affected alternative splicing of the *Dscam* gene transcript [Park et al., 2004]. Third, RNAi-mediated depletion of both U2AF^{35}a and U2AF^{35}b isoforms in HeLa cells altered alternative splicing of Cdc25 transcripts [Pacheco et al., 2006].

Sequence comparisons of U2AF^{35} splicing isoforms and U2AF^{35}-related proteins revealed a striking conservation of the principal signature features of UHMs (Figure 3.3). Moreover, there is biochemical evidence indicating that both U2AF^{35}a and U2AF^{35}b splicing isoforms, U2AF^{26} and U2AF^{35}R2/Urp can interact with U2AF^{65} [Tronchere et al., 1997; Shepard et al., 2002; Pacheco et al., 2004]. U2AF^{35}R2/Urp was further shown to be functionally distinct from U2AF^{35} because U2AF^{35} cannot complement Urp-depleted extracts [Tronchere et al., 1997]. It was therefore proposed that the U2AF^{65} subunit may form diverse heterodimers with the different U2AF^{35}-like proteins, each of them with distinct functional activities. In this regard it is noteworthy that splicing isoform U2AF^{35}a is 9- to 18-fold more abundant than U2AF^{35}b, with distinct tissue-specific patterns of expression [Pacheco et al., 2004], and in the mouse, the *U2AF1L1* gene is expressed predominantly in the brain specially in the pyramidal neurons in the hyppocampus and dental gyrus [Hatada et al., 1993; Hatada et al., 1995]. Identifying the functional uniqueness of each U2AF^{35} protein isoform is clearly an important challenge for future research.

3.6 Concluding remarks

New biological functions are generally acquired through evolutionary redundancy provided by distinct mechanisms. Both the emergence of additional genomic copies by gene duplication and retrotransposition, and an increase in transcript diversity by alternative splicing have contributed to generate new U2AF-related proteins. The

similarity and differences between the U2AF-related proteins imply that they have evolved distinct functions in relation to control of gene expression in complex organisms. Clues to the biological processes in which these proteins participate may be obtained by determining their tissue expression patterns, elucidating their RNA binding specificities and identifying the genes that they control. Ultimately, understanding the function of the diverse U2AF proteins will require deciphering their roles in shaping human development and physiology.

Chapter 4

Recognition of splicing *cis* elements and applications

4.1 Identification of splicing regulatory motifs

As described in 1.1.5, many sequence elements, usually comprising binding sites for splicing factors, act as *cis* regulators of alternative splicing. Over the last decade, there has been a strong effort and some progress in identifying those motifs as an important contribution for the understanding of the mechanisms of alternative splicing.

4.1.1 Experimental and computational approaches

In vitro SELEX (systematic evolution of ligands by exponential enrichment) experiments have been important in revealing the main features of binding sites for several splicing factors [Matlin et al., 2005]. Binding SELEX is an iteractive method for the identification of optimal binding sites for RNA-binding proteins, like SR proteins and hnRNPs. From an initial pool of a random and degenerate cDNAs, sequences undergo several rounds of transcription, protein binding and amplification by RT-PCR until the emergence of a consensus sequence. Although variable, optimal binding sites for SR proteins have been found to generally correspond to canonical purine-rich ESEs. Optimal binding sites for hnRNPs known to be repressors resemble known splicing silencers.

Both *in vitro* and *in vivo* functional SELEX has also revealed different ESEs, including a class of AC-rich elements [Coulter et al., 1997]. Functional SELEX uses a minigene containing a sequence element known to regulate (usually enhance) the splicing of its pre-mRNA. This element is replaced by random sequences and the resulting pool of minigenes is transcribed *in vitro* or transfected into cultured cells, generating a pool of pre-mRNAs. After splicing, the resulting mRNAs are purified and amplified by RT-PCR. The pool of spliced mRNAs, enhancer-enriched, is used to reconstruct new minigenes and the cycle is iteractively repeated, producing sequence "winners" that are supposed to have outstanding splicing enhancing action [Cartegni et al., 2002]. A refinement of this approach, named fluorescence-activated screen for exonic splicing silencers [Wang et al., 2004b], recently allowed the identification of many ESSs, some of them resembling binding sites for hnRNPs. The technique could also be applied to intronic elements [Matlin et al., 2005].

Other techniques have been used to define binding sites. For example, immunoprecipitation of RNA binding proteins from polysomes, followed by RT-PCR and library screening, was shown to successfully identify the in vivo mRNA ligands of RNA binding proteins [Brooks and Rigby, 2000].

Data from SELEX experiments can get statistical treatment and nucleotide scoring matrices have been widely used for ESE prediction. Positive correlation between predicted ESE motifs in natural genes and SR protein specificity of the corresponding pre-mRNAs have been shown [Liu et al., 2000a]. Moreover this approach allowed the successful prediction of mutations that, by disrupting ESEs, can alter splicing and cause disease [Cartegni and Krainer, 2002; Cartegni et al., 2002].

In 2003, the Krainer lab finally released ESEfinder [1], a web-based resource that searches sequences for putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55 [Cartegni et al., 2003]. Its search algorithms are based on the statistical features of motifs obtained from functional SELEX experiments (Figure 4.1). This tool was also designed to predict whether exonic mutations disrupt such regulatory elements.

However, until 2002 there was no published computational tool designed to in-

[1] `http://rulai.cshl.edu/tools/ESE/`

tegrate all the available experimental information and search query sequences for binding motifs. That gap led me to develop a program to predict putative binding sites for SR proteins and hnRNPs, named Splicing Rainbow (described in 4.1.2) due to the color code associated to its output.

Purely computational approaches for motif identification eventually started to emerge. The RESCUE-ESE (relative enhancer and silencer classification by unanimous enrichment) method [Fairbrother et al., 2002] identified 10 predicted human ESE motifs by clustering hexamers that were enriched in exons versus introns and in weak splice site exons versus strong splice site exons. Representatives of the motifs displayed enhancer activity *in vivo*, whereas point mutants of these sequences showed reduced activity. This approach allowed successful prediction of the splicing phenotypes of exonic mutations in human genes. The same method was later applied to a broader range of vertebrates and identified vertebrate-specific ESEs and ISEs [Yeo et al., 2004]. There is also an online ESE analysis tool that annotates RESCUE-ESE hexamers in vertebrate exons and can be used to predict splicing phenotypes by identifying sequence changes that disrupt or alter predicted ESEs [2] [Fairbrother et al., 2004].

A different method, that avoids protein-coding biases, was used to compare the frequency of octamers in internal noncoding exons versus unspliced pseudo exons and 5' UTRs of transcripts of intronless genes [Zhang and Chasin, 2004]. Representa-

[2] Available on `http://genes.mit.edu/burgelab/rescue-ese/` .

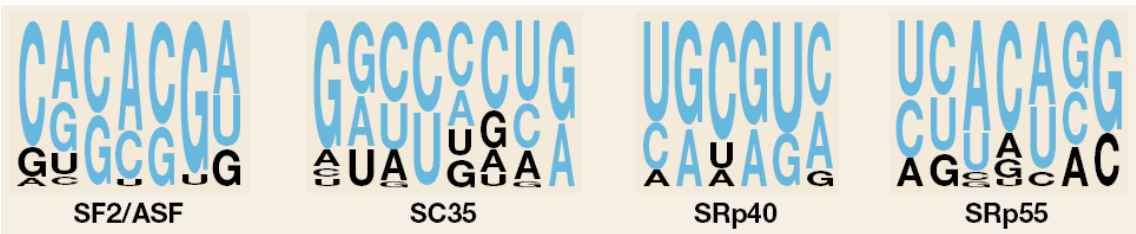

Figure 4.1: Pictograms of functional-SELEX consensus ESE motifs for SR proteins
The height of each letter reflects the frequency of each nucleotide at a given position, after adjusting for background nucleotide composition (blue letters indicate above-background frequencies). (Adapted from [Cartegni et al., 2002].)

tives of each class of motifs found functioned as enhancers or silencers when inserted into a test exon and assayed in transfected mammalian cells. There was significant resemblance between these and the RESCUE-ESE motifs [Fairbrother et al., 2002].

Some other computational analyses specifically focused on alternative exons, namely those associated with tissue specific isoforms. These approaches are still limited by the uncompleteness of alternative splicing events databases but extensive datasets should soon be provided by alternative splicing microarrays [Matlin et al., 2005].

4.1.2 The Splicing Rainbow

I created the Splicing Rainbow in 2002 in the Valcárcel lab at EMBL (European Molecular Biology Laboratory), Heidelberg. It was designed to predict putative binding sites for splicing factors, namely SR proteins and hnRNPs, and to display them in a 'biologist-friendly' way.

The compilation of a exhaustive list of binding sites for splicing factors involved an extensive bibliographic search. More than 50 sequence motifs, for 24 splicing factors, were annotated from a similar number of articles.

Results from SELEX experiments can be statistically addressed through scoring matrices. For each N-mer analysis, the SELEX data is used to calculate a frequency matrix $f_i(a)$, where i is the position of nucleotide a. The scoring matrix is defined by the following formula [Liu et al., 1998]:

$$s_i(a) = log_2 \frac{f_i(a) + \epsilon p(a)}{p(a)(1+\epsilon)} \tag{4.1}$$

where $p(a)$ is the background frequency (when not furnished we take $p(a)=0.25$ for the 4 nucleotides) and $\epsilon=0.5$ is the Bayesian prior parameter [Lawrence et al., 1993]. For each N-mer starting in nucleotide j we take:

$$S_{Nj} = \sum_{i=j}^{j+N} s_i(a) \tag{4.2}$$

A threshold score S_T is defined and the N-mer is considered a putative binding site if $S_{Nj}>S_T$. For most motifs, there was no experimental validation of threshold and the definition of S_T followed 'common sense' criteria. In general, S_T was chosen

as the highest score that would allow the prediction of a binding site in each of the SELEX sequences. This criterium leads to relatively low thresholds. Despite the theoretical abundance of false positives, they were kept to avoid the lost of potential binding sites.

Sometimes the published motif is well defined and non-degenerate and the program just looks for exact matches or, for some cases, the regular expressions are configured to allow a small number of mismatches. Other motifs have definitions that are as ambiguous as 'poly-G' and *ad-hoc* thresholds were chosen, requiring a minimum number of the relevant nucleotide per N-mer.

For all the motifs, criteria and references, see tables A.8, A.9 and A.10.

The Splicing Rainbow is a Perl script, as Perl is an interpreted programming language with a special vocation for text parsing and regular expressions. The input for the program was made as simple as possible. The user is required to provide the query sequence in a FASTA-format file and optional information about the gene structure (based on transcript data) in an EMBL-format file (Figure 4.2).

A color code was defined for the program's output. 'Cold' colors identify putative binding sites for hnRNPs and 'hot' colors represent motifs for SR proteins. The program generates an HTML file (Figure 4.3) where the putative binding sites can be visualized for each factor in separate lines to avoid 'saturation'. The file also includes a header with links for information about criteria and references.

Additionally the Splicing Rainbow generates an EMBL-format file that can be opened with Artemis [Berriman and Rutherford, 2003; Rutherford et al., 2000] (Figure 4.4), an interactive sequence viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence.

A simple tab-delimited text file with the results is also generated (Figure 4.5).

A brief tutorial for the program was created and can be found in A.2.2.

4.1.3 ASD Workbench

The Alternative Splicing Database (ASD) Project [Thanaraj et al., 2004] has been launched with the purpose of creating, maintaining and developing a value-added database of alternative splice events and the resultant isoform splice patterns of genes

A

```
>FAS
ATGTGAACATGGAATCATCAAGGAATGCACACTCACCAGCAACACCAAGTGCAAAGAGGAAGGTAATTATTTTTTTACGGTTATATTCTCCTTTCCCCCAACCCCATGGAAAGATGTGAAGAAAA
```

B

```
ID   Fas   277 BP.
XX
CC   CNSLT0M0V: trimmed hsp from 688 to 687 by g and from 686 to 688 by ag
CC   AV715411: trimmed hsp from 413 to 412 by g and from 411 to 413 by ag
CC   BM455788: trimmed hsp from 679 to 678 by g and from 677 to 679 by ag
CC   BI838027: trimmed hsp from 655 to 654 by g and from 653 to 655 by ag
CC   BI766250: trimmed hsp from 517 to 516 by g and from 515 to 517 by ag
CC   BI463384: No match to query for internal bases 672 to 670. Padded by Xs.
XX
FT   mRNA            join(1..62,215..277)
FT                   /note="EST CNSLT0M0V 1..62,215..277"
FT                   /colour=255 50 50
FT   mRNA            join(1..62,215..277)
FT                   /note="EST AV715411 1..62,215..277"
FT                   /colour=255 50 50
FT   mRNA            join(1..62,215..277)
FT                   /note="EST BM455788 1..62,215..277"
FT                   /colour=255 50 50
FT   mRNA            join(1..62,215..277)
FT                   /note="EST BI838027 1..62,215..277"
FT                   /colour=255 50 50
FT   mRNA            join(1..62,215..273)
FT                   /note="EST BI766250 1..62,215..273"
FT                   /colour=255 50 50
FT   mRNA            217..277
FT                   /note="EST AV651157 217..277"
FT                   /colour=255 50 50
FT   mRNA            221..277
FT                   /note="EST BE070451 221..277"
FT                   /colour=255 50 50
FT   mRNA            join(1..60,214..268)
FT                   /note="EST BI463384 1..60,214..268"
FT                   /colour=255 50 50
FT   mRNA            213..269
FT                   /note="EST BF126149 213..269"
FT                   /colour=255 50 50
FT   mRNA            217..273
FT                   /note="EST BI254532 217..273"
FT                   /colour=255 50 50
FT   mRNA            1..35
FT                   /note="EST BG540571 1..35"
FT                   /colour=255 50 50
XX
SQ   Sequence 277 BP.
     ATGTGAACAT GGAATCATCA AGGAATGCAC ACTCACCAGC AACACCAAGT GCAAAGAGGA        60
     AGGTAATTAT TTTTTTACGG TTATATTCTC CTTTCCCCCA ACCCCATGGA AAGATGTGAA       120
     GAAAAACCAA TCACTCTTGA TTACTAGAAA GTCCTTTATT TAATCTTAAA GATTGCTTAT       180
     TTTCATATAA AATGTCCAAT GTTCCAACCT ACAGGATCCA GATCTAACTT GGGGTGGCTT       240
     TGTCTTCTTC TTTTGCCAAT TCCACTAATT GTTTGGG                                277
//
```

Figure 4.2: Input files for Splicing Rainbow

A - FASTA-format file with query sequence; B - EMBL-format file with transcript and gene structure information.

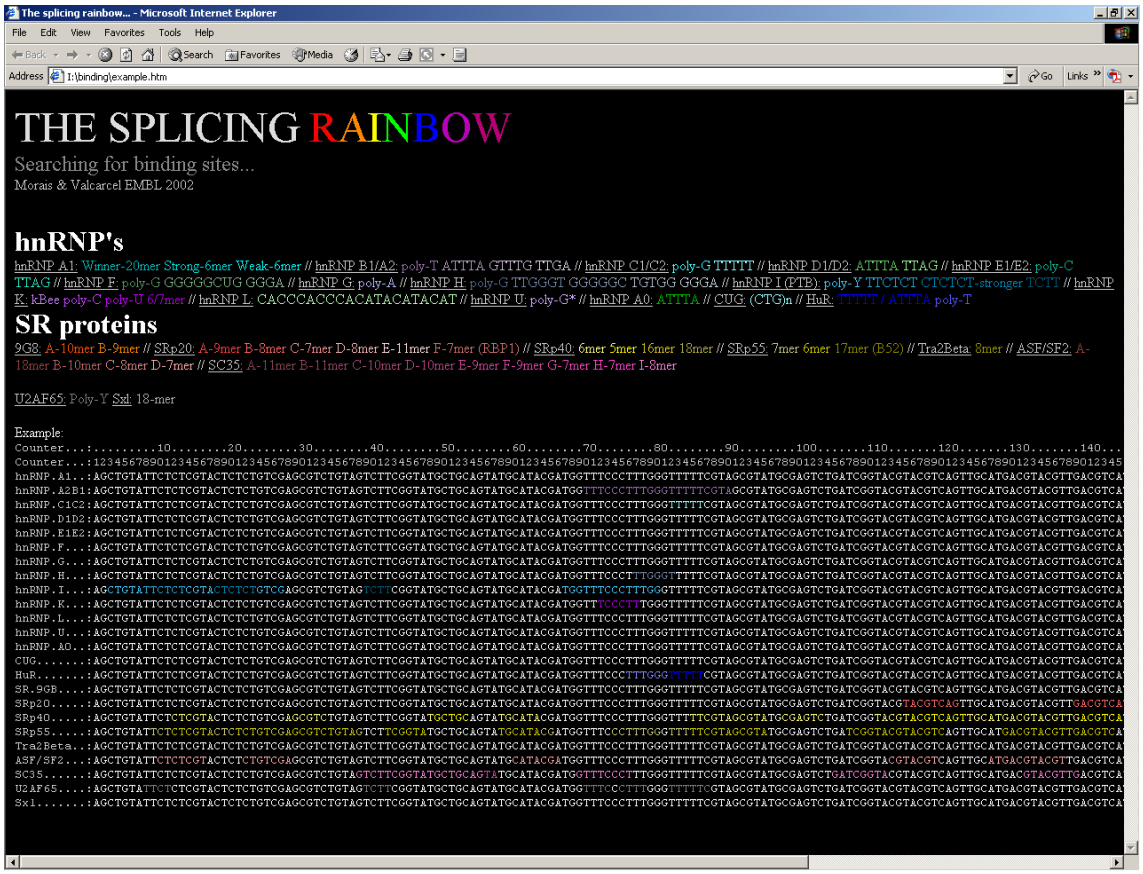

Figure 4.3: HTML output of Splicing Rainbow

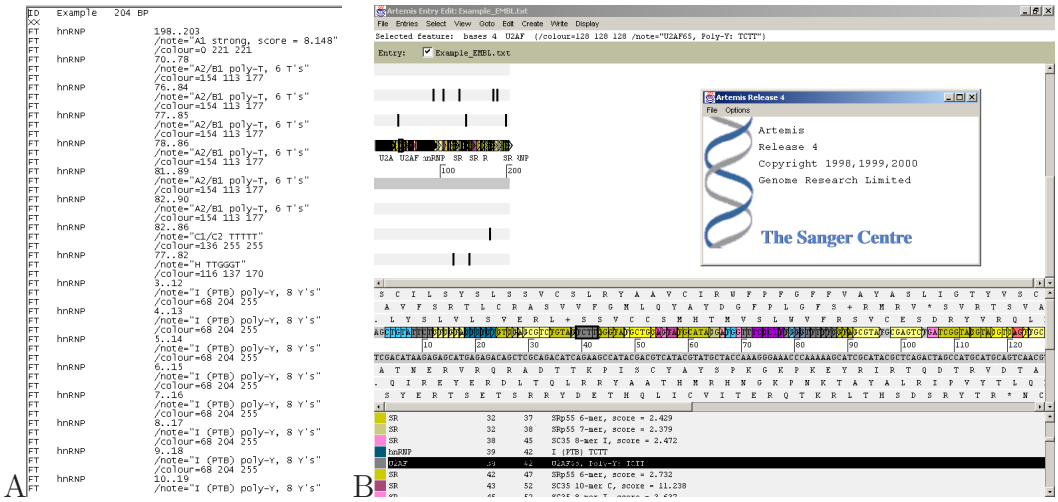

Figure 4.4: Artemis output of Splicing Rainbow

A - EMBL-format file with information on putative binding sites; B - Artemis display.

Sequence name: Example
Length: 204

Nuc_start	Nuc_end	Factor	Consensus_type	Score
198	203	hnRNP A1	Strong UAGGGA/U	8.148
70	78	hnRNP A2/B1	poly-T	6 T's (in 9)
76	84	hnRNP A2/B1	poly-T	6 T's (in 9)
77	85	hnRNP A2/B1	poly-T	6 T's (in 9)
78	86	hnRNP A2/B1	poly-T	6 T's (in 9)
81	89	hnRNP A2/B1	poly-T	6 T's (in 9)
82	90	hnRNP A2/B1	poly-T	6 T's (in 9)
82	86	hnRNP C1/C2	TTTTT	Exact match
77	82	hnRNP H	TTGGGT	Exact match
3	12	hnRNP I (PTB)	poly-Y	8 pyrimidines (in 10)
4	13	hnRNP I (PTB)	poly-Y	8 pyrimidines (in 10)
5	14	hnRNP I (PTB)	poly-Y	8 pyrimidines (in 10)

Figure 4.5: Tabular output of Splicing Rainbow

(from human and other model species) and of experimentally verified regulatory mechanisms that mediate splice variants.

The ASD Project includes a Workbench with online tools relating to alternative splicing (`http://www.ebi.ac.uk/asd-srv/wb.cgi`). The Splicing Rainbow was adapted to an online interactive framework and included in the ASD Workbench. It can be found on `http://www.ebi.ac.uk/asd-srv/wb.cgi?method=8`.

The recent developments in the ASD Project and details on the Workbench have been published [Stamm et al., 2006].

4.2 RNA binding proteins as coordinators of mRNA metabolism

Work from the past few years has begun to reveal mRNA binding proteins as multifunctional entities that act on the mRNA biogenesis pathway from transcription initiation through translation and decay. The polypyrimidine tract binding (PTB) protein was originally identified as an alternative splicing factor but it is also known to be involved in 3' end processing and in the regulation of translation and cytoplasmic localization of some viral and cellular mRNAs. $U2AF^{65}$ is an essential splicing factor, involved in the recognition of introns during the early steps of spliceosome assembly, and is also known to shuttle between the nucleus and the cytoplasm. It has been hypothesized that this shuttling activity may occur in association with a specific subset of mRNA molecules, whose metabolism may be regulated by $U2AF^{65}$.

Association of mRNA binding proteins with mRNA through untranslated sequence elements for regulation (USER codes) has been proposed to constitute a mechanism that allows for the coordination of gene expression at the post-transcriptional level, defining so called post-transcriptional operons [Keene and Tenenbaum, 2002]. This coordination would be expected to be particularly useful in pathways that require rapid activation or shutdown of the expression of specific sets of genes.

Through a genome wide approach coupling RNA immunoprecipitation and microarray analysis, we have identified a subset of mRNA molecules that interact with the pre-mRNA splicing factors $U2AF^{65}$ and PTB, also known to be nucleocytoplasmic

mRNA binding proteins [Gama-Carvalho et al., 2006]. Classification of the mRNAs associated with each protein into Gene Ontology [Ashburner et al., 2000] groups suggests that each protein associates with functionally coherent mRNA populations, supporting a coordinating role in gene expression [3].

To understand whether these RNA populations contain distinctive sequence elements we have performed sequence motif search for consensus U2AF and PTB binding sites in the whole transcript, coding sequence and UTRs and compared to a non-associated mRNA population.

I wrote a Perl script to annotate the mRNA sequences and search them for binding motifs. Each mRNA accession number from the Affimetrix array was used to search UniGene [4] [Wheeler et al., 2003] (v. 176) for a corresponding gene identifier. For each gene, the longest curated transcript available in EMBL [5] [Kanz et al., 2005], GenBank [Wheeler et al., 2003] and RefSeq [6] [Pruitt et al., 2005] databases was retrieved using Bioperl [7] [Stajich et al., 2002] (v.1.4) modules. For each of these transcripts, the annotation of coding sequence and untranslated regions was also performed.

The sequence of these transcripts, corresponding to approximately 80% of the entries from the lists derived from the microarray analysis, was then searched for PTB and U2AF^{65} putative binding sites, using scoring matrices.

The sequence YYYYTCTTYYYY was searched for as a putative motif for PTB [Perez et al., 1997; Singh et al., 1995], using the following scoring matrix (where i is the position of nucleotide a):

[3] We find that U2AF^{65} associated mRNAs are enriched in transcription factors and genes related to transcription regulation and cell cycle regulation. In contrast, a significant proportion of mRNAs enriched in PTB immunoprecipitation experiments encode proteins associated to intracellular transport and cytoplasmic compartments, suggesting that these proteins may be coordinating the metabolism of functionally related subsets of mRNA molecules.

[4] `http://www.ncbi.nlm.nih.gov/UniGene`

[5] `http://www.ebi.ac.uk/embl`

[6] `http://www.ncbi.nlm.nih.gov/RefSeq`

[7] `http://www.bioperl.org`

$s_i(a)$												
a	i											
	1	2	3	4	5	6	7	8	9	10	11	12
T	0.5	0.5	0.5	0.5	1	0	1	1	0.5	0.5	0.5	0.5
C	0.5	0.5	0.5	0.5	0	1	0	0	0.5	0.5	0.5	0.5

Any 12-mer for which

$$S_{12} = \sum_{i=1}^{12} s_i(a) > 5.5 \tag{4.3}$$

was considered a putative binding site for PTB.

For U2AF heterodimer a frequency matrix was derived, based on sequences from a SELEX experiment described in [Wu et al., 1999] (where i is the position of nucleotide a):

$f'_i(a)$														
a	i													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	2	6	1	5	1	1	1	1	4	0	31	0	0	0
C	8	2	5	10	13	16	18	9	9	6	0	0	3	3
G	2	4	1	3	1	2	2	1	9	0	0	31	27	2
T	19	19	24	13	16	12	10	20	9	25	0	0	1	26

The scoring matrix is defined by the following formula [Liu et al., 1998]:

$$s_i(a) = log_2 \frac{f_i(a) + \epsilon p(a)}{p(a)(1+\epsilon)} \tag{4.4}$$

$$f_i(a) = \frac{f'_i(a)}{N_S} \tag{4.5}$$

where *p(a)* is the background frequency (take *p(a)=0.25* for the 4 nucleotides), $\epsilon = 0.5$ is the Bayesian prior parameter [Lawrence et al., 1993] and N_S the number of sequences (in this case 31). For any N-mer we take:

$$S_N = \sum_{i=1}^{N} s_i(a) \tag{4.6}$$

Taking the first 9 positions of the frequency matrix, any 9-mer for which $S_9 > 4.5$ was considered a putative binding site for the $U2AF^{65}$ subunit.

The selected cutoff scores for a positive hit are the highest possible values that produce a Gaussian distribution of the frequency of motifs found in the full length mRNA. Positive hits for binding motifs were scored regarding their location on the coding sequence or untranslated regions. To weight out variations in whole transcript, coding and non-coding region length, the density of putative binding sites per transcript was calculated and used for comparison between the different populations.

Interestingly, we find that the average 3'UTR size of the control (non-associated) populations is 60% smaller than the average size of the U2AF^{65} or PTB associated populations (Figure 4.6). This highly significant difference suggests that these mRNAs are targets for post-transcriptional regulation.

Comparison between the U2AF^{65} and PTB associated mRNAs and the respective control non-associated mRNA population reveals a 1.4 and 1.5 fold increase in the average density of putative binding sites per transcript for the associated proteins (Figures 4.7A and B and 4.8). This difference is highly significant and reflects the clearly distinct frequency distribution of motif densities in the analyzed mRNA populations (Figure 4.9A and E). Analysis of the distribution of putative binding sites in the coding and non-coding regions of the transcript reveals that the highest motif density is found on the 3'UTR (Figure 4.7C and D). However, both the coding sequence and the 3'UTR show significantly a different average motif density between associated and control populations. Considering that the maximum difference between the two populations is found in the analysis of the whole transcript (Figure 4.7), we conclude that both coding and non-coding regions contribute to it. We do not find significant differences in the presence or average density of the searched motifs between associated and control populations for the 5'UTR. Indeed, a large fraction of the transcripts in all populations does not have any putative binding motifs in the 5'UTR, independently of its size (Figure 4.9B and F).

We find that the U2AF and PTB-associated populations are more similar to each other for both motifs searched than to the non-associated population (Figure 4.10). However, this is predicted to occur due to the similar characteristics of the sequence

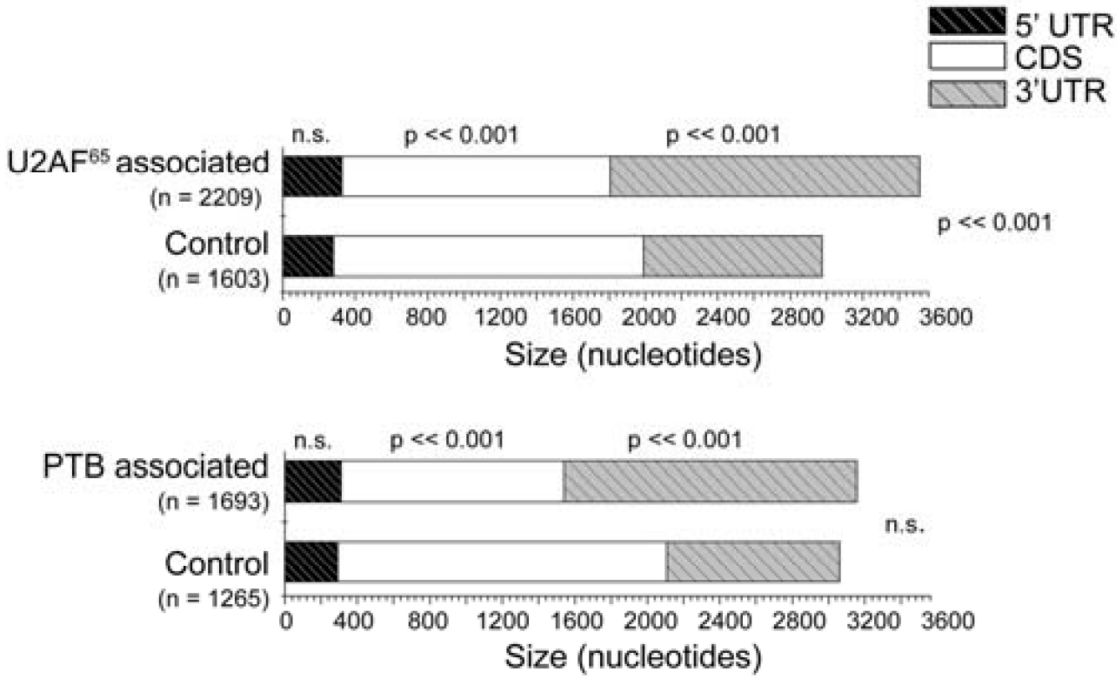

Figure 4.6: Size analysis of coding sequence and untranslated regions of U2AF^{65} and PTB-associated mRNA populations

Average size of 5' and 3' untranslated regions (5' and 3' UTR) and coding sequence (CDS) for mRNAs in the U2AF^{65}-associated or PTB-associated populations and their respective control (not-associated) populations. For this analysis, information for the longest curated transcript available in EMBL [Kanz et al., 2005], GenBank [Wheeler et al., 2003] and RefSeq [Pruitt et al., 2005] databases was retrieved for all entries in each population, when available. Statistically significant differences between the associated and the respective control populations are indicated. n = population size. *n.s.* = not significant.

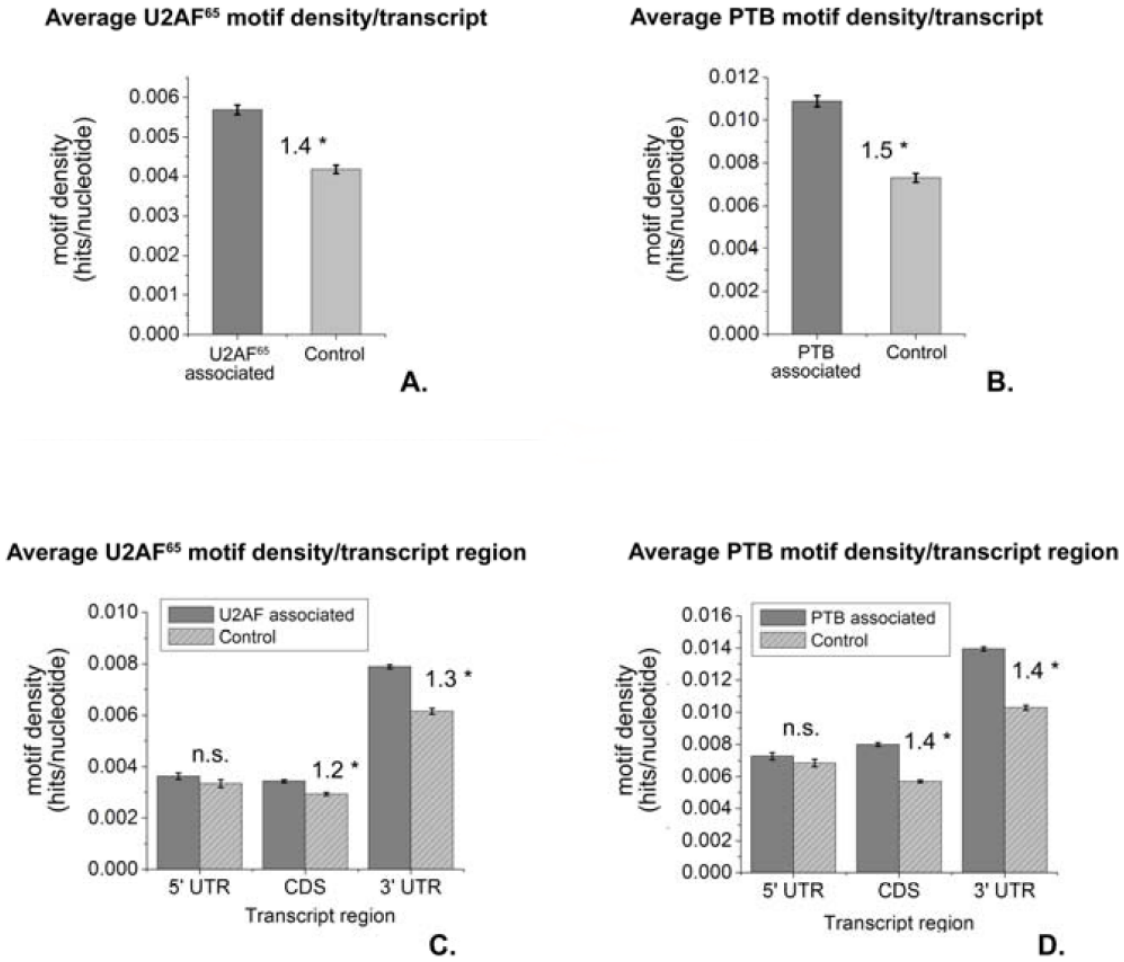

Figure 4.7: Analysis of putative U2AF^{65} and PTB binding motifs in selected mRNA populations

The longest curated transcripts for mRNA accessions in the U2AF^{65}-associated or PTB-associated populations and their respective control (not-associated) populations were searched for putative U2AF65 and PTB binding motifs. **A** - Average U2AF65 motif density in the U2AF^{65}-associated and U2AF^{65}-control populations. **B** - Average PTB motif density the PTB-associated and PTB-control populations. **C** - Average U2AF^{65} motif density by transcript region in the U2AF^{65}-associated and U2AF65-control populations. **D** - Average PTB motif density by transcript region in the PTB-associated and PTB-control populations. The ratio between values for each associated/control pair is shown. * $p \ll 0.001$. *n.s.* = not significant.

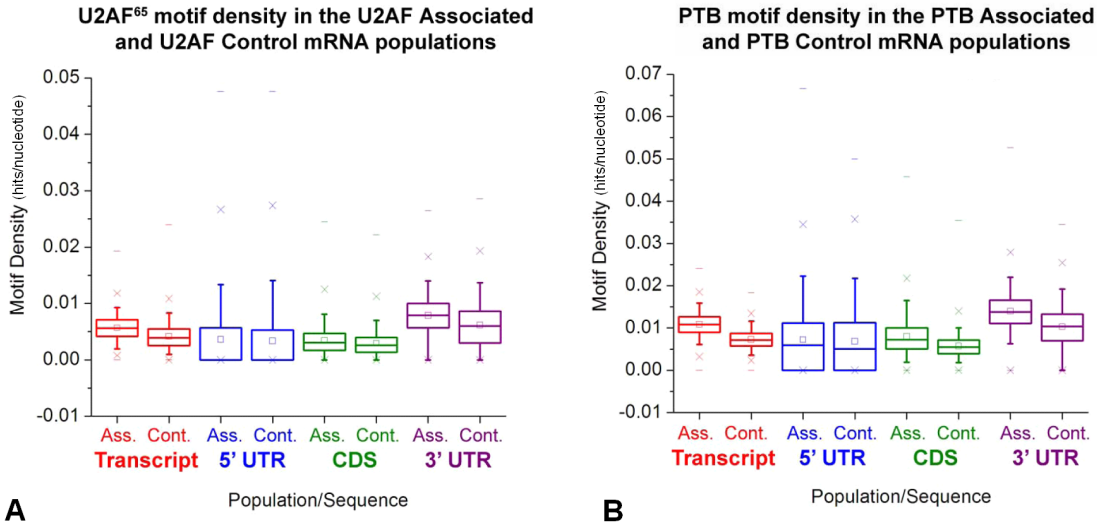

Figure 4.8: Motif density distributions in mRNA populations

Boxplots representing the U2AF^{65} motif density distribution in U2AF associated (Ass.) and control (Cont.) mRNA populations (**A**) and the PTB motif density distribution in PTB associated and control mRNA populations (**B**). Distributions for whole transcript populations are depicted in red, for 5'UTR populations in blue, for coding sequence (CDS) populations in green and for 3'UTR populations in purple.

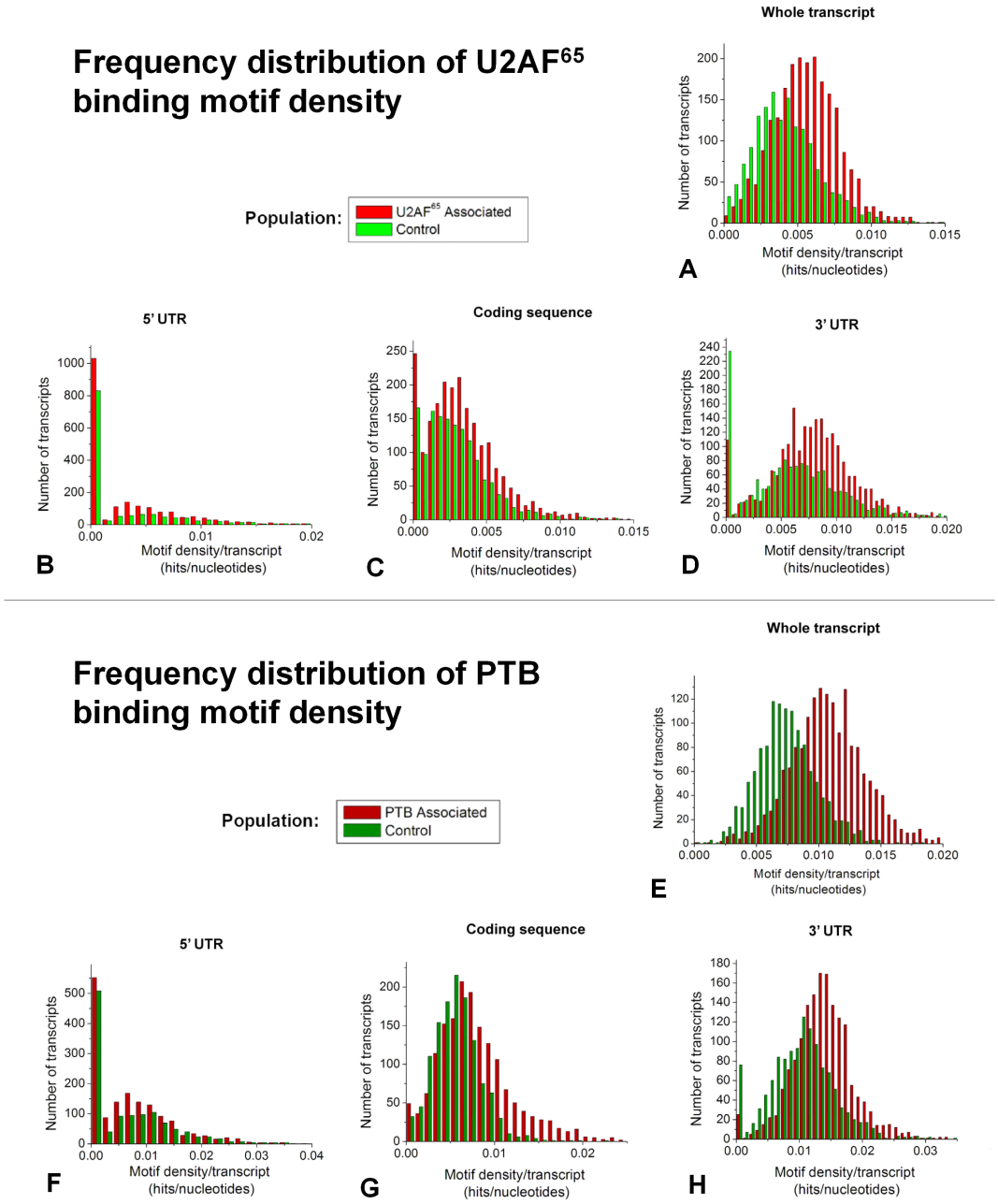

Figure 4.9: Motif frequency distributions in mRNA populations

Histograms representing the U2AF^{65} motif density frequency distribution in U2AF associated (in red) and control (in green) mRNA populations (**A** - whole transcript; **B** - 5'UTR; **C** - coding sequence; **D** - 3'UTR) and the PTB motif density frequency distribution in PTB associated (in dark red) and control (in dark green) mRNA populations (**E** - whole transcript; **F** - 5'UTR; **G** - coding sequence; **H** - 3'UTR)

motifs bound by these proteins. Indeed, we find that an average of 75% of the identified U2AF motifs overlap with PTB motifs whereas only 40% of the PTB motifs are selected as putative U2AF binding sites (data not shown).

Comparison of the U2AF^{65} and PTB-associated mRNA populations revealed that a large proportion of the transcripts are common to both datasets. These transcripts may either correspond to non-specific immunoprecipitation noise or alternatively may be true targets for both proteins. The second possibility is supported by the results obtained in the qRT-PCR quantification of independent immunoprecipitation assays for the GAS2L transcript, which was found to be enriched in the microarray data from both PTB and U2AF^{65} immunoprecipitations. Separate analysis of this dataset revealed an enrichment in putative binding sites comparable to or higher than the one observed for the whole population or for the non-overlapping transcript population, supporting the view that they are targets for interaction with both proteins (Figure 4.10).

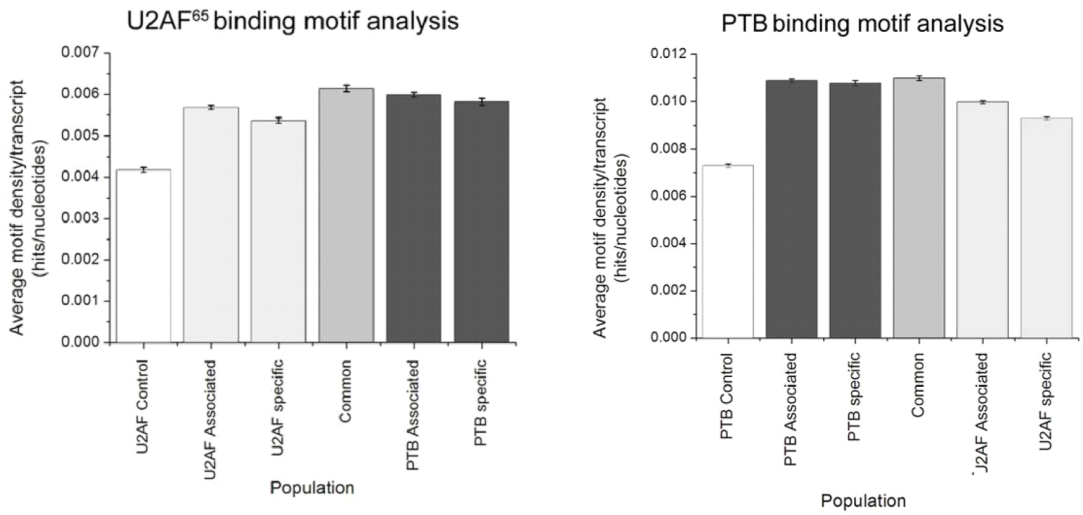

Figure 4.10: Average motif densities in mRNA populations

The search for U2AF and PTB putative binding motifs in the large transcript datasets defined by the genomewide immunoprecipitation studies reveals that the isolated mRNA populations have distinctive sequence characteristics, supporting their predicted association to these proteins. Hence the results support the model of a

differential interaction between functionally related mRNA populations and specific regulatory RNA binding proteins through the presence of USER codes.

In summary, the genomewide analysis of U2AF^{65} and PTB associated mRNAs performed in this work provides strong evidence for new cellular functions for these proteins and more generally, contributes to our understanding of RNA-protein interaction networks that regulate mRNP metabolism and control gene expression at the post-transcriptional level [Gama-Carvalho et al., 2006].

4.3 Alternative splicing regulation and apoptosis

4.3.1 TIA-1, Fas and the regulation of apoptosis

TIA-1 is an RNA-binding protein known to be involved in cell apoptosis [Tian et al., 1991] and implicated in RNA metabolism events occurring both in the nucleus and the cytoplasm [Anderson and Kedersha, 2002]. It comprises three highly similar RNA recognition motifs and a C-terminal glutamine-rich domain [Dember et al., 1996]. TIA-1 acts as a splicing regulator by binding to uridine-rich sequences downstream of weak 5' splice sites and contributing for the recruitment of the U1 snRNP (through protein-protein interactions involving its glutamine-rich domain of TIA-1 and the U1-C protein) [Forch et al., 2002].

Alternative splicing of the human *Fas* gene is regulated by TIA-1 [Forch et al., 2000]. The *Fas* receptor encodes not only a transmembrane protein that mediates apoptosis, upon ligation of the Fas ligand (FasL) [Krammer, 2000], but also soluble forms of the receptor, lacking exon 6 and the encoded transmembrane domain, that can act as inhibitors of Fas signaling [Cheng et al., 1994].

The human *Fas* gene exhibits particular sequence features that may help to explain the regulation of splicing of its exon 6 by TIA-1: a "weak" 5' splice site followed by two poly-uridine tracts in intron 5 and another poly-U tract (putative binding site for TIA-1) in exon 6 (Figure 4.11). The Fas-activated serine/threonine kinase (FAST) is known to interact with TIA-1, phosphorylating it during Fas-mediated apoptosis [Tian et al., 1995]. This information led to a model for regulation of *Fas* splicing by TIA-1, illustrated and described in Figure 4.12.

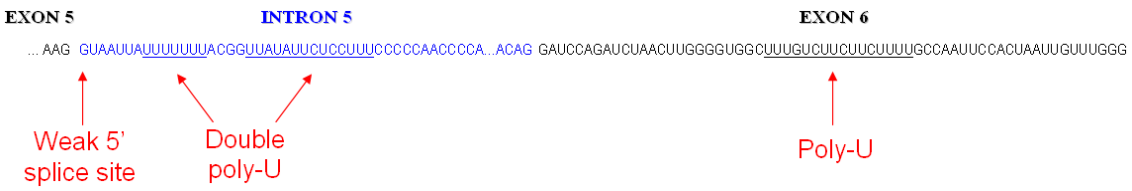

Figure 4.11: Special features of *Fas* intron 5 and exon 6 sequences

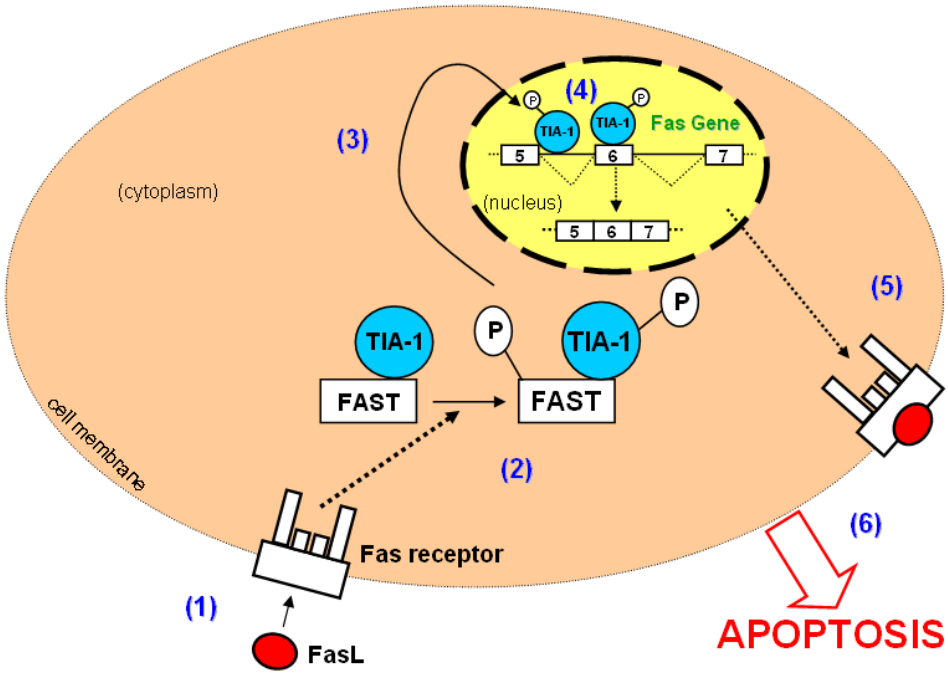

Figure 4.12: Model for regulation of *Fas* splicing in Jurkat cells

Binding of FasL (that acts as an apoptotic signal) to Fas receptor (transmembrane protein) **(1)** induces the phosphorylation of FAST and TIA-1 **(2)**, allowing TIA-1 to get into the nucleus **(3)** and promote the splicing of *Fas* exon 6 **(4)**. Through this positive feedback loop process there is a proliferation of Fas receptors in the cell membrane **(5)**, which triggers a subsequent signalling cascade that leads to apoptosis **(6)**.

I have questioned if TIA-1 could, in a similar way, regulate the expression of other genes involved in apoptotic pathways. To address the issue, I have developed a computational pipeline, based on Perl scripts, to select introns, from apoptosis-related genes, whose splicing was susceptible of being regulated by TIA-1 (as described and illustrated in Figure 4.13).

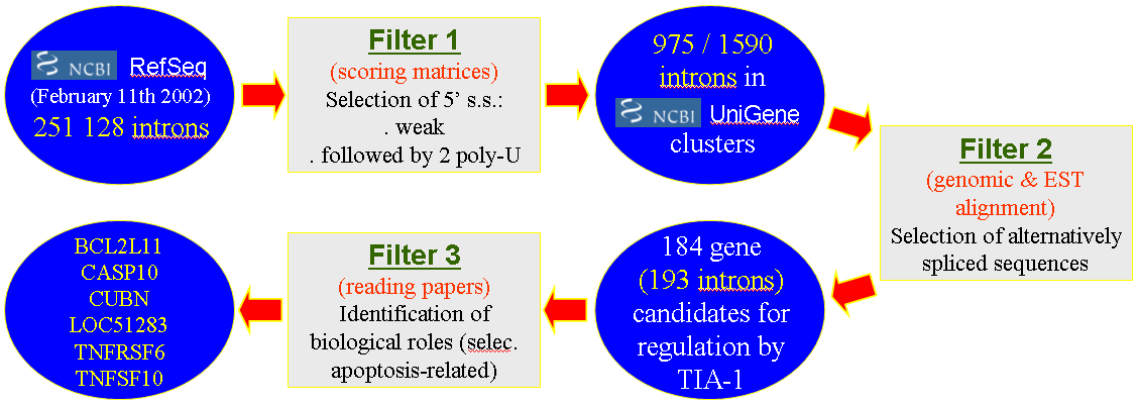

Figure 4.13: Schematics of the Perl-based computational pipeline flow

Sequences for all 251128 human introns annotated in RefSeq [Pruitt et al., 2005] were extracted and searched (using scoring matrices and regular expressions) for features indicating putative regulation by TIA-1: weak 5' splice site followed by two poly-uridine tracts. 1590 matched the criteria, 975 of which were annotated in UniGene [Wheeler et al., 2003] and could be checked for involvement in alternative splicing, by aligning genomic and EST sequences. 193 introns were selected and the biological role of the respective 184 genes was assessed. 6 of those genes were described as being involved in apoptosis.

The analysis provided 6 candidate genes: *BCL2L11* (Bcl-2-like protein 11 / Bcl2 interacting mediator of cell death), *CASP10* (Caspase-10 precursor / ICE-like apoptotic protease 4 / Apoptotic protease Mch-4 / FAS-associated death domain protein interleukin-1B-converting enzyme 2), *CUBN* (cubilin - intrinsic factor-cobalamin receptor), *LOC51283* (*BFAR* - bifunctional apoptosis regulator), *TNFSF10* (Tumor necrosis factor ligand superfamily member 10 / TNF-related apoptosis-inducing ligand / TRAIL protein / Apo-2 ligand/ Apo-2L) and, of course, *TNFRSF6* (*FAS* - TNF receptor superfamily, member 6).

In vitro experiments on the candidate genes revealed no evidence for splicing regulation by TIA-1. Literature on the genes corroborates the experimental results.

Further experiments validated a new model for the gulation of *Fas* alternative splicing, in which TIA-1 and PTB have antagonistic effects on the definition of exon 6 [Izquierdo et al., 2005] (illutrated in Figure 4.14).

Figure 4.14: Model for regulation of *Fas* splicing by TIA-1 and PTB
A - TIA-1 assists U1 snRNP in the recognition of the 5' splice site of exon 6 and downstream sequences, stabilizing the binding of U2AF to the upstream 3' splice site (arrow) and leading to exon definition. **B** - PTB binds to the poly-U tract in exon 6 to inhibit exon definition and U2AF binding to the upstream 3' splice site. **C** (variant of B) - PTB represses U1 snRNP's competence to establish exon definition interactions. (Adapted from [Izquierdo et al., 2005].)

4.3.2 AGAG introns and ALPS

Defects in genes regulating apoptosis are known to cause the autoimmune lymphoproliferative syndrome (ALPS). Its main clinical features are recurrent or more often chronic, benign, sometime massive lymphoadenopathy; splenomegaly of early onset; autoimmune phenomena such as thrombocytopenia or hemolytic anemia; less frequently, malignant lymphoma; frequent presence of CD4/CD8 double-negative, α/β receptor–positive T cells [Roesler et al., 2005].

The most common form of ALPS is associated with mutations in the *Fas* gene that lead to a defective FasL-induced apoptosis. One ALPS patient has been shown to harbor a mutation causing the skipping of *Fas* exon 6 (where the transmembrane domain is encoded), leading to an excessive production of the soluble form (sFas), which antagonizes FasL and inhibits the previously mentioned Fas-mediated apoptosis signaling cascade (Figure 4.15) [Roesler et al., 2005].

The C→G point mutation at position -3 of intron 5 leads to an alteration of the splicing *cis* regulatory region (cctacag/G→cctagag/G), generating a putative premature cryptic or ambiguous splice site (cctacag/G→cctag/AGG). No model for the

effect of the -3G mutation on spliceosome assembly has been tested. Nevertheless, as the mutation falls on the 3' splice site, it is likely to affect the binding of the U2 Auxiliary Factor or, at least, its capacity of recruiting the U2 snRNP. The introduction of a premature 3'ss would shorten its distance to the core of the poly-pyrimidine tract, in a way that could interfere with the binding of $U2AF^{65}$.

To address these questions, I have developed a computational pipeline, based on Perl scripts, in which sequences for all 251128 human introns annotated in RefSeq [Pruitt et al., 2005] were extracted and searched (using regular expressions) for "ag/AG", "agag/..." and normal "ag/..." 3' splice sites. For each intron, the uridine and pyrimidine contents of the 20 nucleotides upstream of the 3'ss were analyzed. Results are illustrated in Figure 4.16.

My analysis show that less than 2% of the 13995 human RefSeq "AGAG" introns have "agag/..." 3' splice sites. Confining the analysis to introns associated with curated genes, only 16 out of 4592 (0.35%) "AGAG" introns have are "agag/...". These results strongly suggest that the spliceosomal machinery performs the cleavage

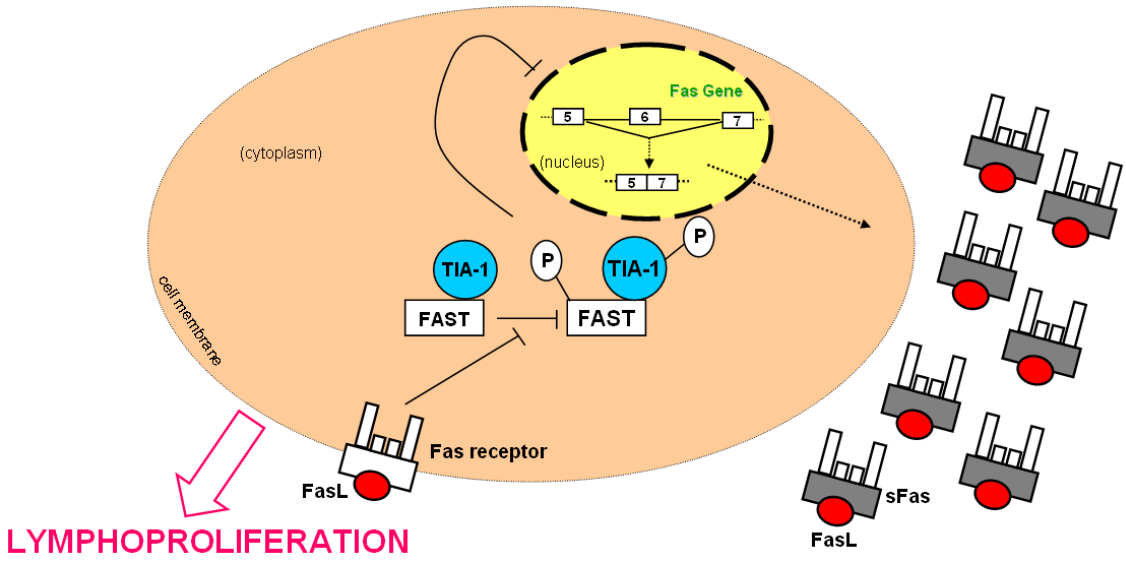

Figure 4.15: Model for abnormal sFas induced lymphoproliferation in ALPS patient
Splicing of *Fas* exon 6 (encoding the transmembrane domain) is inhibited and there is no synthesis of Fas receptors to bind FasL and trigger the apoptotic signaling cascade. Instead, there is the production of the soluble form of Fas (sFas), which antagonizes FasL.

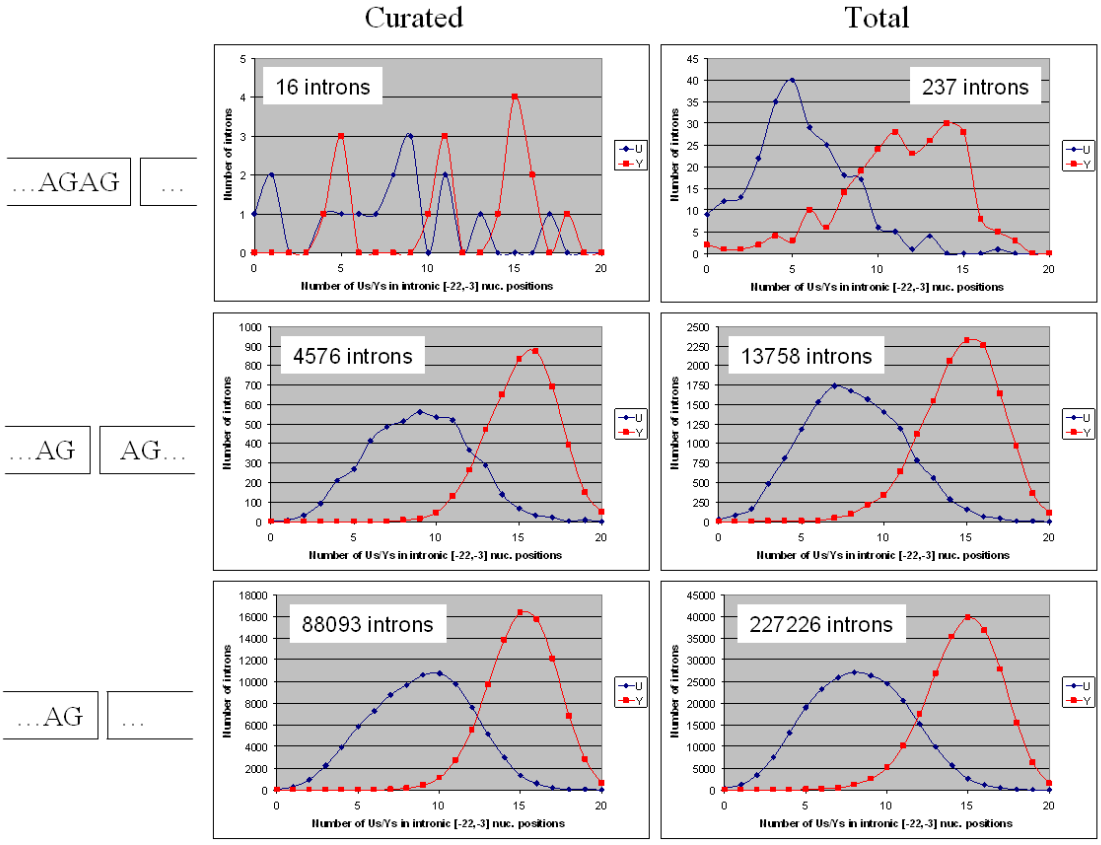

Figure 4.16: Distribution of pyrimidines near the 3'ss of "AGAG" introns (See text for details).

after the first "AG". This is not surprising, as a "YAG" 3' splice site is expected.

The distribution of adjacent uridines/pyrimidines for "ag/AG" 3'ss does not differ from the distribution associated with general "ag/..." 3'ss (average of 9-10 uridines and 15-16 pyrimidines amongst the adjacent 20 nucleotides). For the 16 curated "agag/..." introns the distribution is not conclusive. Taking all 237 "agag/..." introns in RefSeq, the analysis results show a more irregular distribution of pyrimidines and suggest weaker poly-Y tracts. I believe many of those introns are spurious (resulting from erroneous annotation) or maybe alternative, as they would exhibit weak 3'ss, not constitutively recognizable.

4.4 No evidence for a "hybrid" spliceosome

U12-type (so-called "minor") introns are known to coexist with U2-type introns in the same gene, showing no positional bias. Moreover, the protein composition and the pathways of assembly and catalysis of the major-class and minor-class spliceosomes are very similar, despite some mechanistic differences (namely U11 and U12 snRNPs entering the spliceosome as a two-snRNP complex) [Tarn and Steitz, 1997; Patel and Steitz, 2003]. These features legitimate the question whether there are "hybrid" spliceosomes, comprising simultaneously either U1 and U12 or U2 and U11 snRNPs. The possibility existence of such machineries has been raised by the finding that, for the human *MAPK8*[8] gene, the 3'ss of exon 6 is U11-type and the 5's of exon 7 is U2-type, suggesting a "hybrid" intron 6 (Figure 4.17).

To tackle this question, I have developed a computational pipeline, based on Perl scripts, in which sequences for all human introns annotated in RefSeq [Pruitt et al., 2005] were extracted and searched for U1-type and U11-type splice signals at the 5' end and U2-type and U12-type splice signals at the 3' end. I found no "hybrid" intron and I have also found that, for *MAPK8*, exons 6 and 7 are mutually exclusive. There appears to be no intron 6-7. Alternative introns 5-6 and 5-7 are processed by the constitutive machinery and introns 6-8 and 7-8 are spliced out by the minor

[8]Mitogen-activated protein kinase 8 (Stress-activated protein kinase JNK1) (c-Jun N-terminal kinase 1) (JNK-46)

spliceosome (Figure 4.17).

4.5 Intron clustering

U12-type introns, known to be spliced out by a specific spliceosome, were first recognized on the basis of unusual and conserved splicing signals [Burge et al., 1998]. Would it be possible to discriminate other classes of introns, with particular biological characteristics, based on their sequence features?

The scoring matrices used in sequence motif definition show consensus and overall positional abundance of each nucleotide but do not reveal specific associations between nucleotides. I have developed a simple statistical method, inspired in the concept of conditional probability. R is the ratio between the frequencies f of two specific nucleotides (A and B) in certain positions (i and j):

$$R(A_i \leftrightarrow B_j) = \frac{f(A_i|B_j)}{f(A_i)} = \frac{f(B_j|A_i)}{f(B_j)} \tag{4.7}$$

If the appearance of nucleotide A in position i is strongly associated with B in j,

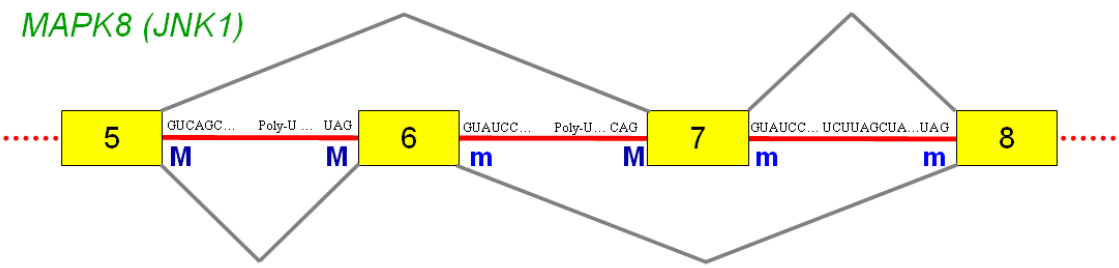

Figure 4.17: Intron-exon structure of human *MAPK8* gene

Virtual intron 6-7 (as exons 6 and 7 are mutually exclusive) exhibits a U11-type 5' splice site (|RUAUCC) and a U2-type 3' splice site (strong poly-uridine tract (poly-U) followed by YAG|). Intron 7-8 exhibits minor-type features |RUAUCC 5'ss and UCUUAGCU branch site (very close to the consensus UCCUUAACU). Intron 5-6 shows typical major-type splice signals. Intron 5-7 is therefore major-type and intron 6-8 is minor-type. (Exons are represented by yellow boxes, introns by red solid lines, splicing patterns by gray solid lines, major-type splice sites by dark blue **M**, minor-type splice sites by blue **m**.)

$R(A_i \leftrightarrow B_j) \gg 1$ is expected. If the two events are independent, then $R(A_i \leftrightarrow B_j) \approx 1$.

This approach was applied to the study of nucleotides +3 to +6 at the 5' end of all human introns annotated in RefSeq [Pruitt et al., 2005]. For example, it was possible to verify a strong dependence between the appearance of C in position +4 and C in position +5: $f(C_{+4}) = 0.0928$, $f(C_{+5}) = 0.0738$, $f(C_{+4}\ ,\ C_{+5}) = 0.0198$, $f(C_{+4} \mid C_{+5}) = 0.268$, $f(C_{+5} \mid C_{+4}) = 0.213 \Rightarrow \underline{f(C_{+4} \leftrightarrow C_{+5}) = 2.88}$. The biological meaning of this association remains unclear.

This work is merely a proof of principle. The method can acquire further sophistication and be applied to a much wider range of sequences.

Chapter 5

Microarrays and Sequence Annotation

One of the most important components of the information associated with a microarray experiment is the existence of detailed descriptions of all the genes involved. However the probe annotation provided to the experimentalist is very variable across platforms. Gene/probe identifiers (and data formats in general) are not uniform and annotations are usually poor in sequence information. This can become a particularly limiting factor in cross-laboratory experimental research and cross-platform data comparison.

Complete sequence information can be obtained from annotation databases, such as Ensembl [1] [Hubbard et al., 2002] and the UCSC Genome Browser Database [2] [Karolchik et al., 2003]. Scripting programming languages, namely Perl, can be used to automate the extraction of data from those repositories. Thus some bioinformatics skills and intimate knowledge of the data format are required to obtain detailed sequence information.

Microarray experiments can therefore greatly benefit from some experience in automated sequence analysis and annotation. In my case, the bioinformatics expertise acquired in the study of splicing regulatory sequences was an added value to sev-

[1] http://www.ensembl.org
[2] http://www.genome.ucsc.edu

eral microarray projects developed in the Oncology Department of the University of Cambridge, namely those involving genomic mapping of clone sequences, extraction of transcriptomic annotation and cross-platform meta-analysis of data. The most relevant of these collaborations are briefly described in this chapter.

5.1 RNA amplification and labelling

Reliability and reproducibility of expression microarray data crucial, as this tool is known to have an increasing clinical potential. Ali Naderi (Department of Oncology, University of Cambridge) leaded the development of a protocol that provides targets generating highly reproducible microarray data [Naderi et al., 2004]. Such protocol was obtained by evaluating the purification steps in indirect labelling of amplified RNA and experimentally determining the best method for each step.

Size distribution of transcripts was one of the tested features. Analysing the representation of transcripts across the size spectrum involved the determination of the transcript size range for each of the genes represented in the expression microarrays (6528 pairs of cDNA spots, CR-UK DMF Human 6.5k genome-wide array). The genes were non-redundantly annotated with a Perl script that also determined the longest and shortest annotated transcript associated with each gene, by combining the information available in RefSeq [3] [Pruitt et al., 2005] and Ensembl [Hubbard et al., 2002] (Human v.16.33.1) databases. This computational search was automated by using BioPerl [4] [Stajich et al., 2002] and Ensembl Perl modules on a Linux platform.

5.2 Large-scale Meta-analysis of Breast Cancer Microarray Data

Meta-analyses of cancer microarray data sets have revealed metasignatures associated with neoplastic transformation and histological grade. Analogous robust prognostic metasignatures have been more elusive, particularly in breast cancer. Andrew

[3] http://www.ncbi.nlm.nih.gov/RefSeq
[4] http://www.bioperl.org

Teschendorff (Department of Oncology, University of Cambridge) developed and applied unbiased semi-supervised models, based on Cox-regression and clustering, to homogeneous ER+ and ER- subgroups of patients within each of the three major current breast cancer microarray data sets in an attempt to derive robust prognostic metagene sets for outcome [Teschendorff et al., 2006b]. It is shown that prognostic feature selection using a Cox-regression model outperforms a standard t-test method based on dichotomising the outcome variable. Moreover, for the ER+ subgroup the derived prognostic metagene sets are strongly associated with outcome and validate the results using an independent external data set.

This identification of robust prognostic metagene sets for outcome in breast cancer required careful collation and preparation of the external microarray data sets. The microarray breast cancer data sets considered in this work were [van de Vijver et al., 2002; Sotiriou et al., 2003; Sorlie et al., 2003; Wang et al., 2005; Naderi et al., 2006]. I created an automated computational pipeline (Perl scripts on a Linux platform) to cross-link the annotation provided for each data set with UniGene [5] [Wheeler et al., 2003] (v. 176). For some data sets, the linkage relied on Ensembl [Hubbard et al., 2002] external database identifiers. Thus each probe was associated with an universal gene name. This procedure generated a non-redundant set of gene identifiers for the subsequent meta analysis.

The annotation of [van de Vijver et al., 2002; Wang et al., 2005; Sotiriou et al., 2003] datasets was used in the external validation of supervised clustering and genetic algorithm methods developed to identify prognostic gene-signatures for overall survival of patients with breast cancer [Naderi et al., 2006]. In this study, the transcript size range for each of the annotated genes was determined through a procedure like the one described in 5.1.

Likewise, the annotation of [Sotiriou et al., 2003; van 't Veer et al., 2002] datasets was used in the testing of a new variational Bayesian algorithm for cluster analysis of gene expression data, developed by Andrew Teschendorff, as described in [Teschendorff et al., 2005].

Finally, the annotation of [van de Vijver et al., 2002; Wang et al., 2005; Naderi

[5] http://www.ncbi.nlm.nih.gov/UniGene

et al., 2006] datasets was used in the development and validation of a feature selection method, based on a mixture model and a non-gaussianity measure of a genes expression profile, to find molecular classifiers in cancer. The procedure was given the name PACK (Profile Analysis using Clustering and Kurtosis) and was also developed by Andrew Teschendorff [Teschendorff et al., 2006a].

5.3 Molecular portraits of primary breast cancers using array-CGH

Breast cancer is the most common malignancy in women and several studies suggest the potential use of copy number profiling as an alternative to expression analysis to subtype breast cancers. Suet-Feung Chin (Department of Oncology, University of Cambridge) and colleagues used array-CGH to define molecular portraits of primary breast cancers, evaluating the copy number changes in 148 well-characterized breast cancers (the largest sample set studied to date using array-CGH) and examining the associations between genomic alterations and clinical phenotype of the tumors [Chin et al., 2006].

The results of this work were compared with array-CGH data sets published in [Nessling et al., 2005; Loo et al., 2004]. I have written a Perl script to cross-link the clone annotation for the arrays used in our study (Vysis Genosensor Array 300 [6]) with the annotation associated with the external data sets. The program relied on the clone annotation tables available at the UCSC Genome Browser Database [Karolchik et al., 2003].

Another Perl script was written to re-annotate all the Vysis clones. For each clone, the provided genomic coordinates were used to determine which curated genes are covered by the clone. For the purpose, we have relied on the gene annotation tables available at the UCSC Genome Browser Database [Karolchik et al., 2003]. UniGene [Wheeler et al., 2003] (v.176) annotation was used to ensure non-redundant gene sets.

[6] http://www.vysis.com/PDF/GenoSensor300ClonesAndKey_July2004.pdf

5.4 Profiling of CpG Islands

Epigenetic changes are heritable changes that include potentially reversible covalent modifications of histone proteins and methylation of DNA. The vast majority of mammalian DNA methylation is located at the cytosine of CpG dinucleotides which are particularly frequent within CpG islands. About 70% of mammalian genomic CpG dinucleotides are methylated and commonly occur within repetitive elements. In contrast, most unmethylated CpG islands span the promoter regions of house-keeping genes and tumour suppressor genes and are critical in gene expression regulation and cell differentiation. The number of cancer-related genes inactivated by epigenetic modifications may equal or exceed the number inactivated by genetic mutations or allele loss.

The identification of abnormal patterns of methylation requires a practical and reliable high-throughput method for identifying CpG methylation in independent samples. We have developed an improved array-based method called Microarray-based Methylation Assessment of Single Samples (MMASS) for identifying genome-wide CpG island methylation which directly compares methylated to unmethylated sequences within a single sample using digestion with methylation sensitive enzymes [Ibrahim et al., 2006].

The development and validation of the MMASS method involved the use of bioinformatic tools to provide detailed annotation of all probes on a publicly available CpG island array. Indeed we annotated a CpG island array with 13,056 features and compared an improved choice of methylation enzyme and enrichment by subtraction for methylated sequences against results from previously published protocols.

Perl scripts were used in the derivation and annotation of probe sequences. End sequences for the CpG island probes were obtained from the Sanger Centre [7]. They were BLASTed [Altschul et al., 1990] against the NCBI v.35 human genome assembly. Each probe sequence was then predicted from contiguous sequence tag alignments containing two "TTAA" *Mse*I recognition sites (as *Mse*I digestion was used to create the CpG island library) [8]. Bioperl [Stajich et al., 2002] and Ensembl [Hubbard

[7] `http://www.sanger.ac.uk/HGP/cgi.shtml`

[8] The majority of the library was subsequently fully sequenced by the Uni-

et al., 2002] (v.31) Perl modules were used in the genomic annotation of sequences. Repetitive sequences were identified using RepeatMasker [9].

Perl scripts were also used in the estimation of the number of restriction sites, per probe sequence, for *Mcr*BC, used to restrict the sample for representation of unmethylated sequences in the hybridization process. Likewise, to optimize the combination of enzymes for the representation of methylated sequences, the restriction sites for all commercially available methylation-sensitive enzymes were identified for unique probes together with the distance to the nearest neighboring genes and the percentage and type of included repetitive sequences.

After BLAST comparison to the human genome, 5435 out of 13056 (41.6%) probes had a percentage identity of >97% and <30% masked repeat elements and these were annotated as single copy sequences. A further 1190 probes (9.1%) contained 100% repeat sequences and the remainder were either not identifiable or had intermediate percentage of repetitive sequences.

We also found that 4160 out of 5435 (76.5%) of the probes on the CpG array would be informative when using the previously described combination of *Bst*UI, *Hpa*II and *Hha*I [Yan et al., 2002] enzymes to digest target DNA. We predicted that using a novel combination of four enzymes (*Aci*I, *Hpa*II, *Hin*P1I and *Hpy*CH4IV) would utilize 4403 out of 5435 (81%) of the array probes and therefore improve utility. In addition this optimized combination of enzymes was more convenient as all four enzymes could digest efficiently in the same buffer. In contrast the standard method digestion required *Bst*UI, *Hpa*II and *Hha*I in a two-step digestion protocol.

It is shown that MMASS offers improved sensitivity to profile methylated as well as unmethylated CpG islands from a single sample [Ibrahim et al., 2006].

versity Health Network Microarray Centre, Toronto (sequences available at `http://derlab.med.utoronto.ca/CpGIslands/`).

[9] `http://www.repeatmasker.org`

Chapter 6

Conclusions

This work aimed to shed some light on the mechanisms associated with complexity in eukaryotic gene expression, through computational approaches. I believe my research has given important and original scientific contributions, namely to the understanding the evolution of splicing and its relation to the complexity of organisms. Moreover my findings raise relevant questions that could trigger new lines of research.

By studying the complete machinery of splicing across eukaryotes, I have revealed differential gene family expansion. This striking phenomenon deserves further analysis, as it appears to have strong implications in eukaryotic gene expression and development. For instance, the remarkable apparent expansion (i.e. selective retention of duplicates) of the hnRNP content in the vertebrate lineage, which is disproportionate amongst splicing factors, may be explained by the diversity of functions of hnRNPs. Some hnRNPs are known to be transcription factors and therefore play a key role in gene expression regulation [Krecic and Swanson, 1999]. It is also possible that the larger number of cell types is correlated with expansion (or selection) for these duplicated proteins. Indeed, some hnRNPs are known to have tissue-specific (particularly neuron-specific) functions [Ashiya and Grabowski, 1997; Chan and Black, 1997; Chou et al., 1999; Wollerton et al., 2001; Zhang et al., 1999]. Previous reports already indicate that analogous expansion and selective retention of duplicates in other gene families is rare and appears to have a key role in development and speciation of vertebrates, as described for HOX [Amores et al., 1998; Amores et al., 2004] and sodium

channel [Lopreato et al., 2001] gene clusters.

Moreover, hundreds of non-coding sequences recently shown to be highly conserved in vertebrates are not found in invertebrates. These conserved sequences are associated to transcription factor and developmental genes and believed to be part of gene regulatory networks in vertebrates. Functional studies have demonstrated, for most of them, tissue-specific regulatory action [Woolfe et al., 2005]. Interestingly subsets of the conserved non-coding elements share sequence similarity and are associated with genes from transcription factors from the same families. Based on this paralogous relationship between similar elements and assuming that sequence similarity corresponds to functional similarity, it is suggested that those duplicated elements might act as *cis*-regulators directing tissue-specific expression and it is reasonable to expect them to be shared between paralogous genes with related expression patterns. Computational comparative studies show retention of regulatory elements between some gene duplicates over evolution and a particularly strong association between the those elements and duplicated transcription factors [Vavouri et al., 2006]. A model for the evolution of conserved non-coding elements, in the context of other major genomic events during the early vertebrate radiation, has been proposed recently (Figure 6.1). Whole-genome duplications and the resulting expansion of the gene repertoire, together with the appearance of a new set of rapidly evolving *cis*-regulatory elements, coincides with fundamental and persistent changes in morphological complexity in vertebrate stem. Given the association between conserved non-coding elements and developmental genes, it is very likely that these events are directly connected [McEwen et al., 2006]. I believe hnRNPs may have been involved in this sequence modelling process, as they are involved in gene expression regulation and their expansion seems to coincide with the appearance of the described regulatory elements. The functional features of these elements have not been fully determined yet and it would be interesting to assess if some of them can be targets for hnRNPs or, at least, be involved in the same expression pathways. It should be noticed that hnRNPs are splicing factors known to bind to introns and the action of many intronic conserved sequence elements is still to be evaluated. Moreover some splicing factors are known to self-regulate their alternative splicing, with important biological consequences [Wollerton

et al., 2004]. Models to explain the evolution of alternative splicing must consider the selective pressure of splicing regulatory factors not only on splice sites but also on other *cis*-elements.

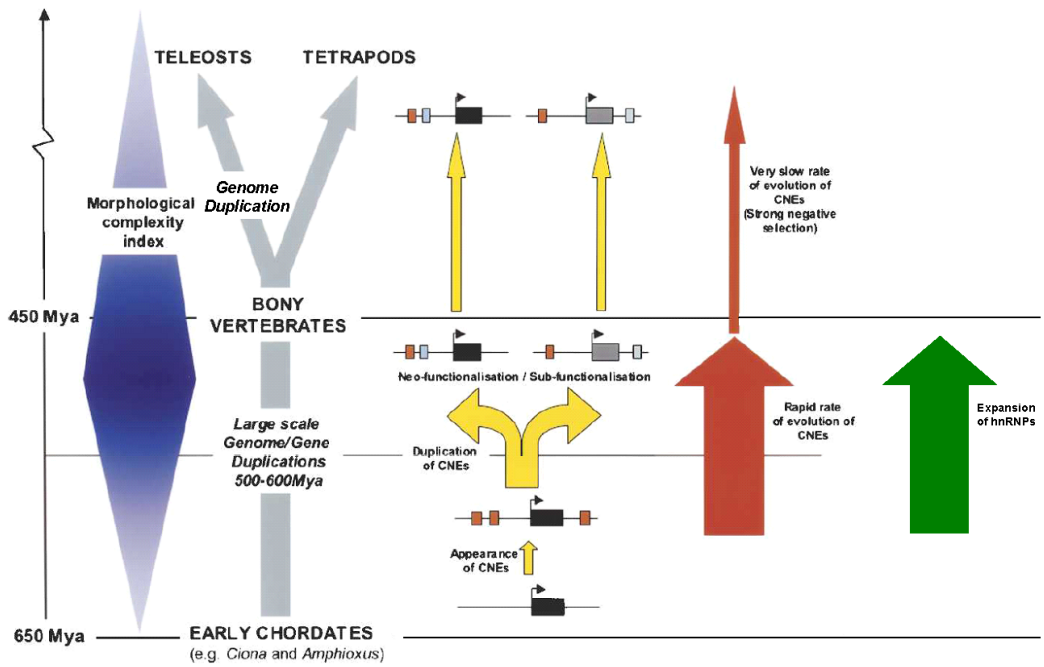

Figure 6.1: Model of the evolution of conserved non-coding elements (CNEs)
Extant vertebrates evolved from the chordate lineage undergoing a period of rapid morphological change (in blue, based on [Aburomia et al., 2003]). During this period (between 650 and 450 Mya) the early ancestral vertebrate underwent one or two whole-genome duplications, which may have contributed to this increase in morphological complexity, by expanding the repertoire of genes. hnRNP families must have been expanded with these events (green arrow). CNEs (red boxes adjacent to gene loci, depicted as dark boxes) are likely to have appeared in vertebrate genomes prior to those large-scale duplications, as most of the duplicated CNEs are associated with paralogous genes involved in transcriptional regulation and/or development, deriving from these ancient duplications (yellow arrows). The duplication of gene loci together with associated *cis*-regulatory modules provides the plasticity for genes to undergo neofunctionalization and/or subfunctionalization. This evolution is believed to have occurred rapidly following duplication over a relatively short evolutionary period ($\sim$50$-$150 Myr) during which duplicated CNEs evolved in length and sequence. In the period since the teleosttetrapod divergence ($\sim$450 Mya), they have had a remarkably slow mutation rate and have remained practically unchanged. (Adapted from [McEwen et al., 2006].)

Some of the questions and hypothesis raised by my work could be addressed by correlating functional specificity of individual splicing factors for their isoforms with the recognition motifs in different species or tissues. In that context, microarrays have been used, for example, to evaluate simultaneously the levels of expression of splicing factors and the patterns of alternative splicing of genes involved in tumor progression [Relogio et al., 2005]. Recently, a systems approach, using splicing oligonucleotide microarrays to find broad relationships between regulation of alternative splicing and sequence conservation, revealed unusual intronic sequence conservation near tissue-regulated exons and identified new sequence motifs implicated in brain and muscle splicing regulation [Sugnet et al., 2006].

Despite the relative success of the described assays in linking the actions of *trans* and *cis* splicing regulators, a great effort in improving the definition of binding motifs for splicing factors is still required. Tools like the Splicing Rainbow (section 4.1) [Stamm et al., 2006] have already been used to try to correlate the expression of splicing factors with alternative splicing profiles [Relógio, 2002] but their accuracy is limited, as they tend to generate many false positives and provide little information about the environmental or tissue context. New approaches are needed and here we suggest that specific spatial associations between nucleotides may provide some extra insight in resolving sequence motifs (section 4.5). Nevertheless, our work clearly shows the potential of sequence motif search in shedding some light over gene expression regulatory mechanisms, as we were able to reveal the influence of untranslated sequence regulatory elements on the differential interaction between functionally related mRNA populations and specific regulatory RNA binding proteins (section 4.2) [Gama-Carvalho et al., 2006].

The evolution of the overall abundance of alternative isoforms and its relation with the spliceosome's evolution is still an open question. Currently, we can not assess if the expansion of splicing regulators contributes to more alternative splices in vertebrates and therefore to potential further complexity. The notion of increased alternative splicing in vertebrates is still somewhat contentious. The relatively low number of human genes [Lander et al., 2001; Venter et al., 2001], when compared with simpler species, led, among many other hypothesis (greater gene modularity in human, post-

translational modifications [Banks et al., 2000]), to the idea that alternative splicing may be responsible for more transcripts per gene and therefore a much larger proteome in human than in other species [Ewing and Green, 2000]. However, different large scale EST studies lead to different results. A recent estimate indicates greater amount of alternative splicing in mammals than in vertebrates [Kim et al., 2004] but those results were immediately disputed by the authors of a previous analysis which suggests that the total amount of alternative splicing is comparable among animals (mammals, insects and worms) [Brett et al., 2002]. Furthermore, a recent study suggests levels of alternative splicing in Drosophila similar to those in Human [Stolc et al., 2004]. I am therefore led to believe that the influence of alternative splicing on complexity is not purely quantitative and a few additional key isoforms can significantly broaden the spectrum of protein activities in some physiologically important tissues. Specific alternative splicing patterns in certain genes or subtle sophistication on the splicing regulatory pathways, in which some hnRNPs are involved, may contribute to an organism's complexity. Developed brains and nervous systems are the distinguishing physiological features of higher organisms and it has been suggested that alternative splicing is indeed extensive in neurons and optimizes the activity of key neuronal proteins [Lipscombe, 2005], consistently with reports of neuron-specific functions of hnRNPs [Ashiya and Grabowski, 1997; Chan and Black, 1997; Chou et al., 1999; Zhang et al., 1999].

My analysis of the spliceosomal evolution reveals additional lineage-specific features, within vertebrates, that deserve further research. The functional consequences of the teleost-specific duplication of some splicing factors have not been assessed, in part due to the lack of curated transcriptomic data for those species. Were the duplicates retained by neofunctionalization or subfunctionalization? This question is of particular interest, given, first, the conservation of most vertebrate-specific CNEs in teleost duplicates [McEwen et al., 2006] and, second, the apparent resemblance between teleost gene duplication and mammalian alternative splicing (suggesting subfunctionalization) we report for U2AF^{35} [Pacheco et al., 2004], also described for other genes [Altschmied et al., 2002; Yu et al., 2003].

It is also outstanding that retrotransposition introduced an additional level of

diversity to the mammalian splicing machinery, given that he majority of retrotransposons are non-functional [Goncalves et al., 2000] and lineage specific, created after human and rodents diverged [Zhang et al., 2004]. Moreover, intronless genes do not undergo alternative splicing and do not benefit from the consequent variability in their expression. It is also believed that intronless genes tend to be transcribed less efficiently than their intron-containing homologs [Le Hir et al., 2003]. They are not supposed to benefit from the same set of evolutionary selected regulatory elements, as most of these are not retrotransposed with the trancript. It is therefore remarkable that those monoexonic factors were positively selected and increase the diversity of families that are already very diverse, comprising factors with subtle and specific regulatory functions and whose expression is also subtly regulated (sometimes by alternative splicing). They actually play relevant roles in key cellular activities: SRp46 is shown to be a trans-acting splicing factor, exhibiting the general features of SR proteins [Soret et al., 1998]; hnRNP E1 is involved in cell spreading [de Hoog et al., 2004], telomere functioning [Bandiera et al., 2003], translational regulation [Antony et al., 2004; Krecic and Swanson, 1999; Leffers et al., 1995; Persson et al., 2003; Reimann et al., 2002] and mRNA stability [Ostareck-Lederer et al., 1998; Morris et al., 2004]; hnRNP G-T appears to be important for germ cell development [Elliott et al., 2000]; smPTB, expressed in some types of smooth muscle in rodents, shares some of the features of PTB and is able to act as a regulator in some alternative splicing events [Gooding et al., 2003]; the imprinted U2AF1-RS1 is involved in tissue-specific transcription regulation [Wang et al., 2004a]. The number of putative retrotransposons is particularly high for some families of Sm proteins and for the hnRNP A family (Tables A.5 and A.7). Previous studies actually show that, for human and mouse, genes which have multiple copies of processed pseudogenes are mostly housekeeping genes with high expression in germ and embryonic cells [Zhang et al., 2003; Zhang et al., 2004]. In the same studies, ribosomal proteins, DNA and RNA binding proteins, structural molecules and metabolic enzymes emerge as the most represented groups of the classification of pseudogenes based on Gene Ontology [Ashburner et al., 2000] functional categories of the functional genes. All intronless mammalian-specific splicing factors (except hnRNP G-T), the hnRNP A family members and most of the

Sm proteins with multiple homologous pseudogenes are "RNA binding".

This discussion actually suggests that, by trying to tackle some important problems in the evolution of gene expression, my work has opened many fundamental questions. Nevertheless, we are now closer to understanding the coordinated action of splicing *cis* and *trans* elements and we can draw a draft overview of the evolution of splicing, as illustrated in Figure 6.2. In summary, although self-splicing occurs in bacteria [Ferat and Michel, 1993] and we found a couple of putative Sm proteins in archaea, the emergence of spliceosomal splicing and the corresponding machinery appear to fall in the roots of eukaryotes. snRNP protein genes are conserved across the eukaryotic lineage but multicellular organisms benefit from more genes implicated in the regulation of splicing than unicellular ones. The evolution of splicing regulatory factors releases the selective pressure from splice sites and indeed alternative splicing appears to have arisen with multicellularity. Whole-genome duplications at the vertebrate stem allowed for the expansion of hnRNPs, the emergence of more regulatory elements and extra specificity and subtlety in the mechanisms of gene expression (and splicing, in particular) regulation, causing important changes in development and morphological complexity of organisms. Further lineage-specific events of gene duplication (whole-genome duplication in teleosts, retrotransposition in mammals, polyploidization in plants) introduced additional diversity to the splicing machinery and contributed to speciation.

Finally, this work shows that complete genome-wide studies on evolution, function and expression can be integrated in one consistent bioinformatics framework. The same set of computational tools for sequence analysis and annotation were used in a complete phylogenetic study, in motif searches and in the cross-annotation of microarray data. Addressing fundamental questions at the level of gene expression involves all those types of information so research in this field clearly demands pipelines capable of integrating and making sense of data provided by such diverse sources. Biology became a multidisciplinary science and requires tight coordination and symbiosis between *in silico* and "wet lab" approaches.

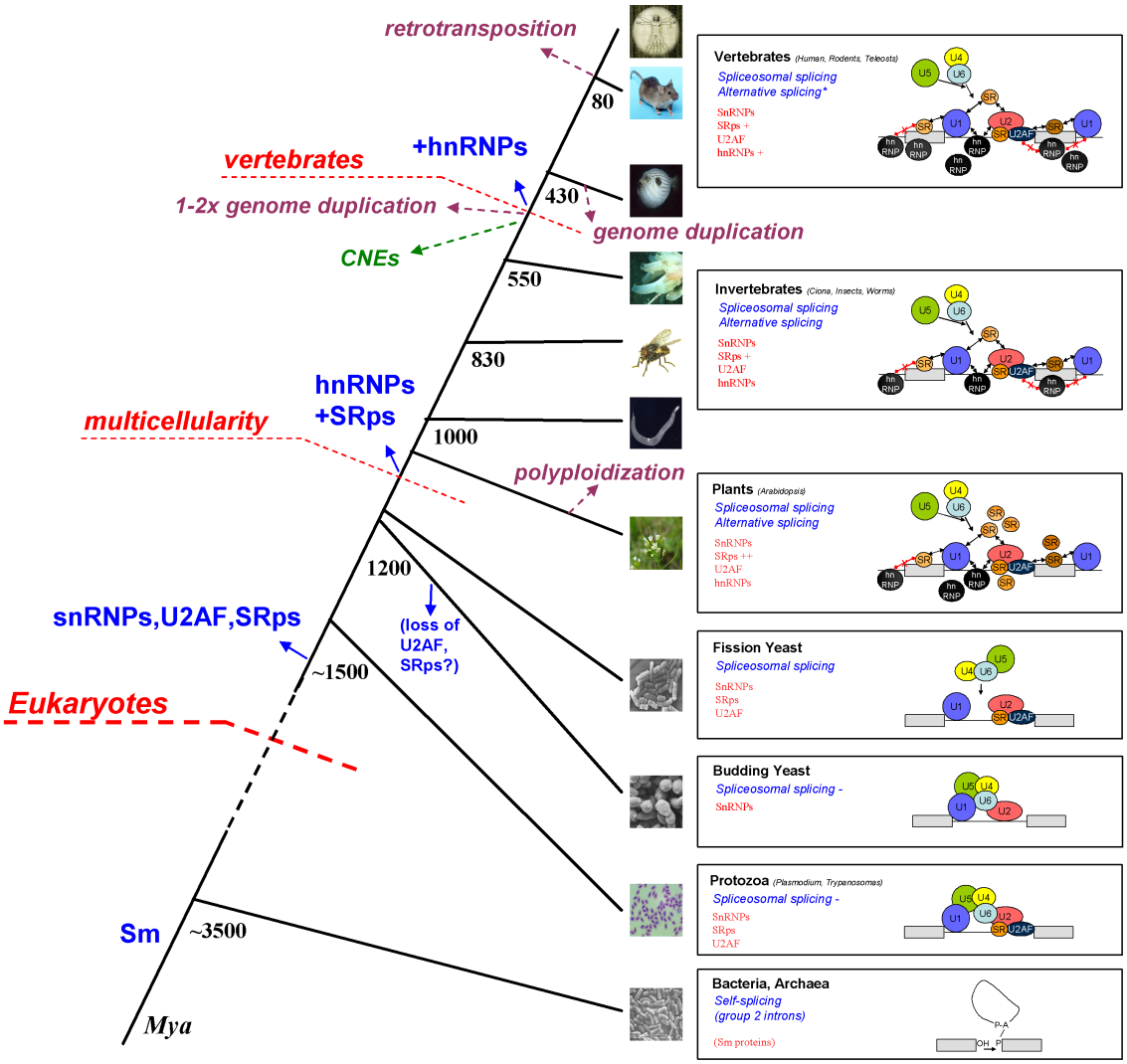

Figure 6.2: The evolution of splicing

The symbolic evolutionary tree of splicing is depicted in black lines, with black text near the nodes representing corresponding approximate dates of divergence (in million years ago). Adjacent blue text and arrows represent appearance and expansion (+) of splicing factors families, purple text and dashed arrows represent major events of gene duplication, red text and dashed lines sign important steps in the evolution of organic complexity, green text and dashed arrow signs the emergence of conserved non-coding elements. Boxes on the right are simplified schematics of the splicing machinery for the corresponding species. Exons are represented by boxes and introns by lines connecting them. For the type of splicing (text in blue italic), (−) represents low frequency and (∗) putative extra subtlety in patterns of alternative splicing. For the families of factors (text in red), (+) and (++) represent extra abundance of factors of a particular type. See text for details.

Bibliography

Abovich, N., Liao, X. C., and Rosbash, M. (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev*, 8(7):843–54.

Aburomia, R., Khaner, O., and Sidow, A. (2003). Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. *J Struct Funct Genomics*, 3(1-4):45–52.

Albertson, D. G. and Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*, 12 Spec No 2:R145–52.

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J. N., and Schartl, M. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics*, 161(1):259–67.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

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(17):3389–402.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., and Postlethwait, J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282(5394):1711–4.

Amores, A., Suzuki, T., Yan, Y. L., Pomeroy, J., Singer, A., Amemiya, C., and Postlethwait, J. H. (2004). Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res*, 14(1):1–10.

Anderson, P. and Kedersha, N. (2002). Stressful initiations. *J Cell Sci*, 115(Pt 16):3227–34.

Antony, A., Tang, Y. S., Khan, R. A., Biju, M. P., Xiao, X., Li, Q. J., Sun, X. L., Jayaram, H. N., and Stabler, S. P. (2004). Translational upregulation of folate receptors is mediated by homocysteine via RNA-heterogeneous nuclear ribonucleoprotein E1 interactions. *J Clin Invest*, 113(2):285–301.

Aparicio, S. (2000). Vertebrate evolution: recent perspectives from fish. *Trends Genet*, 16(2):54–6.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297(5585):1301–10.

Ars, E., Serra, E., Garcia, J., Kruyer, H., Gaona, A., Lazaro, C., and Estivill, X. (2000). Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum Mol Genet*, 9(2):237–47.

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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9.

Ashiya, M. and Grabowski, P. J. (1997). A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. *Rna*, 3(9):996–1015.

Ast, G. (2004). How did alternative splicing evolve? *Nat Rev Genet*, 5(10):773–82.

Auboeuf, D., Dowhan, D. H., Kang, Y. K., Larkin, K., Lee, J. W., Berget, S. M., and O'Malley, B. W. (2004). Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. *Proc Natl Acad Sci U S A*, 101(8):2270–4.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33(Database issue):D154–9.

Bandiera, A., Tell, G., Marsich, E., Scaloni, A., Pocsfalvi, G., Akintunde Akindahunsi, A., Cesaratto, L., and Manzini, G. (2003). Cytosine-block telomeric type DNA-binding activity of hnRNP proteins from human cell lines. *Arch Biochem Biophys*, 409(2):305–14.

Banerjee, H., Rahn, A., Davis, W., and Singh, R. (2003). Sex lethal and U2 small nuclear ribonucleoprotein auxiliary factor (U2AF65) recognize polypyrimidine tracts using multiple modes of binding. *Rna*, 9(1):88–99.

Banks, R. E., Dunn, M. J., Hochstrasser, D. F., Sanchez, J. C., Blackstock, W., Pappin, D. J., and Selby, P. J. (2000). Proteomics: new perspectives, new biomedical opportunities. *Lancet*, 356(9243):1749–56.

Barbosa-Morais, N. L., Carmo-Fonseca, M., and Aparicio, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res*, 16(1):66–77.

Barrass, J. D. and Beggs, J. D. (2003). Splicing goes global. *Trends Genet*, 19(6):295–8.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res*, 30(1):276–80.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. *Nucleic Acids Res*, 32 Database issue:D23–6.

Berriman, M. and Rutherford, K. (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform*, 4(2):124–32.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 72:291–336.

Blencowe, B. J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci*, 25(3):106–10.

Blencowe, B. J., Issner, R., Nickerson, J. A., and Sharp, P. A. (1998). A coactivator of pre-mRNA splicing. *Genes Dev*, 12(7):996–1009.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–70.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST–database for "expressed sequence tags". *Nat Genet*, 4(4):332–3.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8.

Brent, M. R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res*, 15(12):1777–86.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30.

Brooks, S. A. and Rigby, W. F. (2000). Characterization of the mRNA ligands bound by the RNA binding protein hnRNP A2 utilizing a novel in vivo technique. *Nucleic Acids Res*, 28(10):E49.

Burd, C. G. and Dreyfuss, G. (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *Embo J*, 13(5):1197–204.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94.

Burge, C., Tuschl, T., and Sharp, P. (1999). Splicing of Precursors to mRNAs by the Spliceosomes. In Gesteland, R., Cech, T., and Atkins, J., editors, *The RNA World, Second Edition*, pages 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2nd edition.

Burge, C. B., Padgett, R. A., and Sharp, P. A. (1998). Evolutionary fates and origins of U12-type introns. *Mol Cell*, 2(6):773–85.

Caceres, J. F. and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*, 18(4):186–93.

Caputi, M. and Zahler, A. M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *J Biol Chem*, 276(47):43850–9.

Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4):285–98.

Cartegni, L. and Krainer, A. R. (2002). Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet*, 30(4):377–84.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., and Krainer, A. R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, 31(13):3568–71.

Cavaloc, Y., Bourgeois, C. F., Kister, L., and Stevenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *Rna*, 5(3):468–83.

Chan, R. C. and Black, D. L. (1995). Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol Cell Biol*, 15(11):6377–85.

Chan, R. C. and Black, D. L. (1997). The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol Cell Biol*, 17(8):4667–76.

Chen, C. D., Kobayashi, R., and Helfman, D. M. (1999). Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev*, 13(5):593–606.

Cheng, J., Zhou, T., Liu, C., Shapiro, J. P., Brauer, M. J., Kiefer, M. C., Barr, P. J., and Mountz, J. D. (1994). Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science*, 263(5154):1759–62.

Chin, S.-F., Wang, Y., Thorne, N. P., Teschendorff, A. E., Pinder, S. E., Vias, M., Barbosa-Morais, N. L., Roberts, I., Naderi, A., Garcia, M., Iyer, N. G., Kranjac, T., Robertson, J., Ruffalo, T., Aparicio, S., Tavare, S., Ellis, I., Brenton, J., and Caldas, C. (2006). Using array-CGH to define molecular potraits of primary breast cancer. *Oncogene*, (in press).

Chou, M. Y., Rooke, N., Turck, C. W., and Black, D. L. (1999). hnRNP H is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. *Mol Cell Biol*, 19(1):69–77.

Christoffels, A., Koh, E. G., Chia, J. M., Brenner, S., Aparicio, S., and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol*, 21(6):1146–51.

Collins, L. and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol*, 22(4):1053–66.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, 16(22):10881–90.

Coulter, L. R., Landree, M. A., and Cooper, T. A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol*, 17(4):2143–50.

de Hoog, C. L., Foster, L. J., and Mann, M. (2004). RNA and RNA binding proteins participate in early stages of cell spreading through spreading initiation centers. *Cell*, 117(5):649–62.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., Harafuji, N., Hastings, K. E., Ho, I., Hotta, K., Huang, W., Kawashima, T., Lemaire, P., Martinez, D., Meinertzhagen, I. A., Necula, S., Nonaka, M., Putnam, N., Rash, S., Saiga, H., Satake, M., Terry, A., Yamada, L., Wang, H. G., Awazu, S., Azumi, K., Boore, J., Branno, M., Chin-Bow, S., DeSantis, R., Doyle, S., Francino, P., Keys, D. N., Haga, S., Hayashi, H., Hino, K., Imai, K. S., Inaba, K., Kano, S., Kobayashi, K., Kobayashi, M., Lee, B. I., Makabe, K. W., Manohar, C., Matassi, G., Medina, M., Mochizuki, Y., Mount, S., Morishita, T., Miura, S., Nakayama, A., Nishizaka, S., Nomoto, H., Ohta, F., Oishi, K., Rigoutsos, I., Sano, M., Sasaki, A., Sasakura, Y., Shoguchi, E., Shin-i, T., Spagnuolo, A., Stainier, D., Suzuki, M. M., Tassy, O., Takatori, N., Tokuoka, M., Yagi, K., Yoshizaki, F., Wada, S., Zhang, C., Hyatt, P. D., Larimer, F., Detter, C., Doggett, N., Glavina, T., Hawkins, T., Richardson, P., Lucas, S., Kohara, Y., Levine, M., Satoh, N., and Rokhsar, D. S. (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–67.

DeMaria, C. T. and Brewer, G. (1996). AUF1 binding affinity to A+U-rich elements correlates with rapid mRNA degradation. *J Biol Chem*, 271(21):12179–84.

Dember, L. M., Kim, N. D., Liu, K. Q., and Anderson, P. (1996). Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities. *J Biol Chem*, 271(5):2783–8.

Deonier, R. C., Tavaré, S., and Waterman, M. S. (2005). *Computational genome analysis : an introduction*. Springer, New York.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M., and Bourne, P. E. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*, 33(Database issue):D233–7.

Domon, C., Lorkovic, Z. J., Valcarcel, J., and Filipowicz, W. (1998). Multiple forms of the U2 small nuclear ribonucleoprotein auxiliary factor U2AF subunits expressed in higher plants. *J Biol Chem*, 273(51):34603–10.

Dowhan, D. H., Hong, E. P., Auboeuf, D., Dennis, A. P., Wilson, M. M., Berget, S. M., and O'Malley, B. W. (2005). Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERalpha and CAPERbeta. *Mol Cell*, 17(3):429–39.

Duncan, P. I., Stojdl, D. F., Marius, R. M., Scheit, K. H., and Bell, J. C. (1998). The Clk2 and Clk3 dual-specificity protein kinases regulate the intranuclear distribution of SR proteins and influence pre-mRNA splicing. *Exp Cell Res*, 241(2):300–8.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–63.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Monographs on statistics and applied probability ; 57. Chapman & Hall, New York. Bradley Efron and Robert J. Tibshirani. ill. ; 23 cm.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet*, 17(11):661–9.

Eldridge, A. G., Li, Y., Sharp, P. A., and Blencowe, B. J. (1999). The SRm160/300 splicing coactivator is required for exon-enhancer function. *Proc Natl Acad Sci U S A*, 96(11):6125–30.

Elliott, D. J., Venables, J. P., Newton, C. S., Lawson, D., Boyle, S., Eperon, I. C., and Cooke, H. J. (2000). An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. *Hum Mol Genet*, 9(14):2117–24.

Epstein, J. A., Glaser, T., Cai, J., Jepeal, L., Walton, D. S., and Maas, R. L. (1994). Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing. *Genes Dev*, 8(17):2022–34.

Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet*, 25(2):232–4.

Fairbrother, W. G., Yeh, R. F., Sharp, P. A., and Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13.

Fairbrother, W. G., Yeo, G. W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P. A., and Burge, C. B. (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res*, 32(Web Server issue):W187–90.

Fast, N. M. and Doolittle, W. F. (1999). Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8. *Mol Biochem Parasitol*, 99(2):275–8.

Faustino, N. A. and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes Dev*, 17(4):419–37.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–76.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791.

Felsenstein, J. (1989). PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166.

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–60.

Ferat, J. L. and Michel, F. (1993). Group II self-splicing introns in bacteria. *Nature*, 364(6435):358–61.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45.

Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P., and Valcarcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. *Mol Cell*, 6(5):1089–98.

Forch, P., Puig, O., Martinez, C., Seraphin, B., and Valcarcel, J. (2002). The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. *Embo J*, 21(24):6882–92.

Francino, M. P. (2005). An adaptive radiation model for the origin of new gene functions. *Nat Genet*, 37(6):573–7.

Gama-Carvalho, M., Barbosa-Morais, N. L., Brodsky, A. R., Silver, P., and Carmo-Fonseca, M. (2006). Genome wide identification of functionally distinct subsets of cellular mRNAs associated with the mammalian splicing factors U2AF65 and PTB. *(Submitted)*.

Gama-Carvalho, M. H. (2002). *Nuclear Compartmentalisation of Splicing Factors: Characterisation of Molecular Signals and Role in Alternative Splicing Regulation*. Phd, Universidade de Lisboa.

Garcia-Fernandez, J. and Holland, P. W. (1996). Amphioxus Hox genes: insights into evolution and development. *Int J Dev Biol*, Suppl 1:71S–72S.

Gemund, C., Ramu, C., Altenberg-Greulich, B., and Gibson, T. J. (2001). Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucleic Acids Res*, 29(6):1272–7.

Golling, G., Amsterdam, A., Sun, Z., Antonelli, M., Maldonado, E., Chen, W., Burgess, S., Haldi, M., Artzt, K., Farrington, S., Lin, S. Y., Nissen, R. M., and Hopkins, N. (2002). Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nat Genet*, 31(2):135–40.

Goncalves, I., Duret, L., and Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. *Genome Res*, 10(5):672–8.

Gooding, C., Kemp, P., and Smith, C. W. (2003). A novel polypyrimidine tract-binding protein paralog expressed in smooth muscle cells. *J Biol Chem*, 278(17):15201–7.

Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, 15(4):305–8.

Gozani, O., Feld, R., and Reed, R. (1996). Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev*, 10(2):233–43.

Gozani, O., Potashkin, J., and Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol Cell Biol*, 18(8):4752–60.

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. *Rna*, 6(9):1197–211.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, 17(2):100–7.

Graveley, B. R. (2002). Sex, AGility, and the regulation of alternative splicing. *Cell*, 109(4):409–12.

Green, M. R. (1986). Pre-mRNA splicing. *Annu Rev Genet*, 20:671–708.

Griffin, T. J. and Aebersold, R. (2001). Advances in proteome analysis by mass spectrometry. *J Biol Chem*, 276(49):45497–500.

Gu, X., Wang, Y., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet*, 31(2):205–9.

Gu, X. and Zhang, J. (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol*, 14(11):1106–13.

Guth, S., Martinez, C., Gaur, R. K., and Valcarcel, J. (1999). Evidence for substrate-specific requirement of the splicing factor U2AF(35) and for its function after polypyrimidine tract recognition by U2AF(65). *Mol Cell Biol*, 19(12):8263–71.

Guth, S. and Valcarcel, J. (2000). Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem*, 275(48):38059–66.

Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J*, 9(8):576–96.

Hartmuth, K., Urlaub, H., Vornlocher, H. P., Will, C. L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc Natl Acad Sci U S A*, 99(26):16719–24.

Hastings, M. L. and Krainer, A. R. (2001). Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol*, 13(3):302–9.

Hatada, I., Kitagawa, K., Yamaoka, T., Wang, X., Arai, Y., Hashido, K., Ohishi, S., Masuda, J., Ogata, J., and Mukai, T. (1995). Allele-specific methylation and expression of an imprinted U2af1-rs1 (SP2) gene. *Nucleic Acids Res*, 23(1):36–41.

Hatada, I., Sugama, T., and Mukai, T. (1993). A new imprinted gene cloned by a methylation-sensitive genome scanning method. *Nucleic Acids Res*, 21(24):5577–82.

Hayashizaki, Y., Shibata, H., Hirotsune, S., Sugino, H., Okazaki, Y., Sasaki, N., Hirose, K., Imoto, H., Okuizumi, H., Muramatsu, M., and et al. (1994). Identification of an imprinted U2af binding protein related sequence on mouse chromosome 11 using the RLGS method. *Nat Genet*, 6(1):33–40.

Heinrichs, V. and Baker, B. S. (1995). The Drosophila SR protein RBP1 contributes to the regulation of doublesex alternative splicing by recognizing RBP1 RNA target sequences. *Embo J*, 14(16):3987–4000.

Hertel, K. J. and Maniatis, T. (1998). The function of multisite splicing enhancers. *Mol Cell*, 1(3):449–55.

Holland, P. W. (1997). Vertebrate evolution: something fishy about Hox genes. *Curr Biol*, 7(9):R570–2.

Holland, P. W., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Dev Suppl*, pages 125–33.

Huang, X. and Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics*, 12(3):337–357.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41.

Iborra, F. J., Jackson, D. A., and Cook, P. R. (2001). Coupled transcription and translation within nuclei of mammalian cells. *Science*, 293(5532):1139–42.

Ibrahim, A. E. K., Thorne, N. P., Baird, K., Barbosa-Morais, N. L., Tavare, S., Collins, V. P., Wyllie, A. H., Arends, M. J., and Brenton, J. D. (2006). MMASS: an optimised array-based method for assessing CpG island methylation. *Nucleic Acids Res*, (in press).

Ishikawa, F., Matunis, M. J., Dreyfuss, G., and Cech, T. R. (1993). Nuclear proteins that bind the pre-mRNA 3' splice site sequence r(UUAG/G) and the human telomeric DNA sequence d(TTAGGG)n. *Mol Cell Biol*, 13(7):4301–10.

Ismaili, N., Sha, M., Gustafson, E. H., and Konarska, M. M. (2001). The 100-kda U5 snRNP protein (hPrp28p) contacts the 5' splice site through its ATPase site. *Rna*, 7(2):182–93.

Izquierdo, J. M., Majos, N., Bonnal, S., Martinez, C., Castelo, R., Guigo, R., Bilbao, D., and Valcarcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell*, 19(4):475–84.

Jacquenet, S., Mereau, A., Bilodeau, P. S., Damier, L., Stoltzfus, C. M., and Branlant, C. (2001). A second exon splicing silencer within human immunodeficiency virus type 1 tat exon 2 represses splicing of Tat mRNA and binds protein hnRNP H. *J Biol Chem*, 276(44):40464–75.

Jensen, K. B., Dredge, B. K., Stefani, G., Zhong, R., Buckanovich, R. J., Okano, H. J., Yang, Y. Y., and Darnell, R. B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25(2):359–71.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653):2141–4.

Johnson, P. J. (2002). Spliceosomal introns in a deep-branching eukaryote: the splice of life. *Proc Natl Acad Sci U S A*, 99(6):3359–61.

Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*, 19(2):68–72.

Jung, D. J., Na, S. Y., Na, D. S., and Lee, J. W. (2002). Molecular cloning and characterization of CAPER, a novel coactivator of activating protein-1 and estrogen receptors. *J Biol Chem*, 277(2):1229–34.

Jurica, M. S. and Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell*, 12(1):5–14.

Kajita, Y., Nakayama, J., Aizawa, M., and Ishikawa, F. (1995). The UUAG-specific RNA binding protein, heterogeneous nuclear ribonucleoprotein D0. Common modular structure and binding properties of the 2xRBD-Gly family. *J Biol Chem*, 270(38):22167–75.

Kalyna, M. and Barta, A. (2004). A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? *Biochem Soc Trans*, 32(Pt 4):561–4.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H., and Gingeras, T. R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*, 14(3):331–42.

Kanaar, R., Roche, S. E., Beall, E. L., Green, M. R., and Rio, D. C. (1993). The conserved pre-mRNA splicing factor U2AF from Drosophila: requirement for viability. *Science*, 262(5133):569–73.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2005). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, 33(Database issue):D29–33.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., and Kent, W. J. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–4.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–6.

Kaufer, N. F. and Potashkin, J. (2000). Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Res*, 28(16):3003–10.

Kazazian, H. H., J. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–32.

Keegan, L. P., Gallo, A., and O'Connell, M. A. (2001). The many roles of an RNA editor. *Nat Rev Genet*, 2(11):869–78.

Keene, J. D. and Tenenbaum, S. A. (2002). Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell*, 9(6):1161–7.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–64.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, 12(6):996–1006.

Kielkopf, C. L., Lucke, S., and Green, M. R. (2004). U2AF homology motifs: protein recognition in the RRM world. *Genes Dev*, 18(13):1513–26.

Kielkopf, C. L., Rodionova, N. A., Green, M. R., and Burley, S. K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell*, 106(5):595–605.

Kiledjian, M. and Dreyfuss, G. (1992). Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *Embo J*, 11(7):2655–64.

Kim, H., Klein, R., Majewski, J., and Ott, J. (2004). Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*, 36(9):915–6; author reply 916–7.

Kitagawa, K., Wang, X., Hatada, I., Yamaoka, T., Nojima, H., Inazawa, J., Abe, T., Mitsuya, K., Oshimura, M., Murata, A., and et al. (1995). Isolation and mapping of human homologues of an imprinted mouse gene U2af1-rs1. *Genomics*, 30(2):257–63.

Knudsen, S. (2002). *A biologist's guide to analysis of DNA microarray data*. Wiley-Interscience, New York.

Kopelman, N. M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet*, 37(6):588–9.

Kramer, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem*, 65:367–409.

Krammer, P. H. (2000). CD95's deadly mission in the immune system. *Nature*, 407(6805):789–95.

Krawczak, M., Reiss, J., and Cooper, D. N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet*, 90(1-2):41–54.

Krecic, A. M. and Swanson, M. S. (1999). hnRNP complexes: composition, structure, and function. *Curr Opin Cell Biol*, 11(3):363–71.

Ladd, A. N., Charlet, N., and Cooper, T. A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Mol Cell Biol*, 21(4):1285–96.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lareau, L. F., Green, R. E., Bhatnagar, R. S., and Brenner, S. E. (2004). The evolving roles of alternative splicing. *Curr Opin Struct Biol*, 14(3):273–82.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14.

Le Hir, H., Nott, A., and Moore, M. J. (2003). How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci*, 28(4):215–20.

Leffers, H., Dejgaard, K., and Celis, J. E. (1995). Characterisation of two major cellular poly(rC)-binding human proteins, each containing three K-homologous (KH) domains. *Eur J Biochem*, 230(2):447–53.

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 Database issue:D142–4.

Levine, A. and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res*, 29(19):4006–13.

Lipscombe, D. (2005). Neuronal proteins custom designed by alternative splicing. *Curr Opin Neurobiol*, 15(3):358–63.

Lisbin, M. J., Qiu, J., and White, K. (2001). The neuron-specific RNA-binding protein ELAV regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev*, 15(19):2546–61.

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(2):333–44.

Liu, H. X., Cartegni, L., Zhang, M. Q., and Krainer, A. R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet*, 27(1):55–8.

Liu, H. X., Chew, S. L., Cartegni, L., Zhang, M. Q., and Krainer, A. R. (2000a). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol*, 20(3):1063–71.

Liu, H. X., Zhang, M., and Krainer, A. R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev*, 12(13):1998–2012.

Liu, J., He, L., Collins, I., Ge, H., Libutti, D., Li, J., Egly, J. M., and Levens, D. (2000b). The FBP interacting repressor targets TFIIH to inhibit activated transcription. *Mol Cell*, 5(2):331–41.

Longman, D., Johnstone, I. L., and Caceres, J. F. (2000). Functional characterization of SR and SR-related genes in Caenorhabditis elegans. *Embo J*, 19(7):1625–37.

Loo, L. W., Grove, D. I., Williams, E. M., Neal, C. L., Cousens, L. A., Schubert, E. L., Holcomb, I. N., Massa, H. F., Glogovac, J., Li, C. I., Malone, K. E., Daling, J. R., Delrow, J. J., Trask, B. J., Hsu, L., and Porter, P. L. (2004). Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res*, 64(23):8541–9.

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

Lopreato, G. F., Lu, Y., Southwell, A., Atkinson, N. S., Hillis, D. M., Wilcox, T. P., and Zakon, H. H. (2001). Evolution and divergence of sodium channel genes in vertebrates. *Proc Natl Acad Sci U S A*, 98(13):7588–92.

Lou, H., Helfman, D. M., Gagel, R. F., and Berget, S. M. (1999). Polypyrimidine tract-binding protein positively regulates inclusion of an alternative 3'-terminal exon. *Mol Cell Biol*, 19(1):78–85.

Lou, H., Neugebauer, K. M., Gagel, R. F., and Berget, S. M. (1998). Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol Cell Biol*, 18(9):4977–85.

Lu, X., Timchenko, N. A., and Timchenko, L. T. (1999). Cardiac elav-type RNA-binding protein (ETR-3) binds to RNA CUG repeats expanded in myotonic dystrophy. *Hum Mol Genet*, 8(1):53–60.

Luhrmann, R., Kastner, B., and Bach, M. (1990). Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. *Biochim Biophys Acta*, 1087(3):265–92.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–5.

Makeyev, A. V., Chkheidze, A. N., and Liebhaber, S. A. (1999). A set of highly conserved RNA-binding proteins, alphaCP-1 and alphaCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. *J Biol Chem*, 274(35):24849–57.

Maniatis, T. and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, 418(6894):236–43.

Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–98.

Matsushita, K., Tomonaga, T., Shimada, H., Shioya, A., Higashi, M., Matsubara, H., Harigaya, K., Nomura, F., Libutti, D., Levens, D., and Ochiai, T. (2006). An essential role of alternative splicing of c-myc suppressor FUSE-binding protein-interacting repressor in carcinogenesis. *Cancer Res*, 66(3):1409–17.

Matunis, M. J., Xing, J., and Dreyfuss, G. (1994). The hnRNP F protein: unique primary structure, nucleic acid-binding properties, and subcellular localization. *Nucleic Acids Res*, 22(6):1059–67.

Mazet, F. and Shimeld, S. M. (2002). Gene duplication and divergence in the early evolution of vertebrates. *Curr Opin Genet Dev*, 12(4):393–6.

McEwen, G. K., Woolfe, A., Goode, D., Vavouri, T., Callaway, H., and Elgar, G. (2006). Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res*.

McLysaght, A., Hokamp, K., and Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet*, 31(2):200–4.

Merendino, L., Guth, S., Bilbao, D., Martinez, C., and Valcarcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature*, 402(6763):838–41.

Millard, S. S., Vidal, A., Markus, M., and Koff, A. (2000). A U-rich element in the 5' untranslated region is necessary for the translation of p27 mRNA. *Mol Cell Biol*, 20(16):5947–59.

Miller, L. D. (2004). An Overview of DNA Microarrays: from Technology to Biology and Beyond. *National University of Singapore*.

Miller, W., Makova, K. D., Nekrutenko, A., and Hardison, R. C. (2004). Comparative genomics. *Annu Rev Genomics Hum Genet*, 5:15–56.

Min, H., Turck, C. W., Nikolic, J. M., and Black, D. L. (1997). A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer. *Genes Dev*, 11(8):1023–36.

Modafferi, E. F. and Black, D. L. (1999). Combinatorial control of a neuron-specific exon. *Rna*, 5(5):687–706.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nat Genet*, 30(1):13–9.

Mollet, I., Barbosa-Morais, N. L., Andrade, J., and Carmo-Fonseca, M. (2006). Diversity of human U2AF splicing factors. *(Submitted)*.

Moore, M. J. (2002). Nuclear RNA turnover. *Cell*, 108(4):431–4.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. *Science*, 309(5740):1514–8.

Morris, B. J., Adams, D. J., Beveridge, D. J., van der Weyden, L., Mangs, H., and Leedman, P. J. (2004). cAMP controls human renin mRNA stability via specific RNA-binding proteins. *Acta Physiol Scand*, 181(4):369–73.

Mount, S. M. and Salz, H. K. (2000). Pre-messenger RNA processing factors in the Drosophila genome. *J Cell Biol*, 150(2):F37–44.

Myer, V. E. and Steitz, J. A. (1995). Isolation and characterization of a novel, low abundance hnRNP protein: A0. *Rna*, 1(2):171–82.

Nabetani, A., Hatada, I., Morisaki, H., Oshimura, M., and Mukai, T. (1997). Mouse U2af1-rs1 is a neomorphic imprinted gene. *Mol Cell Biol*, 17(2):789–98.

Naderi, A., Ahmed, A. A., Barbosa-Morais, N. L., Aparicio, S., Brenton, J. D., and Caldas, C. (2004). Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling. *BMC Genomics*, 5(1):9.

Naderi, A., Teschendorff, A. E., Pinder, S. E., Barbosa-Morais, N. L., Paish, C. E., Ellis, I. O., Brenton, J. D., and Caldas, C. (2006). Microarray Expression Signature predicts the outcome of Postmenopausal patients with Breast Cancer. *Oncogene*, (in press).

Nagengast, A. A., Stitzinger, S. M., Tseng, C. H., Mount, S. M., and Salz, H. K. (2003). Sex-lethal splicing autoregulation in vivo: interactions between SEX-LETHAL, the U1 snRNP and U2AF underlie male exon skipping. *Development*, 130(3):463–71.

Nasim, M. T., Chernova, T. K., Chowdhury, H. M., Yue, B. G., and Eperon, I. C. (2003). HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet*, 12(11):1337–48.

Neafsey, D. E. and Palumbi, S. R. (2003). Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. *Genome Res*, 13(5):821–30.

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet*, 30:371–403.

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press, Oxford ; New York. Masatoshi Nei, Sudhir Kumar. ill. ; 27 cm.

Nei, M. and Rooney, A. P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families (*). *Annu Rev Genet*, 39:121–152.

Nessling, M., Richter, K., Schwaenen, C., Roerig, P., Wrobel, G., Wessendorf, S., Fritz, B., Bentz, M., Sinn, H. P., Radlwimmer, B., and Lichter, P. (2005). Candidate genes in breast cancer revealed by microarray-based comparative genomic hybridization of archived tissue. *Cancer Res*, 65(2):439–47.

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet*, 20(1):46–50.

Ngo, J. C., Chakrabarti, S., Ding, J. H., Velazquez-Dones, A., Nolen, B., Aubol, B. E., Adams, J. A., Fu, X. D., and Ghosh, G. (2005). Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Mol Cell*, 20(1):77–89.

Nilsen, T. W. (2001). Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet*, 17(12):678–80.

Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25(12):1147–9.

Nissim-Rafinia, M. and Kerem, B. (2002). Splicing regulation as a potential genetic modifier. *Trends Genet*, 18(3):123–7.

Nissim-Rafinia, M. and Kerem, B. (2005). The splicing machinery is a genetic modifier of disease severity. *Trends Genet*, 21(9):480–3.

Nixon, J. E., Wang, A., Morrison, H. G., McArthur, A. G., Sogin, M. L., Loftus, B. J., and Samuelson, J. (2002). A spliceosomal intron in Giardia lamblia. *Proc Natl Acad Sci U S A*, 99(6):3701–5.

Nornes, S., Clarkson, M., Mikkola, I., Pedersen, M., Bardsley, A., Martinez, J. P., Krauss, S., and Johansen, T. (1998). Zebrafish contains two pax6 genes involved in eye development. *Mech Dev*, 77(2):185–96.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.

Orphanides, G. and Reinberg, D. (2002). A unified theory of gene expression. *Cell*, 108(4):439–51.

Ostareck-Lederer, A., Ostareck, D. H., and Hentze, M. W. (1998). Cytoplasmic regulatory functions of the KH-domain proteins hnRNPs K and E1/E2. *Trends Biochem Sci*, 23(11):409–11.

Ostertag, E. M. and Kazazian, H. H., J. (2001). Biology of mammalian L1 retrotransposons. *Annu Rev Genet*, 35:501–38.

Ostrowski, J., Kawata, Y., Schullery, D. S., Denisenko, O. N., Higaki, Y., Abrass, C. K., and Bomsztyk, K. (2001). Insulin alters heterogeneous nuclear ribonucleoprotein K protein binding to DNA and RNA. *Proc Natl Acad Sci U S A*, 98(16):9044–9.

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(26):27039–49.

Pacheco, T. R., Moita, L. F., Gomes, A. Q., Hacohen, N., and Carmo-Fonseca, M. (2006). RNAi Knockdown of hU2AF35 Impairs Cell Cycle Progression and Modulates Alternative Splicing of Cdc25 Transcripts. *Mol Biol Cell*, (in press).

Pagani, F. and Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet*, 5(5):389–96.

Page-McCaw, P. S., Amonlirdviman, K., and Sharp, P. A. (1999). PUF60: a novel U2AF65-related splicing activity. *Rna*, 5(12):1548–60.

Park, J. W., Parisky, K., Celotto, A. M., Reenan, R. A., and Graveley, B. R. (2004). Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc Natl Acad Sci U S A*, 101(45):15974–9.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003). *The analysis of gene expression data : methods and software*. Statistics for biology and health. Springer, New York. Giovanni Parmigiani ... [et al.] editors. ill. ; 25 cm.

Patel, A. A. and Steitz, J. A. (2003). Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, 4(12):960–70.

Patel, N. H. and Prince, V. E. (2000). Beyond the Hox complex. *Genome Biol*, 1(5):REVIEWS1027.

Pearsall, R. S., Shibata, H., Brozowska, A., Yoshino, K., Okuda, K., deJong, P. J., Plass, C., Chapman, V. M., Hayashizaki, Y., and Held, W. A. (1996). Absence of imprinting in U2AFBPL, a human homologue of the imprinted mouse gene U2afbp-rs. *Biochem Biophys Res Commun*, 222(1):171–7.

Perez, I., Lin, C. H., McAfee, J. G., and Patton, J. G. (1997). Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. *Rna*, 3(7):764–78.

Persson, P. B., Skalweit, A., Mrowka, R., and Thiele, B. J. (2003). Control of renin synthesis. *Am J Physiol Regul Integr Comp Physiol*, 285(3):R491–7.

Pollard, A. J., Krainer, A. R., Robson, S. C., and Europe-Finner, G. N. (2002). Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice Site. *J Biol Chem*, 277(18):15241–51.

Potashkin, J., Naik, K., and Wentz-Hunter, K. (1993). U2AF homolog required for splicing in vivo. *Science*, 262(5133):573–5.

Prasad, J., Colwill, K., Pawson, T., and Manley, J. L. (1999). The protein kinase Clk/Sty directly modulates SR protein activity: both hyper- and hypophosphorylation inhibit splicing. *Mol Cell Biol*, 19(10):6991–7000.

Proudfoot, N. J., Furger, A., and Dye, M. J. (2002). Integrating mRNA processing with transcription. *Cell*, 108(4):501–12.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–4.

Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res*, 12(8):1231–45.

Reddy, A. S. (2004). Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. *Trends Plant Sci*, 9(11):541–7.

Reed, R. (1989). The organization of 3' splice-site sequences in mammalian introns. *Genes Dev*, 3(12B):2113–23.

Reed, R. (1990). Protein composition of mammalian spliceosomes assembled in vitro. *Proc Natl Acad Sci U S A*, 87(20):8031–5.

Reed, R. and Magni, K. (2001). A new view of mRNA export: separating the wheat from the chaff. *Nat Cell Biol*, 3(9):E201–4.

Reimann, I., Huth, A., Thiele, H., and Thiele, B. J. (2002). Suppression of 15-lipoxygenase synthesis by hnRNP E1 is dependent on repetitive nature of LOX mRNA 3'-UTR control element DICE. *J Mol Biol*, 315(5):965–74.

Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R. B., and Valcarcel, J. (2005). Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*, 280(6):4779–84.

Relógio, A. M. B. (2002). *Analysis of alternative pre-mRNA splicing regulation using DNA microarrays*. Phd, EMBL Universidade de Lisboa.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–7.

Ringrose, L. and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet*, 38:413–43.

Robberson, B. L., Cote, G. J., and Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol*, 10(1):84–94.

Rocak, S. and Linder, P. (2004). DEAD-box proteins: the driving forces behind RNA metabolism. *Nat Rev Mol Cell Biol*, 5(3):232–41.

Roesler, J., Izquierdo, J. M., Ryser, M., Rosen-Wolff, A., Gahr, M., Valcarcel, J., Lenardo, M. J., and Zheng, L. (2005). Haploinsufficiency, rather than the effect of an excessive production of soluble CD95 (CD95DeltaTM), is the basis for ALPS Ia in a family with duplicated 3' splice site AG in CD95 intron 5 on one allele. *Blood*, 106(5):1652–9.

Rudner, D. Z., Kanaar, R., Breger, K. S., and Rio, D. C. (1996). Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects. *Proc Natl Acad Sci U S A*, 93(19):10333–7.

Ruskin, B., Zamore, P. D., and Green, M. R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell*, 52(2):207–19.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–5.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–25.

Salgado-Garrido, J., Bragado-Nilsson, E., Kandels-Lewis, S., and Seraphin, B. (1999). Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *Embo J*, 18(12):3451–62.

Schaal, T. D. and Maniatis, T. (1999a). Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol*, 19(1):261–73.

Schaal, T. D. and Maniatis, T. (1999b). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol*, 19(3):1705–19.

Seraphin, B. (1995). Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *Embo J*, 14(9):2089–98.

Sharp, P. A. (1994). Split genes and RNA splicing. *Cell*, 77(6):805–15.

Shatkin, A. J. and Manley, J. L. (2000). The ends of the affair: capping and polyadenylation. *Nat Struct Biol*, 7(10):838–42.

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(3):363–73.

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(3):367–76.

Shepard, J., Reick, M., Olson, S., and Graveley, B. R. (2002). Characterization of U2AF(6), a splicing factor related to U2AF(35). *Mol Cell Biol*, 22(1):221–30.

Shi, H., Hoffman, B. E., and Lis, J. T. (1997). A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52. *Mol Cell Biol*, 17(5):2649–57.

Shih, S. C. and Claffey, K. P. (1999). Regulation of human vascular endothelial growth factor mRNA stability in hypoxia by heterogeneous nuclear ribonucleoprotein L. *J Biol Chem*, 274(3):1359–65.

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 99(21):13627–32.

Simpson, A. G., MacQuarrie, E. K., and Roger, A. J. (2002). Eukaryotic evolution: early origin of canonical introns. *Nature*, 419(6904):270.

Singh, R., Valcarcel, J., and Green, M. R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, 268(5214):1173–6.

Smith, C. W. and Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci*, 25(8):381–8.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical taxonomy; the principles and practice of numerical classification*. W. H. Freeman, San Francisco,. [by] Peter H. A. Sneath [and] Robert R. Sokal. illus. 26 cm. A Series of books in biology.

Sokolowski, M., Furneaux, H., and Schwartz, S. (1999). The inhibitory activity of the AU-rich RNA element in the human papillomavirus type 1 late 3' untranslated region correlates with its affinity for the elav-like HuR protein. *J Virol*, 73(2):1080–91.

Soltaninassab, S. R., McAfee, J. G., Shahied-Milam, L., and LeStourgeon, W. M. (1998). Oligonucleotide binding specificities of the hnRNP C protein tetramer. *Nucleic Acids Res*, 26(14):3410–7.

Soret, J., Gattoni, R., Guyon, C., Sureau, A., Popielarz, M., Le Rouzic, E., Dumon, S., Apiou, F., Dutrillaux, B., Voss, H., Ansorge, W., Stevenin, J., and Perbal, B. (1998). Characterization of SRp46, a novel human SR splicing factor encoded by a PR264/SC35 retropseudogene. *Mol Cell Biol*, 18(8):4924–34.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A. L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–23.

Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–8.

Soulard, M., Della Valle, V., Siomi, M. C., Pinol-Roma, S., Codogno, P., Bauvy, C., Bellini, M., Lacroix, J. C., Monod, G., Dreyfuss, G., and et al. (1993). hnRNP G: sequence and characterization of a glycosylated RNA-binding protein. *Nucleic Acids Res*, 21(18):4210–7.

Spangberg, K., Wiklund, L., and Schwartz, S. (2000). HuR, a protein implicated in oncogene and growth factor mRNA decay, binds to the 3' ends of hepatitis C virus RNA of both polarities. *Virology*, 274(2):378–90.

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., and Birney, E. (2004). The Ensembl core software libraries. *Genome Res*, 14(5):929–33.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8.

Staley, J. P. and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, 92(3):315–26.

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(Database issue):D46–55.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nat Rev Genet*, 2(7):493–503.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P. E., Bussemaker, H. J., and White, K. P. (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, 306(5696):655–60.

Stoughton, R. B. (2005). Applications of DNA microarrays in biology. *Annu Rev Biochem*, 74:53–82.

Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., and Ares, M. (2006). Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol*, 2(1):e4.

Swanson, M. S. and Dreyfuss, G. (1988). Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities. *Mol Cell Biol*, 8(5):2237–41.

Tacke, R., Chen, Y., and Manley, J. L. (1997). Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci U S A*, 94(4):1148–53.

Tacke, R. and Manley, J. L. (1999). Determinants of SR protein specificity. *Curr Opin Cell Biol*, 11(3):358–62.

Tacke, R., Tohyama, M., Ogawa, S., and Manley, J. L. (1998). Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. *Cell*, 93(1):139–48.

Takahashi, N., Sasagawa, N., Suzuki, K., and Ishiura, S. (2000). The CUG-binding protein binds specifically to UG dinucleotide repeats in a yeast three-hybrid system. *Biochem Biophys Res Commun*, 277(2):518–23.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol*, 12(5):823–33.

Tarn, W. Y. and Steitz, J. A. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, 84(5):801–11.

Tarn, W. Y. and Steitz, J. A. (1997). Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem Sci*, 22(4):132–7.

Teraoka, S. N., Telatar, M., Becker-Catania, S., Liang, T., Onengut, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R. A., and Concannon, P. (1999). Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am J Hum Genet*, 64(6):1617–31.

Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L., and Caldas, C. (2006a). PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer. *Bioinformatics*.

Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L., Pinder, S. E., Ellis, I. O., Aparicio, S., Brenton, J. D., and Caldas, C. (2006b). A consensus molecular prognostic classifier for ER positive breast cancer. *(Submitted)*.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., and Caldas, C. (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21(13):3025–33.

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(Database issue):D64–9.

Thisted, T., Lyakhov, D. L., and Liebhaber, S. A. (2001). Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and alphaCP-2KL, suggest Distinct modes of RNA recognition. *J Biol Chem*, 276(20):17484–96.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Tian, Q., Streuli, M., Saito, H., Schlossman, S. F., and Anderson, P. (1991). A polyadenylate binding protein localized to the granules of cytolytic lymphocytes induces DNA fragmentation in target cells. *Cell*, 67(3):629–39.

Tian, Q., Taupin, J., Elledge, S., Robertson, M., and Anderson, P. (1995). Fas-activated serine/threonine kinase (FAST) phosphorylates TIA-1 during Fas-mediated apoptosis. *J Exp Med*, 182(3):865–74.

Tisdall, J. D. (2001). *Beginning Perl for bioinformatics*. O'Reilly, Beijing ; Sebastopol, CA, 1st edition.

Tronchere, H., Wang, J., and Fu, X. D. (1997). A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA. *Nature*, 388(6640):397–400.

Tupler, R., Perini, G., and Green, M. R. (2001). Expressing the human genome. *Nature*, 409(6822):832–3.

Valcarcel, J., Gaur, R. K., Singh, R., and Green, M. R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science*, 273(5282):1706–9.

Van Buskirk, C. and Schupbach, T. (2002). Half pint regulates alternative splice site selection in Drosophila. *Dev Cell*, 2(3):343–53.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009.

Van Seuningen, I., Ostrowski, J., and Bomsztyk, K. (1995). Description of an IL-1-responsive kinase that phosphorylates the K protein. Enhancement of phosphorylation by selective DNA and RNA motifs. *Biochemistry*, 34(16):5644–50.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6.

Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A., and Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A*, 101(6):1638–43.

Vavouri, T., McEwen, G. K., Woolfe, A., Gilks, W. R., and Elgar, G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet*, 22(1):5–10.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J.,

Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–51.

Wang, Y., Joh, K., Masuko, S., Yatsuki, H., Soejima, H., Nabetani, A., Beechey, C. V., Okinami, S., and Mukai, T. (2004a). The mouse Murr1 gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented U2af1-rs1 gene. *Mol Cell Biol*, 24(1):270–9.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–9.

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., and Burge, C. B. (2004b). Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45.

Washburn, M. P., Wolters, D., and Yates, J. R., r. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242–7.

Wentz-Hunter, K. and Potashkin, J. (1996). The small subunit of the splicing factor U2AF is conserved in fission yeast. *Nucleic Acids Res*, 24(10):1849–54.

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(1):28–33.

Will, C. L. and Luhrmann, R. (2001). Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol*, 13(3):290–301.

Will, C. L., Schneider, C., Reed, R., and Luhrmann, R. (1999). Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science*, 284(5422):2003–5.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, 379(6564):466–9.

Wollerton, M. C., Gooding, C., Robinson, F., Brown, E. C., Jackson, R. J., and Smith, C. W. (2001). Differential alternative splicing activity of isoforms of polypyrimidine tract binding protein (PTB). *Rna*, 7(6):819–32.

Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., and Smith, C. W. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell*, 13(1):91–100.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., et al. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature*, 415(6874):871–80.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J., Cooke, J. E., and Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7.

Woychik, N. A. and Hampsey, M. (2002). The RNA polymerase II machinery: structure illuminates function. *Cell*, 108(4):453–63.

Wu, S., Romfo, C. M., Nilsen, T. W., and Green, M. R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature*, 402(6763):832–5.

Yan, P. S., Efferth, T., Chen, H. L., Lin, J., Rodel, F., Fuzesi, L., and Huang, T. H. (2002). Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. *Methods*, 27(2):162–9.

Yeo, G., Hoon, S., Venkatesh, B., and Burge, C. B. (2004). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*, 101(44):15700–5.

Yu, W. P., Brenner, S., and Venkatesh, B. (2003). Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. *Trends Genet*, 19(4):180–3.

Zamore, P. D. and Green, M. R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci U S A*, 86(23):9243–7.

Zamore, P. D., Patton, J. G., and Green, M. R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, 355(6361):609–14.

Zhang, L., Liu, W., and Grabowski, P. J. (1999). Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. *Rna*, 5(1):117–30.

Zhang, M., Zamore, P. D., Carmo-Fonseca, M., Lamond, A. I., and Green, M. R. (1992). Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. *Proc Natl Acad Sci U S A*, 89(18):8769–73.

Zhang, W., Liu, H., Han, K., and Grabowski, P. J. (2002). Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. *Rna*, 8(5):671–85.

Zhang, X. H. and Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50.

Zhang, Z., Carriero, N., and Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet*, 20(2):62–7.

Zhang, Z., Harrison, P. M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res*, 13(12):2541–58.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev*, 63(2):405–45.

Zhou, Z., Licklider, L. J., Gygi, S. P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–5.

Zhu, W. and Brendel, V. (2003). Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. *Nucleic Acids Res*, 31(15):4561–72.

Zorio, D. A. and Blumenthal, T. (1999a). Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature*, 402(6763):835–8.

Zorio, D. A. and Blumenthal, T. (1999b). U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. *Rna*, 5(4):487–94.

Zuo, P. and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev*, 10(11):1356–68.

Web Site References

http://www.ensembl.org ; Ensembl

http://www.genome.ucsc.edu ; UCSC Genome Browser

http://www.bioperl.org ; BioPerl

http://hmmer.wustl.edu ; HMMER - Biological sequence analysis using profile hidden Markov models

http://www.fao.org ; Food a Agriculture Organization of the United Nations

http://www.ebi.ac.uk/Wise2 ; Wise2 - Intelligent algorithms for DNA searches (EBI)

http://woody.embl-heidelberg.de/gene2est ; Gene2EST BLAST Server

http://www.ncbi.nlm.nih.gov/BLAST ; NCBI BLAST

http://www.es.embnet.org/Doc/SNAP ; SNAP.pl (Synonymous Nonsynonymous Analysis Program)

http://www.repeatmasker.org ; RepeatMasker

http://www.sanger.ac.uk/Software/Pfam ; Pfam - Protein families database of alignments and HMMs

http://smart.embl-heidelberg.de ; SMART - Simple Modular Architecture Research Tool

http://www.gene.ucl.ac.uk/nomenclature/ ; HUGO Gene Nomenclature Committee

http://www.pymol.org ; The PyMOL Molecular Graphics System

http://www.ncbi.nlm.nih.gov/RefSeq ; NCBI Reference Sequence (RefSeq)

http://bioinfo.mbi.ucla.edu/ASAP/ ; ASAP

http://hollywood.mit.edu ; Hollywood RNA Alternative Splicing Database

http://rulai.cshl.edu/tools/ESE/ ; ESEfinder
http://genes.mit.edu/burgelab/rescue-ese/ ; RESCUE-ESE Web Server
http://www.ebi.ac.uk/asd-srv/wb.cgi ; ASD - Alternative Splicing Workbench
http://www.ebi.ac.uk/asd-srv/wb.cgi?method=8 ; Splicing Rainbow
http://www.ncbi.nlm.nih.gov/UniGene ; UniGene
http://www.ebi.ac.uk/embl ; EMBL Nucleotide Sequence Database
http://derlab.med.utoronto.ca/CpGIslands/ ; University Health Network Microarray Centre, Toronto - Der Laboratory - CpG Island Microarray Bioinformatics
http://www.vysis.com/PDF/GenoSensor300ClonesAndKey_July2004.pdf ; Vysis Genosensor Array 300 Clone Annotation
http://www.sanger.ac.uk/HGP/cgi.shtml ; Sanger Centre - CpG Island Tagging Project
http://www.genome.org/ ; Genome Research
http://us.expasy.org/sprot ; Swiss-Prot and TrEMBL
http://www.jgi.doe.gov; DOE Joint Genome Institute
http://www.sanger.ac.uk/Projects/S_pombe ; The Sanger Institute - The S. pombe Genome Project
http://www.yeastgenome.org ; Saccharomyces Genome Database
http://plasmodb.org ; PlasmoDB - The Plasmodium Genome Resource
http://www.sanger.ac.uk/Projects/T_brucei ; The Sanger Institute - The Trypanosoma brucei Genome Project
http://tcruzidb.org ; TcruziDB - The Trypanosoma cruzi Genome Resource
http://www.ncbi.nlm.nih.gov ; NCBI - National Center of Biotechnology Information
http://www.iupac.org ; International Union of Pure and Applied Chemistry

Appendix

Appendix A

Supplementary information

A.1 Selective expansion of splicing regulatory factors

This section presents the supplementary tables associated with the work described in Chapter 2, except for a table entitled "Putative eukaryotic (and archaeal) splicing factors identified by the pipeline", not shown here due to its size (1920 rows). The missing table, all the phylogenetic trees and alignments and the original files for the tables presented in this section can be found, as Supplemental Material, on the Genome Research website (where the work is published [Barbosa-Morais et al., 2006]):

`http://www.genome.org/`

A.1.1 Human splicing factors and splicing related proteins

Method:

254 human splicing factors and splicing-related proteins were initially identified in a splicing factors database [Relogio et al., 2005], in the literature [Burge et al., 1999; Black, 2003; Hartmuth et al., 2002; Jurica and Moore, 2003; Neubauer et al., 1998; Rappsilber et al., 2002; Zhou et al., 2002] and by searching SwissProt [Boeckmann et al., 2003] (v47.2, `http://us.expasy.org/sprot/`) with appropriate keywords. This search also provided many splicing factors for other species.

Table A.1: Human splicing factors and splicing related proteins

Functional Group	SwissProt Accession	Description
U1 + U2 snRNP	P09012	U1 small nuclear ribonucleoprotein A (U1 snRNP A protein)
	P08579	U2 small nuclear ribonucleoprotein B"
	P09234	U1 small nuclear ribonucleoprotein C (U1-C)
	P08621	U1 small nuclear ribonucleoprotein 70 kDa (U1 snRNP 70 kDa) (snRNP70) (U1-70K)
	O75400	Formin binding protein 3 (Huntingtin yeast partner A) (Huntingtin-interacting protein HYPA/FBP11) (Fas-ligand associated factor 1) (NY-REN-6 antigen) (HSPC225)
	Q8NCZ1	Hypothetical protein DKFZp434O1520
	P09661	U2 small nuclear ribonucleoprotein A' (U2 snRNP-A')
	Q15459	Splicing factor 3 subunit 1 (Spliceosome associated protein 114) (SAP 114) (SF3a120)
	Q15428	Splicing factor 3A subunit 2 (Spliceosome associated protein 62) (SAP 62) (SF3a66)
	Q12874	Splicing factor 3A subunit 3 (Spliceosome associated protein 61) (SAP 61) (SF3a60)
	O75533	Splicing factor 3B subunit 1 (Spliceosome associated protein 155) (SAP 155) (SF3b155) (Pre-mRNA splicing factor SF3b 155 kDa subunit)
	Q13435	Splicing factor 3B subunit 2 (Spliceosome associated protein 145) (SAP 145) (SF3b150) (Pre-mRNA splicing factor SF3b 145 kDa subunit)
	Q15393	Splicing factor 3B subunit 3 (Spliceosome associated protein 130) (SAP 130) (SF3b130) (Pre-mRNA splicing factor SF3b 130 kDa subunit)
	Q15427	Splicing factor 3B subunit 4 (Spliceosome associated protein 49) (SAP 49) (SF3b50) (Pre-mRNA splicing factor SF3b 49 kDa subunit)
	Q9Y3B4	Pre-mRNA branch site protein p14 (CGI-110) (HSPC175) (Ht006)
U4/U6 snRNP	O43395	U4/U6 small nuclear ribonucleoprotein Prp3 (Pre-mRNA splicing factor 3) (U4/U6 snRNP 90 kDa protein) (hPrp3)
	O43172	U4/U6 small nuclear ribonucleoprotein Prp4 (U4/U6 snRNP 60 kDa protein) (WD splicing factor Prp4) (hPrp4).
	O43447	Peptidyl-prolyl cis-trans isomerase H (EC 5.2.1.8) (PPIase H) (Rotamase H) (U-snRNP-associated cyclophilin SnuCyp-20) (USA-CYP) (Small nuclear ribonucleoprotein particle-specific cyclophilin H) (CypH)
U5 snRNP	O95320	U5 snRNP-specific 40 kDa protein
	O75643	U5 small nuclear ribonucleoprotein 200 kDa helicase (U5 snRNP-specific 200 kDa protein) (U5-200KD)
	Q15029	116 kDa U5 small nuclear ribonucleoprotein component (U5 snRNP- specific protein, 116 kDa) (U5-116 kDa)
	O94906	U5 snRNP-associated 102 kDa protein (U5-102 kDa protein)
	O43188	Prp28, U5 snRNP 100 kDa protein
	O14834	Spliceosomal U5 snRNP-specific 15 kDa protein (DIM1 protein homolog) (Thioredoxin-like U5 snRNP protein U5-15kD)
	O14547	PRP8 protein
U4/U6.U5 tri-snRNP	O43290	SART-1 (Squamous cell carcinoma antigen RECOGNISED BY T cells) (U4/U6.U5 TRI-snRNP-associated 110 kDa protein)
	P55769	NHP2-like protein 1 (High mobility group-like nuclear protein 2 homolog 1) ([U4/U6.U5] tri-snRNP 15.5 kDa protein) (OTK27)
	Q96RK9	U4/U6.U5 tri-snRNP-associated 65 kDa protein
U11 + U12 snRNP	Q9UDW3	Hypothetical protein (U11/U12 snRNP 20K) (Em:AC005529.5 protein) (LOC55954 protein)
	Q9BV90	U11/U12 snRNP 25K protein (Minus-99 protein)
	Q96TA6	MADP-1 protein (U11/U12 snRNP 31K)
	Q16560	U1-snRNP binding protein homolog (U11/U12 snRNP 35K, isoform a).
	Q6IEG0	U11/U12 snRNP 48K
	Q96LT9	Hypothetical protein FLJ25070 (U11/U12 snRNP 65K) (RNA recognition protein) (Novel protein)
Sm	P14678	Small nuclear ribonucleoprotein associated proteins B and B' (snRNP-B) (Sm protein B/B') (Sm-B/Sm-B') (SmB/SmB')
	P14648	Small nuclear ribonucleoprotein associated protein N (snRNP-N) (Sm protein N) (Sm-N) (SmN) (Sm-D) (Tissue-specific splicing protein)
	P13641	Small nuclear ribonucleoprotein Sm D1 (snRNP core protein D1) (Sm-D1) (Sm-D autoantigen)
	P43330	Small nuclear ribonucleoprotein Sm D2 (snRNP core protein D2) (Sm-D2)
	P43331	Small nuclear ribonucleoprotein Sm D3 (snRNP core protein D3) (Sm-D3)
	P08578	Small nuclear ribonucleoprotein E (snRNP-E) (Sm protein E) (Sm-E) (SmE)
	Q15356	Small nuclear ribonucleoprotein F (snRNP-F) (Sm protein F) (Sm-F) (SmF)
	Q15357	Small nuclear ribonucleoprotein G (snRNP-G) (Sm protein G) (Sm-G) (SmG)
	O15116	U6 snRNA-associated Sm-like protein LSm1
	Q9Y333	U6 snRNA-associated Sm-like protein LSm2 (Small nuclear ribonuclear protein D homolog) (G7b) (SnRNP core SM-like protein SM-x5)
	Q9Y4Z1	U6 snRNA-associated Sm-like protein LSm3 (MDS017)
	Q9Y4Z0	U6 snRNA-associated Sm-like protein LSm4 (Glycine-rich protein) (GRP)
	Q9Y4Y9	U6 snRNA-associated Sm-like protein LSm5
	Q9Y4Y8	U6 snRNA-associated Sm-like protein LSm6
	Q9UK45	U6 snRNA-associated Sm-like protein LSm7
	O95777	U6 snRNA-associated Sm-like protein LSm8
	Q969L4	U7 snRNA-associated Sm-like protein LSm10
	Q8N4M0	Hypothetical protein
U2AF	P26368	Splicing factor U2AF 65 kDa subunit (U2 auxiliary factor 65 kDa subunit) (U2 snRNP auxiliary factor large subunit) (hU2AF(65))
	Q01081	Splicing factor U2AF 35 kDa subunit (U2 auxiliary factor 35 kDa subunit) (U2 snRNP auxiliary factor small subunit)
	Q8WU68	U2 AUXILIARY FACTOR 26
	Q15695	U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit related-protein 1
	Q15696	U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit related-protein 2
SR	Q9UQ35	RNA binding protein
	Q15410	Nucleic acid binding protein (Fragment).
	Q01130	Splicing factor, arginine/serine-rich 2 (Splicing factor SC35) (SC-35) (Splicing component, 35 kDa) (PR264 protein)
	Q9BRL6	Similar to splicing factor, arginine/serine-rich 2 (SC-35) (SRp46 splicing factor)
	P23152	Splicing factor, arginine/serine-rich 3 (Pre-mRNA splicing factor SRP20) (X16 protein).
	Q16629	Splicing factor, arginine/serine-rich 7 (Splicing factor 9G8)
	Q13242	Splicing factor, arginine/serine-rich 9 (Pre-mRNA splicing factor SRp30C)
	Q07955	Splicing factor, arginine/serine-rich 1 (pre-mRNA splicing factor SF2, P33 subunit) (Alternative splicing factor ASF-1)
	Q13243	Splicing factor, arginine/serine-rich 5 (Pre-mRNA splicing factor SRP40) (Delayed-early protein HRS)
	Q13247	Splicing factor, arginine/serine-rich 6 (Pre-mRNA splicing factor SRP55)
	Q08170	Splicing factor, arginine/serine-rich 4 (Pre-mRNA splicing factor SRP75) (SRP001LB)
	Q05519	Splicing factor arginine/serine-rich 11 (Arginine-rich 54 kDa nuclear protein) (p54)
	Q8WXA9	Splicing factor, arginine/serine-rich 12 (Serine-arginine-rich splicing regulatory protein 86) (SRrp86) (Splicing regulatory protein 508) (SRrp508)
	Q13595	Transformer-2 protein homolog (TRA-2 alpha)
	Q15815	Arginine/serine-rich splicing factor 10 (Transformer-2-beta) (HTRA2- beta) (Transformer 2 protein homolog) (Silica-induced protein 41) (RA301)
	Q9UNR9	Topoisomerase I-binding RS protein
hnRNP	Q13151	Heterogeneous nuclear ribonucleoprotein A0 (hnRNP A0)
	P09651	Heterogeneous nuclear ribonucleoprotein A1 (Helix-destabilizing protein) (Single-strand binding protein) (hnRNP core protein A1)
	P22626	Heterogeneous nuclear ribonucleoproteins A2/B1 (hnRNP A2 / hnRNP B1)
	P51991	Heterogeneous nuclear ribonucleoprotein A3 (hnRNP A3) (D10S102)
	P07910	Heterogeneous nuclear ribonucleoproteins C1/C2 (hnRNP C1 / hnRNP C2).
	O60812	Heterogeneous nuclear ribonucleoprotein C-like dJ845O24.4 (hnRNP core protein C-like)
	Q9UKM9	RNA-binding protein Raly (hnRNP associated with lethal yellow homolog) (Autoantigen p542)
	Q8N1C2	LOC138046 protein
	Q14103	Heterogeneous nuclear ribonucleoprotein D0 (hnRNP D0) (AU-rich element RNA-binding protein 1)
	Q99729	Heterogeneous nuclear ribonucleoprotein A/B (hnRNP A/B) (APOBEC-1 binding protein 1) (ABBP-1)
	O14979	JKTBP2 (Heterogeneous nuclear ribonucleoprotein D-like) (Hypothetical protein) (HNRPDL protein)
	O43347	WUGSC:H_166H1.2 protein (Musashi)
	Q96DH6	Musashi 2, isoform a.
	Q15365	Poly(rC)-binding protein 1 (Alpha-CP1) (hnRNP-E1) (Nucleic acid binding protein SUB2.3)

	Q15366	Poly(rC)-binding protein 2 (Alpha-CP2) (hnRNP-E2)
	P57721	Poly(rC)-binding protein 3 (Alpha-CP3)
	P57723	Poly(rC)-binding protein 4 (Alpha-CP4)
	P52597	Heterogeneous nuclear ribonucleoprotein F (hnRNP F)
	P31943	Heterogeneous nuclear ribonucleoprotein H (hnRNP H)
	P55795	Heterogeneous nuclear ribonucleoprotein H' (hnRNP H') (FTP-3)
	P31942	Heterogeneous nuclear ribonucleoprotein H3 (hnRNP H3) (hnRNP 2H9)
	Q12849	G-rich sequence factor-1 (GRSF-1)
	P38159	Heterogeneous nuclear ribonucleoprotein G (hnRNP G) (RNA binding motif protein, X chromosome) (Glycoprotein p43)
	O75526	Testes specific heterogenous nuclear ribonucleoprotein G-T.
	Q14011	Cold-inducible RNA-binding protein (Glycine-rich RNA-binding protein CIRP) (A18 hnRNP)
	P98179	Putative RNA-binding protein 3 (RNA binding motif protein 3) (RNPL)
	P26599	Polypyrimidine tract-binding protein 1 (PTB) (Heterogeneous nuclear ribonucleoprotein I) (hnRNP I) (57 kDa RNA-binding protein PPTB-1)
	Q969N9	PTB-like protein L (Polypyrimidine tract binding protein 2)
	O95758	Rod1
	Q07244	Heterogeneous nuclear ribonucleoprotein K (hnRNP K) (DC-stretch binding protein) (CSBP)
	P14866	Heterogeneous nuclear ribonucleoprotein L (hnRNP L).
	Q8WVV9	Hypothetical protein
	P52272	Heterogeneous nuclear ribonucleoprotein M (hnRNP M)
	Q9H922	Myelin gene expression factor
	O60506	Gry-rbp (hnRNP Q3)
	O43390	Heterogeneous nuclear ribonucleoprotein R (hnRNP R)
	Q00839	Heterogenous nuclear ribonucleoprotein U (hnRNP U) (Scaffold attachment factor A) (SAF-A)
	O76022	E1B-55kDa-associated protein
	Q8N3B3	Hypothetical protein DKFZp762N1910 (Fragment)
TIA	P31483	Nucleolysin TIA-1 (RNA-binding protein TIA-1) (p40-TIA-1) [Contains: Nucleolysin TIA-1 isoform p15 (p15-TIA-1)]
	Q01085	Nucleolysin TIAR (TIA-1 related protein)
CELF/CUG-BP	Q92879	CUG triplet repeat RNA-binding protein 1 (CUG-BP1) (RNA-binding protein BRUNOL-2) (Deadenylation factor CUG-BP) (50 kDa Nuclear polyadenylated RNA-binding protein) (EDEN-BP)
	Q92950	Apoptosis-related RNA binding protein (ETR-3)
	Q9BZC0	CUG-BP and ETR-3 like factor 5
	Q9BZC1	CUG-BP and ETR-3 like factor 4
	Q9BZC2	CUG-BP and ETR-3 like factor 3
	Q96J87	BRUNO-like 6 RNA-binding protein (RNA-binding protein CELF6)
CLK	P49759	Protein kinase CLK1 (EC 2.7.1.-) (CLK)
	P49760	Protein kinase CLK2 (EC 2.7.1.-) (CDC-like kinase 2)
	P49761	Protein kinase CLK3 (EC 2.7.1.-) (CDC-like kinase 3)
	Q9HAZ1	Protein serine threonine kinase Clk4
SRPK	Q96SB4	SRPK1a protein kinase
	P78362	Serine kinase SRPK2
	Q9UPE1	Serine/threonine-protein kinase 23 (EC 2.7.1.37) (Muscle-specific serine kinase 1) (MSSK-1)
hprp4	Q13523	Serine/threonine-protein kinase PRP4 homolog (EC 2.7.1.37) (PRP4 pre- mRNA processing factor 4 homolog) (PRP4 kinase)
CRK7	Q9NYV4	Cell division cycle 2-related protein kinase 7 (EC 2.7.1.37) (CDC2-related protein kinase 7) (CrkRS).
Skip	Q13575	Nuclear protein SkiP (Ski-interacting protein) (SNW1 protein) (Nuclear receptor coactivator NCoA-62).
NOVA	P51513	RNA-binding protein Nova-1 (Neuro-oncological ventral antigen 1) (Onconeural ventral antigen-1) (Paraneoplastic Ri antigen) (Ventral neuron-specific protein 1)
	Q9UNW9	RNA-binding protein Nova-2 (Neuro-oncological ventral antigen 2) (Astrocytic NOVA1-like RNA-binding protein)
DEAD	P17844	Probable RNA-dependent helicase p68 (DEAD-box protein p68) (DEAD-box protein 5)
	Q92841	Probable RNA-dependent helicase p72 (DEAD-box protein p72) (DEAD-box protein 17)
	Q9H5Z1	Probable ATP-dependent helicase DHX35 (DEAH-box protein 35)
	Q96EI0	DEAD (Asp-Glu-Ala-Asp) box polypeptide 46
	O00571	DEAD-box protein 3 (Helicase-like protein 2) (HLP2) (DEAD-box, X isoform)
	O15523	DEAD-box protein 3, Y-chromosomal
	P38919	Probable ATP-dependent helicase DDX48 (DEAD-box protein 48) (Eukaryotic initiation factor 4A-like NUK-34) (Nuclear matrix protein 265) (hNMP 265) (Eukaryotic translation initiation factor 4A isoform 3)
	Q6IPS3	DDX26B protein
	Q9UL03	Candidate tumor suppressor protein DICE1 (DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 26) (OTTHUMP00000018439)
	O94894	KIAA0801 protein
	Q9UJV9	DEAD-box protein abstrakt homolog (DEAD-box protein 41)
	O43143	Putative pre-mRNA splicing factor RNA helicase (DEAH box protein 15) (ATP-dependent RNA helicase #46)
	Q13838	Spliceosome RNA helicase BAT1 (DEAD-box protein UAP56) (56 kDa U2AF65 associated protein) (ATP-dependent RNA helicase p47) (HLA-B associated transcript-1)
	O00148	ATP-dependent helicase DDX39 (DEAD-box protein 39) (Nuclear RNA helicase URH49)
	Q14562	ATP-dependent helicase DHX8 (RNA helicase HRH1) (DEAH-box protein 8)
	O60231	Putative pre-mRNA splicing factor RNA helicase (ATP-dependent RNA helicase #3) (DEAH-box protein 16)
	Q92620	Pre-mRNA splicing factor ATP-dependent RNA helicase PRP16 (EC 3.6.1.-) (ATP-dependent RNA helicase DHX38) (DEAH-box protein 38)
	Q08211	ATP-dependent RNA helicase A (Nuclear DNA helicase II) (NDH II) (DEAH-box protein 9)
	P42285	KIAA0052 protein
Cyclophilins	Q96BP3	Hypothetical protein KIAA0073 (EC 5.2.1.8) (Peptidyl-prolyl cis-trans isomerase) (PPIase) (Rotamase)
	Q9H2H8	Cyclophilin-like protein PPIL3b
	Q9Y3C6	Peptidyl-prolyl cis-trans isomerase like 1 (EC 5.2.1.8) (PPIase) (Rotamase) (CGI-124) (UNQ2425/PRO4984)
	Q13356	Peptidyl-prolyl cis-trans isomerase like 2 (EC 5.2.1.8) (PPIase) (Rotamase) (Cyclophilin-60) (Cyclophilin-like protein Cyp-60)
	Q9UNP9	Peptidyl-prolyl cis-trans isomerase E (EC 5.2.1.8) (PPIase E) (Rotamase E) (Cyclophilin E) (Cyclophilin 33)
HeatShock	P08107	Heat shock 70 kDa protein 1 (HSP70.1) (HSP70-1/HSP70-2)
	P11142	Heat shock cognate 71 kDa protein
	P11021	78 kDa glucose-regulated protein precursor (GRP 78) (Immunoglobulin heavy chain binding protein) (BiP) (Endoplasmic reticulum lumenal Ca(2+) binding protein grp78)
p52/p75	Q9UER6	Transcriptional coactivator p75
ELAV	P26378	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D)
	Q14576	ELAV-like protein 3 (Hu-antigen C) (HuC) (Paraneoplastic cerebellar degeneration-associated antigen) (Paraneoplastic limbic encephalitis antigen 21)
	Q12926	ELAV-like protein 2 (Hu-antigen B) (HuB) (ELAV-like neuronal protein 1) (Nervous system-specific RNA binding protein Hel-N1)
	Q15717	ELAV-like protein 1 (Hu-antigen R) (HuR)
P52/P100	P23246	Splicing factor, proline-and glutamine-rich (Polypyrimidine tract- binding protein-associated splicing factor) (PTB-associated splicing factor) (PSF) (DNA-binding P52/P100 complex, 100 kDa subunit)
	Q15233	54 kDa nuclear RNA- and DNA-binding protein (p54(nrb)) (p54nrb) (55 kDa nuclear protein) (NMT55) (Non-POU domain-containing octamer- binding protein) (DNA-binding P52/P100 complex, 52 kDa subunit)
FUSE	Q92945	KSRP
	Q92946	FUSE binding protein 3 (Fragment)
ColdShock	P16989	DNA-binding protein A (Cold shock domain protein A) (Single-strand DNA binding protein NF-GMB)
	P16991	Nuclease sensitive element binding protein 1 (Y box binding protein-1) (Y-box transcription factor) (YB-1) (CCAAT-binding transcription factor I subunit A) (CBF-A) (Enhancer factor I subunit A) (EFI-A) (DNA-binding protein B) (DBPB)

FBP	Q9NZA0	FBP-interacting repressor (Siah binding protein 1, FBP interacting repressor, pyrimidine tract binding splicing factor, Ro ribonucleoprotein-binding protein 1)
P32	Q07021	Complement component 1, Q subcomponent binding protein, mitochondrial precursor (Glycoprotein gC1qBP) (GC1q-R protein) (Hyaluronan-binding protein 1) (p32) (p33)
SNP70	Q9Y2W2	SH3 domain-binding protein SNP70 (NPW38-binding protein NPWBP) (Similar to WW domain binding protein 11)
CBP	P52298	Nuclear cap binding protein subunit 2 (20 kDa nuclear cap binding protein) (NCBP 20 kDa subunit) (CBP20) (NCBP interacting protein 1) (NIP1)
	Q09161	80 kDa nuclear cap binding protein (NCBP 80 kDa subunit) (CBP80)
ALY	O43672	THO complex subunit 4 (Tho4) (Ally of AML-1 and LEF-1) (Transcriptional coactivator Aly/REF) (bZIP enhancing factor BEF)
SLU7	O95391	Step II splicing factor SLU7
PRP18	Q99633	Pre-mRNA splicing factor 18 (PRP18 homolog).
CA150	O14776	Putative transcription factor CA150
RDP	P18615	Negative elongation factor E (NELF-E) (RD protein).
ACIN	Q9UKV3	Apoptotic chromatin condensation inducer in the nucleus (Acinus).
ILF3	Q12906	Interleukin enhancer-binding factor 3
CRN	Q9BZJ0	Crooked neck-like protein 1 (Crooked neck homolog)
WTAP	Q15007	Wilms' tumor 1-associating protein (WT1-associated protein) (Putative pre-mRNA splicing regulator female-lethal(2D) homolog)
PRP17	O60508	Pre-mRNA splicing factor PRP17 (hPRP17) (Cell division cycle 40 homolog) (EH-binding protein 3)
Others	Q9ULR0	KIAA1160 protein
	O75937	DNAJC8 protein
	P35637	RNA-binding protein FUS (Oncogene FUS) (Oncogene TLS) (Translocated in liposarcoma protein) (POMp75) (75 kDa DNA-pairing protein)
	Q92804	TATA-binding protein associated factor 2N (RNA-binding protein 56) (TAFII68) (TAF(II)68)
	Q8N2M8	Splicing factor, arginine/serine-rich 16 (Suppressor of white-apricot homolog 2)
	Q14498	RNA-binding region containing protein 2 (Hepatocellular carcinoma protein 1) (Splicing factor HCC1)
	O43934	ET putative translation product
	O43670	Zinc finger protein 207
	Q9Y5S9	RNA-binding protein 8A (RNA binding motif protein 8A) (Ribonucleoprotein RBM8A) (RNA-binding protein Y14) (Binder of OVCA1-1) (BOV-1)
	Q8IYB3	Ser/Arg-related nuclear matrix protein
	Q15637	Splicing factor 1 (Zinc finger protein 162) (Transcription factor ZFM1) (Zinc finger gene in MEN1 locus) (Mammalian branch point binding protein mBBP) (BBP)
	O15042	Hypothetical protein KIAA0332 (U2-associated SR140 protein)
	P52756	RNA-binding protein 5 (RNA binding motif protein 5) (Putative tumor suppressor LUCA15) (G15 protein)
	Q16630	HPBRII-4 mRNA (HPBRII-7 protein)
	Q96T58	Msx2-interacting protein (SMART/HDAC1 associated repressor protein)
	Q96T37	Putative RNA-binding protein 15 (RNA binding motif protein 15) (One-twenty two protein)
	Q9P2S7	Cisplatin resistance-associated overexpressed protein
	Q06787	Fragile X mental retardation 1 protein (Protein FMR-1) (FMRP)
	O00425	Putative RNA binding protein KOC (Koc1)
	P29558	Single-stranded DNA-binding protein MSSP-1 (RNA binding motif, single-stranded interacting protein 1)
	O15355	Protein phosphatase 2C gamma isoform (EC 3.1.3.16) (PP2C-gamma) (Protein phosphatase magnesium-dependent 1 gamma) (Protein phosphatase 1C)
	O95926	Hypothetical protein DKFZp564O2082 (GCIP-interacting protein p29)
	Q96I25	Splicing factor 45 (45kDa splicing factor) (RNA binding motif protein 17)
	Q14152	Eukaryotic translation initiation factor 3 subunit 10 (eIF-3 theta) (eIF3 p167) (eIF3 p180) (eIF3 p185) (eIF3a)
	Q9P013	HSPC148
	Q9BQ61	Hypothetical protein
	Q9BXP5	Arsenite-resistance protein 2
	Q9BRD0	Hypothetical protein
	P20042	Eukaryotic translation initiation factor 2 subunit 2 (Eukaryotic translation initiation factor 2 beta subunit) (eIF-2-beta)
	Q9NW64	Hypothetical protein FLJ10290
	O75940	Survival of motor neuron-related splicing factor 30 (SMN-related protein) (30 kDa splicing factor SMNrp) (Survival motor neuron domain containing protein 1)
	O75229	R31449_3
	O95400	CD2 antigen cytoplasmic tail-binding protein 2
	O43719	HIV TAT specific factor 1
	Q9HCE1	Potential helicase MOV-10 (EC 3.6.1.-) (Moloney leukemia virus 10 protein)
	Q9HCS7	XPA-binding protein 2 (HCNP protein) (PP3898)
	Q9H5H0	Hypothetical protein FLJ23445
	Q8WYA6	Beta-catenin-like protein 1 (Nuclear associated protein) (NAP) (NYD-SP19) (PP8304)
	Q8WWY3	U4/U6 snRNP-associated 61 kDa protein
	Q96DF8	DGCR14 protein (DiGeorge syndrome critical region 14) (ES2 protein)
	Q9Y6A4	Transcription factor IIB (EVORF)
	P43243	Matrin 3
	Q92973	Transportin 1 (Importin beta-)2 (Karyopherin beta-2) (M9 region interaction protein) (MIP)
	P61326	Mago nashi protein homolog
	O43684	Mitotic checkpoint protein BUB3
	P55081	Microfibrillar-associated protein 1
	Q8NI27	THO complex subunit 2 (Tho2)
	Q9Y5B6	GC-rich sequence DNA-binding factor homolog
	O60306	KIAA0560 protein
	Q13123	Red protein (RER protein) (IK factor) (Cytokine IK)
	Q12905	NF45 protein
	Q99974	Pombe Cdc5-related protein (CDC5 cell division cycle 5-like) (S.pombe) (CDC5-like)
	P41223	G10 protein homolog (EDG-2)
	P05455	Lupus La protein (Sjogren syndrome type B antigen) (SS-B) (La ribonucleoprotein) (La autoantigen)
	Q9HCG8	KIAA1604 protein
	Q969P6	DNA topoisomerase I, mitochondrial precursor (EC 5.99.1.2) (TOP1mt)
	O75934	Putative spliceosome associated protein (DAM1 protein) (Breast carcinoma amplified sequence 2)
	Q9UBB9	Tuftelin-interacting protein 11 (HSPC006)
	O43809	Pre-mRNA cleavage factor I 25 kDa subunit (Cleavage and polyadenylation specific factor 5, 25 kD subunit)
	Q9P2B8	KIAA1429 protein
	O43660	Pleiotropic regulator 1
	Q9UMS4	Nuclear matrix protein NMP200 (PRP19/PSO4 homolog)
	Q96J01	THO complex subunit 3 (Tho3)
	Q9BU59	Homolog of C. elegans smu-1
	Q96FV9	THO complex subunit 1 (Tho1) (Nuclear matrix protein p84)
	Q86W42	MGC2655 protein
	P49768	Presenilin 1 (PS-1) (S182 protein)
	P04720	Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1 A-1) (eEF1A-1) (Elongation factor Tu) (EF-Tu)
	P11940	Polyadenylate-binding protein 1 (Poly(A)-binding protein 1) (PABP 1)

A.1.2 Outgroups for phylogentic tree rooting

Table A.2 legend:

Prot_ID: local Locus ID for proteins within the family; SwissProt ID for proteins external to the family

Status: outgroup chosen within the family or externally; when there was an unambiguous outgroup protein within a family (e.g. when there was only one protozoan factor) that sequence was taken to root the trees; otherwise an external outgroup (designated as ExtRoot in the trees) was chosen

Table A.2: Outgroups for phylogentic tree rooting

Group	Family	Prot_ID	Status	Accession	Source	Species	Description
snRNP	CypH	PPIA_HUMAN	external	P62937	SwissProt	Homo sapiens	Peptidyl-prolyl cis-trans isomerase A (EC 5.2.1.8) (PPIase) (Rotamase) (Cyclophilin A) (Cyclosporin A-binding protein).
	FBP11	TCRG1_HUMAN	external	O14776	SwissProt	Homo sapiens	Transcription elongation regulator 1 (TATA box-binding protein- associated factor 2S) (Transcription factor CA150).
	p14	FUSIP_HUMAN	external	O75494	SwissProt	Homo sapiens	FUS interacting serine-arginine rich protein 1 (TLS-associated protein with Ser-Arg repeats) (TLS-associated protein with SR repeats) (TASR) (TLS-associated serine-arginine protein) (TLS-associated SR protein) (40 kDa SR-repressor protein) (SRrp40) (Splicing factor SRp38).
	PRP8	Tcru_PRP8	within family	TSKTSC_7739.t00032	TcruziDB	Trypanosoma cruzi	-
	S3A1	SF04_HUMAN	external	Q8IWZ8	SwissProt	Homo sapiens	Splicing factor 4 (RNA-binding protein RBP).
	S3A2	GP1_CHLRE	external	Q9FPQ6	SwissProt	Chlamydomonas reinhardtii	Vegetative cell wall protein gp1 precursor (Hydroxyproline-rich glycoprotein 1).
	S3A3	OPTN_CHICK	external	Q90Z16	SwissProt	Gallus gallus	Optineurin (Ag9-C5) (FIP-2).
	S3B1	Tcru_S3B1	within family	TSKTSC_7887.t00010	TcruziDB	Trypanosoma cruzi	-
	S3B2	Tcru_S3B2	within family	TSKTSC_4894.t00005	TcruziDB	Trypanosoma cruzi	-
	S3B3	Tcru_S3B3	within family	TSKTSC_4894.t00005	TcruziDB	Trypanosoma cruzi	-
	S3B4	Tcru_S3B4	within family	TSKTSC_8318.t00008	TcruziDB	Trypanosoma cruzi	-
	Tri-110	TRHY_HUMAN	external	Q07283	SwissProt	Homo sapiens	Trichohyalin.
	Tri-15	Tcru_NHP2l	within family	TSKTSC_8253.t00003	TcruziDB	Trypanosoma cruzi	-
	Tri-65	Scer_Tri65	within family	P43589	SwissProt	S. cerevisiae	Hypothetical 52.2 kDa protein in MPR1-GCN20 intergenic region
	U11U12-20	Dmel_UB20	within family	Q8IPW7	SwissProt	Drosophila melanogaster	CG31922-PA
	U11U12-25	Q5RJ89_HUMAN	external	Q5RJ89	SwissProt	Homo sapiens	Ubiquitin D.
	U11U12-31	SFRS7_HUMAN	external	Q16629	SwissProt	Homo sapiens	Splicing factor, arginine/serine-rich 7 (Splicing factor 9G8).
	U11U12-35	RU17_HUMAN	external	P08621	SwissProt	Homo sapiens	U1 small nuclear ribonucleoprotein 70 kDa (U1 snRNP 70 kDa) (snRNP70) (U1-70K).
	U11U12-48	Cint_UB48	within family	ci0100154801	JGI	Ciona intestinalis	-
	U11U12-65	Q96IZ5_HUMAN	external	Q96IZ5	SwissProt	Homo sapiens	Hypothetical protein FLJ11016.
	U1-70	Tbru_TSR1l	within family	CAB62267	NCBI	Trypanosoma brucei	splicing factor pTSR1 interacting protein
	U1AU2B	Tcru_U1A2B	within family	TSKTSC_7541.t00014	TcruziDB	Trypanosoma cruzi	-
	U1C	SF3A2_HUMAN	external	Q15428	SwissProt	Homo sapiens	Splicing factor 3A subunit 2 (Spliceosome associated protein 62) (SAP 62) (SF3a66).
	U2A	CJ011_HUMAN	external	Q9H2I8	SwissProt	Homo sapiens	Protein C10orf11 (CDA017).
	U4U6-60	Tcru_Prp4	within family	TSKTSC_5741.t00003	TcruziDB	Trypanosoma cruzi	-
	U4U6-90	NFH_HUMAN	external	P12036	SwissProt	Homo sapiens	Neurofilament triplet H protein (200 kDa neurofilament protein) (Neurofilament heavy polypeptide) (NF-H).
	U5-100	DDX17_HUMAN	external	Q92841	SwissProt	Homo sapiens	Probable RNA-dependent helicase p72 (DEAD-box protein p72) (DEAD-box protein 17).
	U5-102	CRNL1_HUMAN	external	Q9BZJ0	SwissProt	Homo sapiens	Crooked neck-like protein 1 (Crooked neck homolog) (hCrn) (CGI-201) (MSTP021).
	U5-116	Spom_U5116	within family	O94316	SwissProt	S. pombe	SPBC215.12 protein (Cwf10 protein) (Spef2 protein) (Snu114 protein)
	U5-15	TXN4B_HUMAN	external	Q9NX01	SwissProt	Homo sapiens	Thioredoxin-like protein 4B (Dim1-like protein).
	U5-200	HELC1_HUMAN	external	Q8N3C0	SwissProt	Homo sapiens	Activating signal cointegrator 1 complex subunit 3 (EC 3.6.1.-) (ASC-1 complex subunit p200) (Trip4 complex subunit p200) (Helicase, ATP binding 1).
	U5-40	Pfal_U540	within family	MAL8P1.43	PlasmoDB	Plasmodium falciparum	u5 snrnp-specific 40 kDa protein, putative
Sm	LSm1	LSM8_HUMAN	external	O95777	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm8.
	LSm10	Dmel_LSm10	within family	Q9V5Q2	SwissProt	Drosophila melanogaster	CG12938-PA
	LSm2	LSM4_HUMAN	external	Q9Y4Z0	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm4 (Glycine-rich protein) (GRP).
	LSm3	SMD2_HUMAN	external	P62316	SwissProt	Homo sapiens	Small nuclear ribonucleoprotein Sm D2 (snRNP core protein D2) (Sm-D2).
	LSm4	SMD3_HUMAN	external	P62318	SwissProt	Homo sapiens	Small nuclear ribonucleoprotein Sm D3 (snRNP core protein D3) (Sm-D3).
	LSm5	LSM3_HUMAN	external	P62310	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm3 (MDS017).
	LSm6	RUXF_HUMAN	external	P62306	SwissProt	Homo sapiens	Small nuclear ribonucleoprotein F (snRNP-F) (Sm protein F) (Sm-F) (SmF).
	LSm7	RUXG_HUMAN	external	P62308	SwissProt	Homo sapiens	Small nuclear ribonucleoprotein G (snRNP-G) (Sm protein G) (Sm-G) (SmG).
	LSm8	LSM1_HUMAN	external	O15116	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm1 (Small nuclear ribonuclear CaSm) (Cancer-associated Sm-like).
	SmBN	SF3B4_HUMAN	external	Q15427	SwissProt	Homo sapiens	Splicing factor 3B subunit 4 (Spliceosome associated protein 49) (SAP 49) (SF3b50) (Pre-mRNA splicing factor SF3b 49 kDa subunit).
	SmD1	SMD3_HUMAN	external	P62318	SwissProt	Homo sapiens	Small nuclear ribonucleoprotein Sm D3 (snRNP core protein D3) (Sm-D3).
	SmD2	LSM3_HUMAN	external	P62310	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm3 (MDS017).
	SmD3	LSM4_HUMAN	external	Q9Y4Z0	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm4 (Glycine-rich protein) (GRP).
	SmE	LSM5_HUMAN	external	Q9Y4Y9	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm5.
	SmF	LSM6_HUMAN	external	P62312	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm6 (Sm protein F).
	SmG	LSM7_HUMAN	external	Q9UK45	SwissProt	Homo sapiens	U6 snRNA-associated Sm-like protein LSm7.
	SmNew	Dmel_SmNew	within family	Q8IPZ7	SwissProt	Drosophila melanogaster	CG31950-PA
U2AF	U2AF35	U2AFL_HUMAN	external	Q15695	SwissProt	Homo sapiens	U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit related-protein 1 (U2(RNU2) small nuclear RNA auxillary factor 1-like 1).
	U2AF35R	Atha_U2R	within family	NP_172503	NCBI	Arabidopsis thaliana	U2 snRNP auxiliary factor-related
	U2AF65	Pfal_U2AF	within family	PF14_0656	PlasmoDB	Plasmodium falciparum	U2 snRNP auxiliary factor, putative
SR	9G8-SRp20	SFRS5_HUMAN	external	Q13243	SwissProt	Homo sapiens	Splicing factor, arginine/serine-rich 5 (Pre-mRNA splicing factor SRP40) (Delayed-early protein HRS).
	p54	Cele_p54	within family	O01159	SwissProt	C. elegans	Probable splicing factor, arginine/serine-rich 7 (p54)
	RY1	Atha_RY1	within family	NP_568856	NCBI	Arabidopsis thaliana	expressed protein
	SC35	SFRS7_HUMAN	external	Q16629	SwissProt	Homo sapiens	Splicing factor, arginine/serine-rich 7 (Splicing factor 9G8).
	SRm300	Atha_SR45	within family	NP_173107	NCBI	Arabidopsis thaliana	arginine/serine-rich protein, putative (SR45)
	SRp30c-ASF	Pfal_SF	within family	PFE0865c	PlasmoDB	Plasmodium falciparum	splicing factor, putative
	SRp40-55-75	SFRS9_HUMAN	external	Q13242	SwissProt	Homo sapiens	Splicing factor, arginine/serine-rich 9 (Pre-mRNA splicing factor SRp30C).
	TopoI-B	SFRS4_HUMAN	external	Q08170	SwissProt	Homo sapiens	Splicing factor, arginine/serine-rich 4 (Pre-mRNA splicing factor SRP75) (SRP001LB).
	Tra2	Cele_Tra2	within family	Q9XTZ2	SwissProt	C. elegans	Hypothetical protein rsp-8
hnRNP	hnRNP-A	ROAA_HUMAN	external	Q99729	SwissProt	Homo sapiens	Heterogeneous nuclear ribonucleoprotein A/B (hnRNP A/B) (APOBEC-1 binding protein 1) (ABBP-1).
	hnRNP-C	Cint_ROC	within family	ci0100140076	JGI	Ciona intestinalis	RNA-binding region RNP-1 (RNA recognition motif)
	hnRNP-D-U2	ROA3_HUMAN	external	P51991	SwissProt	Homo sapiens	Heterogeneous nuclear ribonucleoprotein A3 (hnRNP A3).
	hnRNP-E	Cele_PCB	within family	Q95Y67	SwissProt	C. elegans	Patterned expression site protein 4
	hnRNP-F-H	HNRPD_HUMAN	external	Q14103	SwissProt	Homo sapiens	Heterogeneous nuclear ribonucleoprotein D0
	hnRNP-G	RBY1A_HUMAN	external	Q15414	SwissProt	Homo sapiens	RNA binding motif protein, Y chromosome, family 1 member A1 (RNA- binding motif protein 1).
	hnRNP-I	ELAV2_HUMAN	external	Q12926	SwissProt	Homo sapiens	ELAV-like protein 2
	hnRNP-K	Cele_ROK	within family	P91277	SwissProt	C. elegans	Hypothetical protein F26B1.2
	hnRNP-L	Cele_ROL	within family	Q95QR5	SwissProt	C. elegans	Hypothetical protein C44B7.2
	hnRNP-M	Tcru_ROM	within family	TSKTSC_8485.t00012	TcruziDB	Trypanosoma cruzi	-
	hnRNP-R	DND1_HUMAN	external	Q8IYX4	SwissProt	Homo sapiens	Dead end protein homolog 1 (RNA binding motif, single-stranded interacting protein 4).
	hnRNP-U	Cele_ROU	within family	Q9U2H8	SwissProt	C. elegans	Hypothetical protein Y41E3.11
	Musashi	Spom_MUS	within family	O94432	SwissProt	S. pombe	SPBC660.15 protein
DEAD	ABS	Pfal_ABS	within family	PFE1390w	PlasmoDB	Plasmodium falciparum	RNA helicase-1
	DDX26	Cele_DDX26	within family	F08B4.1b	Ensembl	C. elegans	DEAD H box polypeptide 26 (4J459)
	DDX39	DDX6_HUMAN	external	P26196	SwissProt	Homo sapiens	Probable ATP-dependent RNA helicase p54 (Oncogene RCK) (DEAD-box protein 6).
	DDX3XY	DDX4_HUMAN	external	Q9NQI0	SwissProt	Homo sapiens	DEAD-box protein 4 (VASA homolog).
	DDX46	DDX17_HUMAN	external	Q92841	SwissProt	Homo sapiens	Probable RNA-dependent helicase p72 (DEAD-box protein p72) (DEAD-box protein 17).
	DDX48	IF41_HUMAN	external	P60842	SwissProt	Homo sapiens	Eukaryotic initiation factor 4A-I (eIF4A-I) (eIF-4A-I).
	DHX15	DHX8_HUMAN	external	Q14562	SwissProt	Homo sapiens	ATP-dependent helicase DHX8 (RNA helicase HRH1) (DEAH-box protein 8).
	DHX16	Spom_CDC28	within family	Q10752	SwissProt	S.pombe	Putative ATP-dependent RNA helicase cdc28
	DHX35	DHX16_HUMAN	external	O60231	SwissProt	Homo sapiens	Putative pre-mRNA splicing factor RNA helicase (ATP-dependent RNA helicase #3) (DEAH-box protein 16).
	DHX38	DHX8_HUMAN	external	Q14562	SwissProt	Homo sapiens	ATP-dependent helicase DHX8 (RNA helicase HRH1) (DEAH-box protein 8).
	DHX8	DHX16_HUMAN	external	O60231	SwissProt	Homo sapiens	Putative pre-mRNA splicing factor RNA helicase (ATP-dependent RNA helicase #3) (DEAH-box protein 16).
	DHX9	Cele_DHX9	within family	Q22307	SwissProt	C. elegans	Probable ATP-dependent RNA helicase A (Nuclear DNA helicase II) (NDH II)
	KIAA0052	SKIV2_HUMAN	external	Q15477	SwissProt	Homo sapiens	Helicase SKI2W (Helicase-like protein) (HLP).
	p68p72	Q96EI0_HUMAN	external	Q96EI0	SwissProt	Homo sapiens	DEAD (Asp-Glu-Ala-Asp) box polypeptide 46.
Others	CLK	DYRK4_HUMAN	external	Q9NR20	SwissProt	Homo sapiens	Dual-specificity tyrosine-phosphorylation regulated kinase 4 (EC 2.7.1.-).
	CUG	RBM28_HUMAN	external	Q9NW13	SwissProt	Homo sapiens	RNA-binding protein 28 (RNA binding motif protein 28).
	ELAV	Cele_ELAV	within family	F35H8.5	Ensembl	C. elegans	Transcribed locus [Source:;Acc:Cel.23024]
	FUSE	CO1A2_HUMAN	external	P08123	SwissProt	Homo sapiens	Collagen alpha 2(I) chain precursor.
	NOA	PCBP2_HUMAN	external	Q15366	SwissProt	Homo sapiens	Poly(rC)-binding protein 2 (Alpha-CP2) (hnRNP-E2).
	PRP4	DYRK2_HUMAN	external	Q92630	SwissProt	Homo sapiens	Dual-specificity tyrosine-phosphorylation regulated kinase 2 (EC 2.7.1.112) (EC 2.7.1.37).
	SKIP	INCE_HUMAN	external	Q9NQS7	SwissProt	Homo sapiens	Inner centromere protein.
	SRPK	DYRK2_HUMAN	external	Q92630	SwissProt	Homo sapiens	Dual-specificity tyrosine-phosphorylation regulated kinase 2 (EC 2.7.1.112) (EC 2.7.1.37).
	TIA	PABP3_HUMAN	external	Q9H361	SwissProt	Homo sapiens	Polyadenylate-binding protein 3 (Poly(A)-binding protein 3) (PABP 3) (Testis-specific poly(A)-binding protein).

A.1.3 Molecular clock test

Method:

Given the alignments, the program GAMMA [Gu and Zhang, 1997] was used to calculate the gamma-corrected substitution rate. Rooted and corrected trees were then rebuilt using again the Phylip programs Protdist, Neighbor and Consense for NJ and Proml for ML. The LinearTree [Takezaki et al., 1995] program TPCV (5% significance) was used to apply the two-cluster test of rate constancy and linearized trees were drawn for significant families. This procedure relies on ungapped alignments and for 8 of the 97 families there were too many gaps to perform the test successfully. 61 of the 89 analysed families satisfied the molecular clock hypothesis.

Table A.3 legend:

N: number of sequences

${\chi^2}_{N-2}$: chi-square test with n-1 degrees of freedom (n - the number of sequences under the root $\Rightarrow$ n=N-1)

p: significance (p-value) for ${\chi^2}_{N-2}$ test

Table A.3: Molecular clock test

Group	Family	N	χ^2_{N-2}	p	Test
snRNP	CypH	15	20,150	9,153E-02	Yes
	FBP11	19	11,661	8,202E-01	Yes
	p14	17	9,583	8,451E-01	Yes
	PRP8	17	25,090	4,875E-02	No
	S3A1	18	8,042	9,476E-01	Yes
	S3A2	17	9,131	8,705E-01	Yes
	S3A3	19	9,948	9,058E-01	Yes
	S3B1	17	51,542	6,723E-06	No
	S3B2	15	18,688	1,331E-01	Yes
	S3B3	17	51,212	7,619E-06	No
	S3B4	16	31,517	4,689E-03	No
	Tri-110	15	6,941	9,052E-01	Yes
	Tri-15	19	12,693	7,565E-01	Yes
	Tri-65	11	10,017	3,491E-01	Yes
	U11/U12-20	9	11,666	1,121E-01	Yes
	U11/U12-25	10	6,831	5,550E-01	Yes
	U11/U12-31	11	18,196	3,297E-02	No
	U11/U12-35	10	15,071	5,777E-02	Yes
	U11/U12-48	8	2,004	9,194E-01	Yes
	U11/U12-65	11	4,942	8,393E-01	Yes
	U1-70	16	2,193	9,999E-01	Yes
	U1AU2B	30	22,091	7,771E-01	Yes
	U1C	16	8,815	8,427E-01	Yes
	U2A	19	18,167	3,784E-01	Yes
	U4U6-60	17	12,084	6,727E-01	Yes
	U4U6-90	18	16,163	4,416E-01	Yes
	U5-100	14	26,183	1,011E-02	No
	U5-102	16	24,183	4,355E-02	No
	U5-116	13	35,300	2,211E-04	No
	U5-15	15	47,717	7,308E-06	No
	U5-200	20	39,869	2,175E-03	No
	U5-40	15	10,402	6,608E-01	Yes
Sm	LSm1	15	6,432	9,290E-01	Yes
	LSm10	9	4,071	7,716E-01	Yes
	LSm2	14	21,072	4,934E-02	No
	LSm3	16	11,466	6,491E-01	Yes
	LSm4	21	18,975	4,584E-01	Yes
	LSm5	15	9,509	7,335E-01	Yes
	LSm6	17	19,504	1,918E-01	Yes
	LSm7	15	3,435	9,959E-01	Yes
	LSm8	16	12,972	5,287E-01	Yes
	SmBN	24	14,718	8,741E-01	Yes
	SmD1	18	9,729	8,804E-01	Yes
	SmD2	18	31,649	1,111E-02	No
	SmD3	19	17,525	4,194E-01	Yes
	SmE	18	13,612	6,276E-01	Yes
	SmF	18	8,204	9,425E-01	Yes
	SmG	17	9,189	8,674E-01	Yes
	SmNew	8	5,841	4,412E-01	Yes

Group	Family	N	χ^2_{N-2}	p	Test
U2AF	U2AF35	33	46,776	3,435E-02	No
	U2AF35R	12	15,770	1,064E-01	Yes
	U2AF65	18	21,653	1,547E-01	Yes
SR	9G8-SRp20	37	228,793	2,608E-30	No
	p54	21	26,392	1,196E-01	Yes
	RY1	11	7,373	5,984E-01	Yes
	SC35	22	13,260	8,660E-01	Yes
	SRm300	11	9,590	3,847E-01	Yes
	SRp30c-ASF	25			
	SRp40-55-75	41			
	TopoI-B	13	0,990	1,000E+00	Yes
	Tra2	20	17,929	4,603E-01	Yes
hnRNP	hnRNP-A	37	28,319	7,808E-01	Yes
	hnRNP-C	28			
	hnRNP-D-U2	33	22,897	8,528E-01	Yes
	hnRNP-E	30	81,861	3,515E-07	No
	hnRNP-F-H	39	1349806,123	0,000E+00	No
	hnRNP-G	14	47,140	4,409E-06	No
	hnRNP-I	41			
	hnRNP-K	15	35,958	6,024E-04	No
	hnRNP-L	17			
	hnRNP-M	19	13,983	6,683E-01	Yes
	hnRNP-R	24	10,570	9,804E-01	Yes
	hnRNP-U	27	40,448	2,622E-02	No
	Musashi	20	5,252	9,984E-01	Yes
DEAD	ABS	14	24,658	1,653E-02	No
	DDX26	17	19,812	1,793E-01	Yes
	DDX39	27	768,103	6,397E-146	No
	DDX3XY	27	20,387	7,263E-01	Yes
	DDX46	19	19,470	3,022E-01	Yes
	DDX48	18	58,346	9,917E-07	No
	DHX15	20	108,161	6,917E-15	No
	DHX16	15	22,915	4,271E-02	No
	DHX35	13	17,873	8,458E-02	Yes
	DHX38	17	50,301	1,075E-05	No
	DHX8	20	32,577	1,877E-02	No
	DHX9	12	19,757	3,163E-02	No
	KIAA0052	20			
	p68p72	28	28,018	3,576E-01	Yes
Others	CLK	37	28,967	7,537E-01	Yes
	CUG	56			
	ELAV	40	39,924	3,846E-01	Yes
	FUSE	34	18,645	9,710E-01	Yes
	NOA	19	17,233	4,387E-01	Yes
	PRP4	20	33,324	1,525E-02	No
	SKIP	15	11,267	5,885E-01	Yes
	SRPK	45			
	TIA	37	21,668	9,621E-01	Yes

A.1.4 Database sources for genomic and proteomic sequences

Table A.4: Database sources for genomic and proteomic sequences

Database	Species	Version
Ensembl [Hubbard et al., 2002] http://www.ensembl.org	*Homo sapiens* [Lander et al., 2001; Venter et al., 2001] (genome + proteome)	NCBI35/v30 (369 supercontigs, 33869 peptides, 3272.2 Mb)
	Fugu rubripes [Aparicio et al., 2002] (genome + proteome)	V2.0/v30 (20379 scaffolds, 33003 peptides, 329.1 Mb)
Joint Genome Institute http://www.jgi.doe.gov	*Ciona intestinalis* [Dehal et al., 2002] (genome + proteome)	Release 1.0 (2510 scaffolds, 15852 peptides, 119.1 Mb)
Sanger Institute http://www.sanger.ac.uk/ /Projects/S_pombe	*Schizosaccharomyces pombe* [Wood et al., 2002] (proteome)	V42 (4994 peptides linked to SwissProt [Boeckmann et al., 2003])
Saccharomyces Genome Database http://www.yeastgenome.org	*Saccharomyces cerevisiae* (proteome)	V42 (9747 peptides linked to SwissProt [Boeckmann et al., 2003])
PlasmoDB http://plasmodb.org	*Plasmodium falciparum* (genome + proteome)	V4.3 (19 scaffolds, 5334 annot. peptides, 23.2 Mb)
Sanger Institute http://www.sanger.ac.uk/ /Projects/T_brucei	*Trypanosoma brucei* (genome + proteome)	Jan. 2004 (5 contigs, 4559 proteins, 4.4 Mb, unfinished)
TcruziDB http://tcruzidb.org	*Trypanosoma cruzi* (genome + proteome)	V3.0 (Jul. 2004) (4014 scaffolds, 22273 proteins, 60 Mb, unfinished)
NCBI http://www.ncbi.nlm.nih.gov	*Arabidopsis thaliana* (proteome)	v.5.0
	16 species of Archaea: *Aeropyrum pernix* *Archaeoglobus fulgidus DSM* *Halobacterium sp. NRC-1* *Methanocaldococcus jannaschii* *Methanopyrus kandleri AV19* *Methanosarcina acetivorans C2A* *Methanosarcina mazei Goe1* *Methanothermobacter thermautotrophicus str. Delta H* *Pyrobaculum aerophilum str. IM2* *Pyrococcus abyssi* *Pyrococcus furiosus DSM 3638* *Pyrococcus horikoshii* *Sulfolobus solfataricus* *Sulfolobus tokodaii* *Thermoplasma acidophilum* *Thermoplasma volcanium*	Latest versions (from 1997 to 2002)

A.1.5 Putative pseudo-genes annotated as active genes in Ensembl

Table A.5 legend:

Disruption: appearance of frame disruption events (cryptic stop codons; frameshifts introduced by missing or extra nucleotides in the conserved coding region)

Ref. S1: closest active paralogue used for comparison

Ensembl dS/dN S1: rate of synonymous / non-synonymous substitutions provided by Ensembl for the comparison with the closest active paralogue (Ref. S1)

ds/dn S1: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the closest active paralogue (Ref. S1)

Ref. S2: active orthologue in the alternative species (Human/Mouse) used for comparison

ds/dn S2: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the active orthologue (Ref. S2)

ds/dn R1-R2: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison between the two active orthologues (Ref. S1 and Ref. S2)

Table A.5: Putative pseudo-genes annotated as active genes in Ensembl

Group	Family	Species	Accession	Source	Description	Disruption	Ref. S1	Ensembl dS/dN S1	ds/dn S1	Ref. S2	ds/dn S2	ds/dn R1-R2
snRNP	CypH	Human	ENSP00000330894	Ensembl	[PEPTIDYL PROLYL CIS TRANS ISOMERASE EC_5.2.1.8 PPIASE ROTAMASE CYCLOPHILIN]	Y	Hsap_CypH		2,25	Mmus_CypH	5,59	Infinite
	S3A2	Mouse	ENSMUSP00000054123	Ensembl	[SPLICING FACTOR 3A SUBUNIT 2 SPLICEOSOME ASSOCIATED 62 SAP 62 S]	Y	Mmus_S3A2		2,28	Hsap_S3A2	7,90	51,46
	S3A3	Human	ENSP00000328650	Ensembl	[SPLICING FACTOR 3A SUBUNIT 3 SPLICEOSOME ASSOCIATED 61 SAP 61 S]	Y	Hsap_S3A3		3,55	Mmus_S3A3	16,38	174,02
	Tri-15	Mouse	ENSMUSP00000078771	Ensembl	[NHP2 1 HIGH MOBILITY GROUP NUCLEAR 2 HOMOLOG 1 [U4/U6 U5] TRI SNRNP 15.5 KDA]		Mmus_NHP2l	0,00	0,00	Hsap_NHP2l	91,72	Infinite
	U1AU2B	Mouse	ENSMUSP00000074725	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN A U1 SNRNP A U1 A]	Y	Mmus_U1A		1,63	Hsap_U1A	15,39	27,72
		Mouse	ENSMUSP00000079363	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN A U1 SNRNP A U1 A]	Y	Mmus_U1A		2,90	Hsap_U1A	16,21	27,72
		Mouse	ENSMUSP00000074375	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN A U1 SNRNP A U1 A]	Y	Mmus_U1A		1,40	Hsap_U1A	5,64	27,72
		Mouse	ENSMUSP00000059367	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN A U1 SNRNP A U1 A]	Y	Mmus_U1A		1,67	Hsap_U1A	29,53	27,72
	U1C	Human	ENSP00000305492	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN C U1 SNRNP C U1C U1 C]		Hsap_U1C	2,27	2,37	Mmus_U1C	15,85	253,28
		Mouse	ENSMUSP00000074551	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN C U1 SNRNP C U1C U1 C]		Mmus_U1C	2,22	2,32	Hsap_U1C	6,19	253,28
	U4U6-90	Human	ENSP00000315379	Ensembl	[U4/U6 SMALL NUCLEAR RIBONUCLEOPROTEIN PRP3 PRE SPLICING FACTOR 3]	Y	Hsap_Prp3	1,47	1,86	Mmus_Prp3	6,18	22,11
		Mouse	ENSMUSP00000038836	Ensembl	[U4/U6 SMALL NUCLEAR RIBONUCLEOPROTEIN PRP3 PRE SPLICING FACTOR 3]	Y	Mmus_Prp3		1,76	Hsap_Prp3	3,51	22,11
	U5-15	Human	ENSP00000272762	Ensembl	[THIOREDOXIN 4 THIOREDOXIN U5 SNRNP U5 15KD SPLICEOSOMAL U5 SNRNP SPECIFIC 15 KDA DIM1 HOMOLOG]		Hsap_U5-15	4,31	4,47	Mmus_U5-15	12,02	280,75
Sm	LSm5	Mouse	ENSMUSP00000064031	Ensembl	[U6 SNRNA ASSOCIATED SM LSM5]	Y	Mmus_LSm5		1,68	Hsap_LSm5	60,05	Infinite
	LSm6	Mouse	ENSMUSP00000073010	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN F SNRNP F SM F SM F SMF]		Mmus_LSm6	0,78	0,83	Hsap_LSm6	22,78	Infinite
	LSm7	Mouse	ENSMUSP00000072521	Ensembl	[U6 SNRNA ASSOCIATED SM LSM7]		Mmus_LSm7	3,41	3,60	Hsap_LSm7	27,81	46,24
	SmBN	Mouse	ENSMUSP00000056410	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN ASSOCIATED SNRNP SM SM]	Y	Mmus_SmN	1,18	1,09	Hsap_SmN	9,13	Infinite
	SmD2	Human	ENSP00000277860	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN SM D2 SNRNP CORE D2 SM D2]		Hsap_SmD2	4,10	4,33	Mmus_SmD2	9,76	Infinite
		Mouse	ENSMUSP00000079907	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN SM D2 SNRNP CORE D2 SM D2]		Mmus_SmD2	5,61	5,92	Hsap_SmD2	23,37	Infinite
	SmE	Human	ENSP00000327930	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN E SNRNP E SM E SM E SME]		Hsap_SmE		4,01	Mmus_SmE	32,50	Infinite
		Human	ENSP00000329276	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN E SNRNP E SM E SM E SME]		Hsap_SmE		3,82	Mmus_SmE	40,27	Infinite
	SmF	Mouse	ENSMUSP00000071222	Ensembl	SMALL NUCLEAR RIBONUCLEOPROTEIN F (SNRNP-F) (SM PROTEIN F) (SM-F) (SMF)		Mmus_SmF	1,60	1,68	Hasp_SmF	61,67	Infinite
		Mouse	ENSMUSP00000077958	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN F SNRNP F SM F SM F SMF]		Mmus_SmF	3,55	2,72	Hasp_SmF	11,03	Infinite
	SmG	Mouse	ENSMUSP00000058234	Ensembl	[UNKNOWN]	Y	Mmus_SmG		4,44	Hasp_SmG	7,97	Infinite
U2AF	U2AF35	Human	ENSP00000309888	Ensembl	[SPLICING FACTOR U2AF 35 KDA SUBUNIT U2 AUXILIARY FACTOR 35 KDA SUBUNIT U2 SNRNP AUXILIARY FACTOR SMALL SUBUNIT]		Hsap_U2AG		2,30	Mmus_U2AG	16,44	Infinite
	U2AF35R	Human	ENSP00000342084	Ensembl	Similar to U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit related-protein 2	Y	Mmus_U2R2		3,64	Hsap_U2R2	8,02	6,63
	U2AF65	Mouse	ENSMUSP00000061115	Ensembl	Splicing factor U2AF 65 kDa subunit (U2 auxiliary factor 65 kDa subunit) (U2 snRNP auxiliary factor large subunit)		Mmus_U2AF		3,51	Hsap_U2AF	10,80	181,72
SR	9G8-SRp20	Mouse	ENSMUSP00000056269	Ensembl	SPLICING FACTOR, ARGININE/SERINE-RICH 3 (PRE-MRNA SPLICING FACTOR SRP20) (X16 PROTEIN)		Mmus_SR20		1,54	Hsap_SR20	86,14	Infinite
		Mouse	ENSMUSP00000040235	Ensembl	[SPLICING FACTOR ARGININE/SERINE RICH SPLICING FACTOR]		Mmus_SR20		0,88	Hsap_SR20	10,85	Infinite
	SRp40-55-75	Mouse	ENSMUSP00000055868	Ensembl	[SPLICING FACTOR ARGININE/SERINE RICH PRE SPLICING FACTOR]		Mmus_SR40	3,06	3,24	Hsap_SR40	30,34	69,51
	Tra2	Mouse	ENSMUSP00000060263	Ensembl	[ARGININE/SERINE RICH SPLICING FACTOR 10 TRANSFORMER 2 BETA HTRA2 BETA TRANSFORMER 2 HOMOLOG]		Mmus_Tra2B	1,83	2,40	Hsap_Tra2B	31,85	Infinite
		Mouse	ENSMUSP00000073554	Ensembl	[ARGININE/SERINE RICH SPLICING FACTOR 10 TRANSFORMER 2 BETA HTRA2 BETA TRANSFORMER 2 HOMOLOG]	Y	Mmus_Tra2B	3,14	3,80	Hsap_Tra2B	14,21	Infinite
hnRNP	hnRNP-A	Human	ENSP00000259575	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA1		2,05	Mmus_ROA1	8,44	332,43
		Human	ENSP00000222956	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]		Hsap_ROA1		2,12	Mmus_ROA1	12,51	332,43
		Human	ENSP00000312595	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA1		1,78	Mmus_ROA1	4,26	332,43
		Human	ENSP00000322209	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA1		2,49	Mmus_ROA1	6,31	332,43
		Human	ENSP00000324128	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA1		3,14	Mmus_ROA1	11,77	332,43
		Human	ENSP00000324555	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA1		2,30	Mmus_ROA1	6,16	332,43
		Human	ENSP00000341227	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Hsap_ROA3		1,34	Mmus_ROA3	4,81	Infinite
		Human	ENSP00000346926	Ensembl	PREDICTED: similar to Hnrpa1 protein		Hsap_ROA1	2,19	2,01	Mmus_ROA1	8,64	332,43
		Human	ENSP00000353125	Ensembl	heterogeneous nuclear ribonucleoprotein A1 (HNRPA1), transcript variant 1, mRNA	Y	Hsap_ROA1		2,67	Mmus_ROA1	17,09	332,43
		Human	ENSP00000341285	Ensembl	heterogeneous nuclear ribonucleoprotein A1-like (LOC144983), transcript variant 2, mRNA		Hsap_ROA1		2,16	Mmus_ROA1	28,36	332,43
		Human	ENSP00000350090	Ensembl	heterogeneous nuclear ribonucleoprotein A1 isoform a		Hsap_ROA1	3,49	4,82	Mmus_ROA1	80,35	332,43
		Human	ENSP00000339841	Ensembl	29 kDa protein	Y	Hsap_ROA3		1,24	Mmus_ROA3	6,30	Infinite
		Human	ENSP00000328902	Ensembl	31 kDa protein	Y	Hsap_ROA3		1,39	Mmus_ROA3	6,19	Infinite
		Mouse	ENSMUSP00000072775	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		7,46	Hsap_ROA3	83,44	Infinite
		Mouse	ENSMUSP00000079142	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA1		-	Hsap_ROA1	278,28	332,43
		Mouse	ENSMUSP00000080129	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y			Too short			
		Mouse	ENSMUSP00000076498	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,37	Hsap_ROA3	7,73	Infinite
		Mouse	ENSMUSP00000077780	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		3,69	Hsap_ROA3	23,35	Infinite
		Mouse	ENSMUSP00000080245	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,16	Hsap_ROA3	6,28	Infinite
		Mouse	ENSMUSP00000053413	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA1		2,10	Hsap_ROA1	37,90	332,43
		Mouse	ENSMUSP00000075081	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		2,68	Hsap_ROA3	6,97	Infinite
		Mouse	ENSMUSP00000073549	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,16	Hsap_ROA3	5,45	Infinite
		Mouse	ENSMUSP00000044485	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		2,51	Hsap_ROA3	5,02	Infinite
		Mouse	ENSMUSP00000071738	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		3,64	Hsap_ROA3	16,80	Infinite
		Mouse	ENSMUSP00000078666	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]		Mmus_ROA3		2,47	Hsap_ROA3	28,13	Infinite
		Mouse	ENSMUSP00000058664	Ensembl	Heterogeneous nuclear ribonucleoprotein A1 (Helix-destabilizing protein) (Single-strand binding protein) (hnRNP core protein A1) (HDP- 1) (Topoisomerase-inhibitor suppressed)		Mmus_ROA1		1,76	Hsap_ROA1	56,96	332,43
		Mouse	ENSMUSP00000054675	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		2,01	Hsap_ROA3	6,62	Infinite
		Mouse	ENSMUSP00000079581	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		2,44	Hsap_ROA3	30,49	Infinite
		Mouse	ENSMUSP00000056607	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,20	Hsap_ROA3	41,97	Infinite
		Mouse	ENSMUSP00000080126	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]		Mmus_ROA3		1,92	Hsap_ROA3	16,93	Infinite
		Mouse	ENSMUSP00000072189	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]		Mmus_ROA3	1,63	1,72	Hsap_ROA3	115,41	Infinite
		Mouse	ENSMUSP00000071646	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]		Mmus_ROA3		2,77	Hsap_ROA3	41,61	Infinite
		Mouse	ENSMUSP00000022493	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA1		4,46	Hsap_ROA1	23,81	332,43
		Mouse	ENSMUSP00000045802	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA1		2,62	Hsap_ROA1	23,52	332,43
		Mouse	ENSMUSP00000030569	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,54	Hsap_ROA3	3,70	Infinite
		Mouse	ENSMUSP00000038713	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA1		2,54	Hsap_ROA1	8,44	332,43
		Mouse	ENSMUSP00000071404	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,85	Hsap_ROA3	6,97	Infinite
		Mouse	ENSMUSP00000077190	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,88	Hsap_ROA3	3,96	Infinite
		Mouse	ENSMUSP00000076239	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN]	Y	Mmus_ROA3		1,34	Hsap_ROA3	5,26	Infinite
	hnRNP-C	Human	ENSP00000345190	Ensembl	[HNRNP]	Y	Hsap_ROC		1,59	Mmus_ROC	6,28	39,28
		Mouse	ENSMUSP00000060431	Ensembl	[HNRNP]	Y	Mmus_ROC		3,62	Hsap_ROC	1,36	39,28
		Mouse	ENSMUSP00000071969	Ensembl	[HNRNP]	Y	Mmus_ROC		2,43	Hsap_ROC	1,39	39,28
	hnRNP-F-H	Mouse	ENSMUSP00000051560	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN HNRNP]	Y	Mmus_ROH2	2,00	2,13	Hsap_ROH2	26,96	92,18
	hnRNP-G	Human	ENSP00000341015	Ensembl	30 kDa protein	Y	Hsap_ROG		1,29	Mmus_ROG	3,10	55,39
	hnRNP-I	Human	ENSP00000339100	Ensembl	58 kDa protein	Y	Hsap_PTB1	2,30	2,28	Mmus_PTB1	5,19	16,68
		Mouse	ENSMUSP00000071250	Ensembl	[POLYPYRIMIDINE TRACT BINDING 1 PTB HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN I HNRNP I]		Mmus_PTB1		3,05	Hsap_PTB1	3,63	16,68
	hnRNP-K	Human	ENSP00000333256	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN K]	Y	Hsap_ROK		1,27	Mmus_ROK	9,51	Infinite
		Human	ENSP00000341659	Ensembl	[HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN K]	Y	Hsap_ROK		1,26	Mmus_ROK	3,92	Infinite
	hnRNP-M	Mouse	ENSMUSP00000058811	Ensembl	HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN M HNRNP M		Mmus_Myel		1,00	Hsap_Myel	8,79	8,85
Others	CLK	Human	ENSP00000335433	Ensembl	45 kDa protein	Y	Hsap_CLK3	1,32	1,40	Mmus_CLK3	2,64	71,48
		Human	ENSP00000328452	Ensembl	7 kDa protein	Y			Too short			

A.1.6 Putative novel active retrotransposed genes

Table A.6 legend:

Trans. Act.: EST evidence for transcriptional activity

Ref. S1: closest active paralogue used for comparison

Ensembl dS/dN S1: rate of synonymous / non-synonymous substitutions provided by Ensembl for the comparison with the closest active paralogue (Ref. S1)

ds/dn S1: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the closest active paralogue (Ref. S1)

Ref. S2: active orthologue in the alternative species (Human/Mouse) used for comparison

ds/dn S2: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the active orthologue (Ref. S2)

ds/dn R1-R2: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison between the two active orthologues (Ref. S1 and Ref. S2)

*Factors exhibiting the same transcript sequences as their closest active paralogue (Ref. S1)

Table A.6: Putative novel active retrotransposed genes

Group	Family	Species	Accession	Source	Description	Trans. Act.	Ref. S1	Ensembl dS/dN S1	ds/dn S1	Ref. S2	ds/dn S2	ds/dn R1-R2
snRNP	Tri-15	Mouse	ENSMUSP00000076811	Ensembl	NHP2-like protein 1 (High mobility group-like nuclear protein 2 homolog 1) ([U4/U6.U5] tri-snRNP 15.5 kDa protein) (Sperm specific antigen 1) (Fertilization antigen 1) (FA-1)	Y	Mmus_NHP2l	Infinite	Infinite	Hsap_NHP2l	Infinite	Infinite
	U1C	Mouse	ENSMUSP00000076754	Ensembl	[U1 SMALL NUCLEAR RIBONUCLEOPROTEIN C U1 SNRNP C U1C U1 C]		Mmus_U1C	24,56	25,18	Hsap_U1C	99,69	253,28
Sm	LSm6	Mouse	ENSMUSP00000075036	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN F SNRNP F SM F SM F SMF]	Y	Mmus_LSm6	3,25	3,55	Hsap_LSm6	52,32	Infinite
	LSm7	Mouse	ENSMUSP00000071938	Ensembl	LSM7 homolog, U6 small nuclear RNA associated	Y	Mmus_LSm7	Equal! *	Equal! *	Hsap_LSm7	46,24	46,24
	SmD2	Mouse	ENSMUSP00000079230	Ensembl	[SMALL NUCLEAR RIBONUCLEOPROTEIN SM D2 SNRNP CORE D2 SM D2]	Y	Mmus_SmD2	1,14	1,19	Hsap_SmD2	52,7	Infinite
	SmG	Mouse	ENSMUSP00000050648	Ensembl	Small nuclear ribonucleoprotein G (snRNP-G) (Sm protein G) (Sm-G) (SmG)	Y	Mmus_SmG	Equal! *	Equal! *	Hsap_SmG	Infinite	Infinite
U2AF	U2AF35	Mouse	MG-BL-Q9D883-chr17	Blast	(Blast prediction on chrom.17 [3044018-3044734] based on Swiss protein Q9D883)	Y	Mmus_U2AG		7,03	Hsap_U2AG	391,11	Infinite

A.1.7 Other retrotransposed pseudo-genes

Table A.7 legend:

Disruption: appearance of frame disruption events (cryptic stop codons; frameshifts introduced by missing or extra nucleotides in the conserved coding region)

Ref. S1: closest active paralogue used for comparison

ds/dn S1: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the closest active paralogue (Ref. S1)

Ref. S2: active orthologue in the alternative species (Human/Mouse) used for comparison

ds/dn S2: rate of synonymous / non-synonymous substitutions calculated with SNAP for the comparison with the active orthologue (Ref. S2)

Table A.7: Other retrotransposed pseudo-genes

Group	Family	Species	Accession	Source	Description	Disruption	Ref. S1	ds/dn S1	Ref. S2	ds/dn S2
snRNP	p14	Human	HG-GW-SINFRUP00000137952-ch17	GeneWise	(GeneWise prediction on chrom.17 [55945331-55945673] based on Ensembl protein SINFRUP00000137952)	Y				
	S3A1	Mouse	MG-GW-Q8K4Z5-ch7	GeneWise	(GeneWise prediction on chrom.7 [74858572-74859706] based on SwissProt protein Q8K4Z5 -> ENSMUSP00000052962 in older versions of Ensembl)	Y				
	Tri-15	Mouse	ENSMUST00000023107	Ensembl	PseudoGene	Y	Mmus_NHP2l	0,88	Hsap_NHP2l	14,22
		Mouse	GENSCAN00000097595	Ensembl	(GenScan prediction on chrom.2 [71306700-71307083])		Mmus_NHP2l	3,30	Hsap_NHP2l	38,40
		Mouse	MG-GW-Q9D0T1-ch13	GeneWise	(GeneWise prediction on chrom.13 [82532570-82532953] based on SwissProt protein Q9D0T1)	Y				
		Mouse	MG-GW-Q9D0T1-ch3	GeneWise	(GeneWise prediction on chrom.3 [50618994-50619365] based on SwissProt protein Q9D0T1)	Y				
		Mouse	MG-GW-Q9D0T1-ch16	GeneWise	(GeneWise prediction on chrom.16 [50945250-50945660] based on SwissProt protein Q9D0T1 -> ENSMUSP00000062538 in older versions of Ensembl)	Y				
	U1AU2B	Human	HG-GW-ci0100139455-chr5	GeneWise	(GeneWise prediction on chrom.5 [56307180-56307830] based on JGI protein ci0100139455)	Y				
		Human	OTTHUMT00000042627	Ensembl	PseudoGene	Y	Hsap_U2B	1,60		
	U1C	Mouse	MG-GW-Q62241-ch16	GeneWise	(GeneWise prediction on chrom.16 [38852922-38853419] based on SwissProt protein Q62241)	Y				
		Mouse	MG-GW-Q62241-ch7	GeneWise	(GeneWise prediction on chrom.7 [114198579-114199049] based on SwissProt protein Q62241)		Mmus_U1C	2,73	Hsap_U1C	9,33
Sm	LSm2	Human	ENST00000310584	Ensembl	PseudoGene	Y	Hsap_LSm2	3,33	Mmus_LSm2	5,25
		Human	HG-GW-ci0100150332-chr19	GeneWise	(GeneWise prediction on chrom.19 [10570049-10570642] based on JGI protein ci0100150332)	Y				
		Human	*HG-GW-ci0100150332-chr5*	GeneWise	(GeneWise prediction on chrom.5 [108257947-108258231] based on JGI protein ci0100150332)	Y	Hsap_LSm2	2,02	Mmus_LSm2	7,87
	LSm3	Human	HG-BL-Q9Y4Z1-chr2_AC011236	Blast	(Blast prediction on chrom.2 [85241169-85241472] (clone AC011236) based on Swiss protein Q9Y4Z1)	Y				
		Human	HG-GW-ci0100133377-chr5	GeneWise	(GeneWise prediction on chrom.5 [65276827-65277121] based on JGI protein ci0100133377)	Y				
		Human	HG-GW-SINFRUP00000158728-chr12	GeneWise	(GeneWise prediction on chrom.12 [93487068-93487370]based on Ensembl protein SINFRUP00000158728)	Y				
		Human	HG-GW-SINFRUP00000158728-chr4	GeneWise	(GeneWise prediction on chrom.4 [143595999-143596301] based on Ensembl protein SINFRUP00000158728)	Y				
		Human	*HG-GW-SINFRUP00000158728-chr16*	GeneWise	(GeneWise prediction on chrom.16 [76946930-76947229] based on Ensembl protein SINFRUP00000158728)		Hsap_LSm3	1,78	Mmus_LSm3	11,10
	LSm5	Human	OTTHUMT00000081957	Ensembl	PseudoGene	Y				
		Human	OTTHUMT00000053994	Ensembl	PseudoGene	Y				
		Mouse	MG-GW-Q9Y4Y9-ch11	GeneWise	(GeneWise prediction on chrom.11 [120763476-120763754] based on Swiss protein Q9Y4Y9)	Y				
		Mouse	MG-BL-Q9Y4Y9-chr17_CAAA01192147	Blast	(Blast prediction on chrom.17 [34738177-34738449] (clone CAAA01192147)based on Swiss protein Q9Y4Y8)	Y				
	LSm6	Human	HG-BL-Q9Y4Y8-chr2_AC007179	Blast	(Blast prediction on chrom.2 [59653674-59653912] (clone AC007179) based on Swiss protein Q9Y4Y8)	Y				
		Human	HG-GW-ci0100154610-chr18	GeneWise	(GeneWise prediction on chrom.18 [57371361-57436789] based on JGI protein ci0100154610)	Y				
		Human	HG-GW-SINFRUP00000133790-ch12	GeneWise	(GeneWise prediction on chrom.12 [48451010-48550492] based on Ensembl protein SINFRUP00000133790)	Y				
		Mouse	MG-GW-Q9Y4Y8-ch7	GeneWise	(GeneWise prediction on chrom.7 [85682053-85682288] based on Swiss protein Q9Y4Y8)	Y				
		Mouse	MG-GW-Q9Y4Y8-ch12	GeneWise	(GeneWise prediction on chrom.12 [79654839-79655076] based on Swiss protein Q9Y4Y8)	Y				
		Mouse	MG-GW-Q9Y4Y8-ch18	GeneWise	(GeneWise prediction on chrom.18 [10367606-10367836] based on Swiss protein Q9Y4Y8)		Mmus_LSm6	0,44	Hsap_LSm6	10,51
	LSm7	Mouse	MG-GW-Q9CQQ8-chX	GeneWise	(GeneWise prediction on chrom.X [123734233-123746696] based on Swiss protein Q9CQQ8)	Y				
	SmBN	Human	HG-GW-ci0100151791-chr1	GeneWise	(GeneWise prediction on chrom.1 [2315257-2315656] based on JGI protein ci0100151791)	Y				
		Human	ENST00000333253	Ensembl	PseudoGene	Y	Hsap_SmN	1,65	Mmus_SmN	4,00
	SmD2	Mouse	ENSMUST00000069963	Ensembl	PseudoGene	Y	Mmus_SmD2	5,51	Hsap_SmD2	28,57
		Mouse	ENSMUST00000059224	Ensembl	PseudoGene	Y	Mmus_SmD2	5,54	Hsap_SmD2	8,24
	SmE	Human	ENST00000319409	Ensembl	PseudoGene	Y	Hsap_SmE	8,96	Mmus_SmE	25,37
		Human	ENST00000338402	Ensembl	PseudoGene	Y	Hsap_SmE	1,91	Mmus_SmE	2,69
		Human	HG-BL-P08578-chr1_AC099065	Blast	(Blast prediction on chrom.1 [220138549-220138820] (clone AC099065) based on Swiss protein P08578)	Y				
		Human	HG-BL-P08578-chr2_AC011747	Blast	(Blast prediction on chrom.2 [8693810-8694037] (clone AC011747) based on Swiss protein P08578)	Y				
		Human	HG-BL-P08578-chr2_AC093162	Blast	(Blast prediction on chrom.2 [85400821-85401096] (clone AC093162) based on Swiss protein P08578)	Y				
		Human	HG-BL-P08578-chr5_AC112191	Blast	(Blast prediction on chrom.5 [159686301-159686555] (clone AC112191) based on Swiss protein P08578)	Y				
		Human	OTTHUMT00000042623	Ensembl	PseudoGene	Y				
		Human	HG-GW-NP_111594-chr16	GeneWise	(GeneWise prediction on chrom.16 [20157795-20158070] based on GenBank protein NP_111594)	Y				
		Mouse	ENSMUST00000062417	Ensembl	PseudoGene	Y	Mmus_SmE	1,77	Hsap_SmE	8,34
	SmF	Human	HG-BL-Q15356-chr15_AC105036	Blast	(Blast prediction on chrom.15 [73592168-73592441] (clone AC105036) based on Swiss protein Q15356)	Y				
		Human	OTTHUMT00000079655	Ensembl	PseudoGene	Y				
		Human	HG-GW-SINFRUP00000122927-chr1	GeneWise	(GeneWise prediction on chrom.1 [201095326-201095558] based on Ensembl protein SINFRUP00000122927)	Y				
		Human	HG-GW-SINFRUP00000122927-chr3	GeneWise	(GeneWise prediction on chrom.3 [48177671-48177908] based on Ensembl protein SINFRUP00000122927)	Y				
	SmG	Human	OTTHUMT00000046823	Ensembl	PseudoGene	Y				
		Human	HG-BL-Q15357-chr11	Blast	(Blast prediction on chrom.11 [65038450-65038647] based on Swiss protein Q15357)	Y				
		Human	OTTHUMT00000045030	Ensembl	PseudoGene	Y				
		Human	HG-BL-Q15357-chr2_AC093762	Blast	(Blast prediction on chrom.2 [228468738-228468956] (clone AC093762) based on Swiss protein Q15357)	Y				
		Human	HG-BL-Q15357-chr2_AC104695	Blast	(Blast prediction on chrom.2 [28594655-28594881] (clone AC104695) based on Swiss protein Q15357)	Y				
		Human	*HG-BL-Q15357-chr2_AC010906*	Blast	(Blast prediction on chrom.2 [109326396-109326620] (clone AC010906) based on Swiss protein Q15357)		Hsap_SmG	2,07	Mmus_SmG	26,20
		Human	HG-BL-Q15357-chr21	Blast	(Blast prediction on chrom.21 [38796239-38796406] based on Swiss protein Q15357)	Y				
		Human	OTTHUMT00000082745	Ensembl	PseudoGene	Y				
		Human	HG-BL-Q15357-chr8	Blast	(Blast prediction on chrom.8 [128373660-128373869] based on Swiss protein Q15357)	Y				
		Human	HG-GW-ci010014951-chr1	GeneWise	(GeneWise prediction on chrom.1 [202052032-202052253] based on JGI protein ci0100149951)	Y				
		Human	OTTHUMT00000073702	Ensembl	PseudoGene	Y				
		Human	HG-GW-Q15357-chr17	GeneWise	(GeneWise prediction on chrom.17 [54713271-54713481] based on Swiss protein Q15357)	Y				
		Human	HG-GW-ci010014951-chr17	GeneWise	(GeneWise prediction on chrom.17 [64826527-64826697] based on JGI protein ci0100149951)	Y				
		Human	HG-GW-ci010014951-chr18	GeneWise	(GeneWise prediction on chrom.18 [50187255-50187482] based on JGI protein ci0100149951)	Y				

		Human	HG-GW-ci010014951-chr19	GeneWise	(GeneWise prediction on chrom.19 [14461200-14461421] based on JGI protein ci0100149951)	Y				
		Human	OTTHUMT00000050519	Ensembl	PseudoGene	Y				
		Human	HG-GW-ci0100149951-chr11	GeneWise	(GeneWise prediction on chrom.11 [92310257-92310481] based on JGI protein ci0100149951)	Y				
U2AF	U2AF35	Mouse	ENSMUST00000037657	Ensembl	PseudoGene	Y	Mmus_U2AG	3,68	Hsap_U2AG	37,12
		Mouse	ENSMUST00000050346	Ensembl	PseudoGene	Y	Mmus_U2AG	2,58	Hsap_U2AG	32,05
	U2AF65	Mouse	MG-GW-P26369-ch13	GeneWise	(GeneWise prediction on chrom.13 [9104361-9105705] based on SwissProt protein P26369 -> ENSMUSP00000049639 in older versions of Ensembl)	Y				
		Mouse	ENSMUST00000079393	Ensembl	PseudoGene	Y	Mmus_U2AF	2,11	Hsap_U2AF	19,68
SR	SC35	Human	ENST00000315132	Ensembl	PseudoGene	Y	Hsap_SC35	3,43	Mmus_SC35	4,61
		Human	HG-BL-Q01130-chr11	Blast	(Blast prediction on chrom.11 [94410402-94411073] based on Swiss protein Q01130)		Hsap_SC35	1,69	Mmus_SC35	1,86
		Human	HG-GW-ci0100146984-chrX	GeneWise	(GeneWise prediction on chrom.X [34165496-34166017] based on JGI protein ci0100146984)	Y				
	9G8-SRp20	Mouse	MG-BL-Q8R3E9-chr18	Blast	(Blast prediction on chrom.18 [8442599-8442916] based on SwissProt protein Q8R3E9 -> ENSMUSP00000052956 in older versions of Ensembl)	Y				
		Mouse	MG-BL-Q8R3E9-chr11	Blast	(Blast prediction on chrom.11 [98296117-98296674] based on Swiss protein Q8R3E9)		Mmus_SR20	0,00	Hsap_SR20	26,47
	Tra2	Mouse	MG-BL-ENSMUSP00000023564-chr3	Blast	(Blast prediction on chrom.3 [151104942-151105705] based on Ensembl protein ENSMUSP00000023564)	Y				
		Mouse	MG-BL-ENSMUSP00000023564-chr8	Blast	(Blast prediction on chrom.8 [] based on Ensembl protein ENSMUSP00000023564 -> ENSMUSP00000050950 in older versions of Ensembl)	Y				
hnRNP	hnRNP-A	Human	ENST00000315889	Ensembl	Heterogeneous nuclear ribonucleoprotein A3 pseudogene 1	Y	Hsap_ROA3	2,34	Mmus_ROA3	10,33
		Human	HG-BL-P51991-chr2	Blast	(Blast prediction on chrom.2 [175000259-175001257] based on Swiss protein P51991)	Y				
		Human	HG-BL-P51991-chr12	Blast	(Blast prediction on chrom.12 [50392755-50393657] based on Swiss protein P51991)	Y				
		Human	HG-BL-P51991-chr15	Blast	(Blast prediction on chrom.15 [55326664-55327383] based on Swiss protein P51991)	Y				
		Human	HG-BL-P51991-chr18	Blast	(Blast prediction on chrom.18 [28246202-28247194] based on Swiss protein P51991)	Y				
		Mouse	ENSMUST00000081086	Ensembl	PseudoGene	Y				
		Mouse	ENSMUST00000073495	Ensembl	PseudoGene	Y	Mmus_ROA3	3,46	Hsap_ROA3	73,69
		Mouse	ENSMUST00000045570	Ensembl	PseudoGene	Y	Mmus_ROA3	0,86	Hsap_ROA3	57,52
		Mouse	GENSCAN00000142819	Ensembl	(GenScan prediction on chrom.3 [158432758-158433672])		Mmus_ROA3	2,15	Hsap_ROA3	5,47
		Mouse	MG-BL-ENSMUSP00000078963-chr1a	Blast	(Blast prediction on chrom.1 [143583427-143584578] based on Ensembl protein ENSMUSP00000078963)	Y				
		Mouse	MG-BL-ENSMUSP00000078963-chr1b	Blast	(Blast prediction on chrom.1 [9643123-9644274] based on Ensembl protein ENSMUSP00000078963)		Mmus_ROA3	4,64	Hsap_ROA3	78,71
		Mouse	MG-BL-ENSMUSP00000078963-chr1c	Blast	(Blast prediction on chrom.1 [87451353-87452336] based on Ensembl protein ENSMUSP00000078963)	Y				
		Mouse	MG-BL-ENSMUSP00000078963-chr2	Blast	(Blast prediction on chrom.2 [131848808-131849659] based on Ensembl protein ENSMUSP00000078963)	Y				
	hnRNP-C	Human	HG-BL-P07910-chr11	Blast	(Blast prediction on chrom.11 [86413274-86414466] based on Swiss protein P07910)	Y				
		Human	HG-BL-P07910-chr15	Blast	(Blast prediction on chrom.15 [77315729-77316418] based on Swiss protein P07910)	Y				
		Human	OTTHUMT00000082465	Ensembl	PseudoGene	Y				
		Human	ENST00000329728	Ensembl	PseudoGene	Y	Hsap_ROC	1,36	Mmus_ROC	3,82
		Human	ENST00000317869	Ensembl	PseudoGene	Y	Hsap_ROC	1,06	Mmus_ROC	6,97
		Human	ENST00000323770	Ensembl	PseudoGene	Y	Hsap_ROC	1,16	Mmus_ROC	12,99
		Human	ENST00000357261	Ensembl	PseudoGene	Y	Hsap_ROC	1,33	Mmus_ROC	6,51
	hnRNP-E	Mouse	ENSMUST00000050197	Ensembl	PseudoGene	Y				
	hnRNP-F-H	Human	ENST00000327998	Ensembl	PseudoGene	Y	Hsap_ROF	1,75	Mmus_ROF	2,73
		Human	OTTHUMT00000042919	Ensembl	PseudoGene	Y				
		Human	ENST00000316594	Ensembl	PseudoGene	Y	Hsap_ROH1	2,60	Mmus_ROH1	3,48
		Mouse	ENSMUST00000054664	Ensembl	PseudoGene	Y	Mmus_ROF	1,82	Hsap_ROF	62,44
	hnRNP-G	Human	ENST00000320676	Ensembl	PseudoGene	Y	Hsap_ROG	1,56	Mmus_ROG	7,95
		Human	HG-GW-ci0100135546-chr4	GeneWise	(GeneWise prediction on chrom.4[110625602-110626222] based on JGI protein ci0100135546)	Y				
		Human	OTTHUMT00000040828	Ensembl	PseudoGene	Y				
		Human	OTTHUMT00000051985	Ensembl	PseudoGene	Y				
	hnRNP-K	Human	HG-BL-Q07244-chr2	Blast	(Blast prediction on chrom.2 [136790419-136791805] based on Swiss protein Q07244)	Y				
		Mouse	ENSMUST00000042280	Ensembl	PseudoGene	Y	Mmus_ROK	1,78	Hsap_ROK	13,11
		Mouse	ENSMUST00000051522	Ensembl	PseudoGene	Y	Mmus_ROK	1,42	Hsap_ROK	22,90
	hnRNP-L	Human	ENST00000309714	Ensembl	PseudoGene	Y				
		Human	ENST00000333525	Ensembl	PseudoGene	Y				
	hnRNP-R	Human	OTTHUMT00000077915	Ensembl	PseudoGene	Y				
		Human	ENST00000343438	Ensembl	PseudoGene	Y				
Others	PRP4	Human	OTTHUMT00000053788	Ensembl	PseudoGene	Y				
	SKIP	Human	HG-BL-Q13573-chr1	Blast	(Blast prediction on chrom.1 [78926758-78928362] based on Swiss protein Q13573)	Y				
	SRPK	Human	HG-GW-SINFRUP00000150696-chr8	GeneWise	(GeneWise prediction on chrom.8 [63938207-63939368] based on Ensembl protein SINFRUP00000150696)	Y				

A.2 Splicing Rainbow

A.2.1 Criteria for binding site detection

This section comprises three tables summarizing, for all the analysed splicing factors, the criteria that have been used for putative binding site definition in the Splicing Rainbow. To make the table contents consistent with standard sequence formats, Us are replaced by Ts and, for other nucleotide characters, the IUPAC[1] ambiguous nucleotide code is followed.

[1]International Union of Pure and Applied Chemistry - `http://www.iupac.org`

Table A.8: SR proteins - criteria for binding site detection

Factor	Motif size	Criteria	References
9G8	10-mer	Scoring matrix from SELEX data: S>6	[Cavaloc et al., 1999]
	9-mer	Scoring matrix from SELEX data: S>3.5	
ASF/SF2	18-mer	more than 15 matches with dsx PRE AAAGGACAAAGGACAAAA (*ad-hoc*)	[Hertel and Maniatis, 1998]
	10-mer	Scoring matrix from SELEX data: S>3.5	[Tacke and Manley, 1999]
	8-mer	Scoring matrix from SELEX data: S>2	
	7-mer	Scoring matrix from SELEX data: S>2.2	[Liu et al., 1998; Pollard et al., 2002] [Liu et al., 2001; Cartegni and Krainer, 2002]
SC35	11-mer	Scoring matrix from SELEX data: S>11	[Cavaloc et al., 1999]
	11-mer	Scoring matrix from SELEX data: S>6	
	10-mer	Scoring matrix from SELEX data: S>6	
	10-mer	Scoring matrix from SELEX data: S>6.3	
	7-mer	Scoring matrix from SELEX data: S>4	
	9-mer	Scoring matrix from SELEX data: S>6.5	[Tacke and Manley, 1999]
	9-mer	Scoring matrix from SELEX data: S>5	
	8-mer	Scoring matrix from SELEX data: S>2.1	[Liu et al., 2000a; Pollard et al., 2002] [Liu et al., 2001; Cartegni and Krainer, 2002]
	7-mer	Scoring matrix with threshold S>4, assuming $f_1(T)=f_2(G)=f_3(C)=f_5(G)=1$, $f_4(A)=f_4(G)=f_4(C)=0.125$, $f_4(T)=0.675$, $f_6(C)=f_7(C)=0.25$, $f_6(T)=f_7(T)=0.75$, all others $f_i(a)=0$ (*ad-hoc*)	[Schaal and Maniatis, 1999b] [Schaal and Maniatis, 1999a]
SRp20	9-mer	Scoring matrix from SELEX data: S>5.2	[Cavaloc et al., 1999]
	8-mer	Scoring matrix from SELEX data: S>4.2	
	7-mer	Scoring matrix from SELEX data: S>5.2	
	11-mer	More than 8 matches with GCTCCTCTTCC (*ad-hoc*)	[Lou et al., 1998]
	8-mer	More than 6 matches with CCTCGTCC (*ad-hoc*)	[Schaal and Maniatis, 1999b]
	7-mer	More than 6 matches with ATCTTTA (RBP1 for *Drosophila*) (*ad-hoc*)	[Heinrichs and Baker, 1995]
SRp40	18-mer	Scoring matrix from SELEX data: S>12	[Tacke et al., 1997]
	16-mer	Scoring matrix from SELEX data: S>10	
	6-mer	Scoring matrix from logo data: S>2.5	[Liu et al., 1998; Cartegni et al., 2002]
	5-mer	Scoring matrix from SELEX data: S>1.7	[Liu et al., 1998; Liu et al., 2001; Pollard et al., 2002]
SRp55	17-mer	Exact match: GNTCAACCNGGCGACNG (B52 for *Drosophila*)	[Shi et al., 1997]
	7-mer	Scoring matrix from logo data: S>2	[Liu et al., 1998; Cartegni et al., 2002]
	6-mer	Scoring matrix from SELEX data: S>2	[Liu et al., 1998; Liu et al., 2001; Pollard et al., 2002]
Tra2β	8-mer	more than 5.5 matches with AAGAAGAA (0.5 match $\Rightarrow$ alternative purine) (*ad-hoc*)	[Tacke et al., 1998; Modafferi and Black, 1999]

Table A.9: hnRNPs - criteria for binding site detection

Factor	Motif size	Criteria	References
hnRNP A0	5-mer	Exact match: ATTTA	[Myer and Steitz, 1995]
hnRNP A1	20-mer	Exact match: TATGATAGGGACTTAGGGTG	[Burd and Dreyfuss, 1994]
	6-mer	Scoring matrix from SELEX data: S≥8 ⇒ TAGGGW (strong) 8>S>5.5 (weak)	
hnRNP B1/A2	9-mer	poly-T ⇒ more than 5 Ts (*ad-hoc*)	[Brooks and Rigby, 2000]
	5-mer	Exact match: ATTTA	
	5-mer	Exact match: GTTTG	
	4-mer	Exact match: TTGA	[Ishikawa et al., 1993; Kajita et al., 1995]
hnRNP C	15-mer	poly-G ⇒ more than 11 Gs (*ad-hoc*)	[Soltaninassab et al., 1998]
	5-mer	Exact match: TTTTT	[Soltaninassab et al., 1998; Millard et al., 2000]
hnRNP D	5-mer	Exact match: ATTTA	[DeMaria and Brewer, 1996]
	4-mer	Exact match: TTGA	[Ishikawa et al., 1993; Kajita et al., 1995]
hnRNP E1/E2 (PCB)	9-mer	poly-C ⇒ more than 5 Cs (*ad-hoc*)	[Leffers et al., 1995]
	4-mer	Exact match: TTGA	[Ishikawa et al., 1993]
hnRNP F	9-mer	poly-G ⇒ more than 5 Gs (*ad-hoc*)	[Matunis et al., 1994]
	8-mer	Exact match: GGGGGCUG	[Chou et al., 1999; Min et al., 1997] [Modafferi and Black, 1999]
	4-mer	Exact match: GGGA	[Caputi and Zahler, 2001]
hnRNP G	10-mer	poly-A ⇒ more than 6 As (*ad-hoc*)	[Soulard et al., 1993]
hnRNP H	9-mer	poly-G ⇒ more than 5 Gs (*ad-hoc*)	[Matunis et al., 1994]
	6-mer	Exact match: TTGGGT	[Jacquenet et al., 2001]
	6-mer	Exact match: GGGGGC	[Caputi and Zahler, 2001; Chou et al., 1999] [Modafferi and Black, 1999; Min et al., 1997]
	5-mer	Exact match: TGTGG	[Chen et al., 1999]
	4-mer	Exact match: GGGA	[Caputi and Zahler, 2001]
hnRNP I (PTB)	10-mer	poly-Y ⇒ more than 7 Ys (*ad-hoc*)	[Chan and Black, 1997; Lou et al., 1999]
	6-mer	Exact match: TTCTCT	[Chan and Black, 1997]
	6-mer	Exact match: CTCTCT (stronger)	[Chan and Black, 1997; Ashiya and Grabowski, 1997] [Chan and Black, 1995; Modafferi and Black, 1999]
	4-mer	Exact match: TCTT	[Perez et al., 1997]
hnRNP K	11-mer	more than 8 matches with GGGGACTTTCC (kB enhancer element) (*ad-hoc*)	[Van Seuningen et al., 1995]
	8-mer	poly-C ⇒ more than 5 Cs (*ad-hoc*)	[Ostrowski et al., 2001; Leffers et al., 1995] [Swanson and Dreyfuss, 1988]
	8-mer	poly-T ⇒ more than 5 Ts (*ad-hoc*)	[Ostrowski et al., 2001]
	7-mer	Scoring matrix from SELEX data: S>7	[Thisted et al., 2001]
	6-mer	Scoring matrix from SELEX data: S>3.2	
hnRNP L	21-mer	more than 15 matches with CACCCACCCACATACATACAT (*ad-hoc*)	[Shih and Claffey, 1999]
hnRNP U	10-mer	strong affinity for poly-G, moderate affinity for poly-A and poly-T: S_{10j} >7 for $S_{10j} = \sum_{i=j}^{j+10} s_i(a)$, with s_i(G)=1, s_i(T)=s_i(A)=0.5, s_i(C)=0 (motif starting in nucleotide j) (*ad-hoc*)	[Kiledjian and Dreyfuss, 1992]

Table A.10: Other splicing factors - criteria for binding site detection

Factor	Motif size	Criteria	References
CELF (CUG-BP)	9-mer	strong affinity for CTG repeats: more than 7 matches with $(CTG)_3$ (*ad-hoc*)	[Takahashi et al., 2000; Lu et al., 1999]
HuR	10-mer	poly-T ⇒ more than 6 Ts (*ad-hoc*)	[Spangberg et al., 2000]
	5-mer	Exact match: ATTTA or TTTTT	[Sokolowski et al., 1999]
$U2AF^{65}$	4-mer	poly-Y: 4 Ys, 3 of them Ts (*ad-hoc*)	[Valcarcel et al., 1996]
Sxl	18-mer	Scoring matrix from SELEX data: S>11	[Singh et al., 1995]

A.2.2 Short pseudo-tutorial

1) Login to Windows (make sure Perl is installed).

2) Open Windows Explorer.

3) Go to folder `...\binding`.

4) Drop the file with your genomic sequence of interest (Fasta format, please) into this folder.

5) If available, drop the EMBL file with mRNA information (obtained from Gene2EST [Gemund et al., 2001], to be opened in Artemis [Berriman and Rutherford, 2003; Rutherford et al., 2000]) into the same folder.

6) Double click on `bdfinder.pl` to run the program.

7) A MS-DOS window will be opened and you will be asked to type the name of the file with the sequence. Do it and press `Enter`.

8) The same for the EMBL file... If not available, just press `Enter`.

9) Wait for some seconds (time proportional to the length of the sequence), until the window closes. The program will generate three files: the first has the name of the sequence file and `htm` extension; the second has the name of the sequence itself followed by `_EMBL.txt`; the third has the name of the sequence file followed by `results.txt`.

10) The first file can be opened with Internet Explorer. A colour code is used to visualize the putative binding sites for each splicing factor. Clicking on the underlined links (top) you can access information about the criteria used and the references. If you have included mRNA information, the last line will colourfully illustrate it. In this case the program will just consider introns for potential hnRNP binding sites and exons for SR proteins. Of course, regions that can be alternatively intronic and exonic will be considered suitable for the binding of both types of factors.

11) The second file should be opened in Artemis (together with the file with mRNA information, if desired). The color code is the same but the view is "saturated". Nevertheless, the bottom window gives you information about each putative binding site (type of factor, first and last nucleotides and score).

12) The third is a tab-delimited text file summarizing results.

13) For Linux the procedure is similar from step 3) on, except for: 6) Type `perl bdfinder.pl` to run the program.

Appendix B

Publications

This section presents the printouts of all the published articles associated with work described in this dissertation and whose contents are not integrally printed in the main text:

[Teschendorff et al., 2006a] Teschendorff AE, Naderi A, **Barbosa-Morais NL**, Caldas C. "PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer". *Bioinformatics*, 2006 May 8

[Stamm et al., 2006] Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, **Barbosa-Morais NL**, Thanaraj TA. "ASD: a bioinformatics resource on alternative splicing". *Nucleic Acids Res.*, 2006 Jan 1;34(Database issue):D46-55

[Teschendorff et al., 2005] Teschendorff AE, Wang Y, **Barbosa-Morais NL**, Brenton JD, Caldas C. "A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data". *Bioinformatics*, 2005 Jul 1;21(13):3025-33

[Pacheco et al., 2004] Pacheco TR, Gomes AQ, **Barbosa-Morais NL**, Benes V, Ansorge W, Wollerton M, Smith CW, Valcarcel J, Carmo-Fonseca M. "Diversity of vertebrate splicing factor U2AF^{35}: identification of alternatively spliced U2AF1 mRNAS". *J Biol Chem*, 2004 Jun 25;279(26):27039-49

[Naderi et al., 2004] Naderi A, Ahmed AA, **Barbosa-Morais NL**, Aparicio S, Brenton JD, Caldas C. "Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling". *BMC Genomics*, 2004 Jan 30;5(1):9