



**PROVAS ACADÉMICAS**  
NA FACULDADE DE MEDICINA DA UNIVERSIDADE DE LISBOA  
INSTITUTO DE FORMAÇÃO AVANÇADA

---

**Doutoramento:**

Ciências Biomédicas

**Nome do Aluno:**

Inês Maria de Stoop Camões Guerra Mollet

**Tema da Tese:**

Genome Wide Mining of Alternative Splicing in Metazoan Model Organisms

**Ramo:**

Ciências Biomédicas

**Especialidade:**

Ciências Morfológicas

**Data da Defesa:**

14 de Abril de 2009

**Classificação:**

Aprovada com Distinção e Louvor por Unanimidade

**Júri:**

Presidiu a Vogal Professora Doutora Maria do Carmo Fonseca da FMUL e estiveram presentes os vogais: Professores Doutores Didier Auboeuf, do Institut Universitaire d'Hématologie, Hôpital Saint Louis, Paris, França, Gil Ast, Tel Aviv University, Israel, Mário Ramirez, Margarida Gama-Carvalho, Luis Ferreira Moita e João Eurico da Fonseca da Universidade de Lisboa.



**PROVAS ACADÉMICAS**  
NA FACULDADE DE MEDICINA DA UNIVERSIDADE DE LISBOA  
INSTITUTO DE FORMAÇÃO AVANÇADA

---

## SUMÁRIO

### ***Objectivos***

O trabalho apresentado tem como objectivo o desenvolvimento de um algoritmo que gera uma base de dados de splicing alternativo e que detecta novos elementos envolvidos em splicing alternativo de maneira consistente e reprodutível. Este trabalho foi desenvolvido em duas fases. A primeira fase consistiu no desenvolvimento e optimização do algoritmo que cria uma base de dados de splicing alternativo para qualquer organismo a partir de informação de sequências depositados no GenBank [Mollet et al. 2008]. A segunda foi a aplicação dos dados gerados a análises de splicing alternativo de grande e pequena escala [Pacheco *et al.* 2006; Mollet *et al.* 2006].

### ***Conceitos Teóricos***

A quantidade fenomenal de sequências depositadas no GenBank todos os anos requer o desenvolvimento de algoritmos de triagem e condensação desta informação de modo a que esta seja útil ao biólogo. Só para o humano há mais de 8 milhões de sequências de transcritos depositadas. Um dos aspectos de enorme interesse é o do splicing alternativo que permite a um gene produzir variantes de uma proteína utilizando combinações alternativas de exões [Black 2003]. Para além da geração de variantes de proteínas, a geração de transcritos não codificantes é um mecanismo importante de regulação da expressão genética [Lewis *et al.* 2003; Blencowe 2006; Graveley 2001; Modrek e Lee 2002]. Estima-se que cerca de 73% dos cerca de 20,000 genes humanos sofram splicing alternativo [Johnson *et al.* 2003; Kan *et al.* 2001]. A informação útil retirada de bases de dados como o GenBank sobre splicing alternativo terá de ser útil não só para a análise de pequena escala, pelo biólogo que se dedica a aspectos específicos de alguns genes, mas também deverá ser útil para análises de grande escala, nomeadamente para análises por microarray. Os microarrays permitem monitorizar milhares genes em simultâneo num dado tecido. A comparação por microarray de dados em células de diferentes tecidos, ou de células saudáveis com células doentes, permite detectar os grupos de genes responsáveis pelas diferenças.

### ***Desenvolvimento do algoritmo***

O algoritmo foi desenvolvido na linguagem Perl interagindo com bases de dados construídas

em MySQL [Mollet *et al.* 2008]. Os dados de partida consistem em tabelas de alinhamentos de transcritos (mRNA e ETs) a um genoma pelo programa BLAT [Kent 2002], retiradas do UCSC Genome Browser [Karolchik *et al.* 2008]. A qualidade dos alinhamentos é analisada pelo programa que procede a uma triagem e à condensação de informação repetida. A informação condensada é utilizada para gerar tabelas com todo o splicing alternativo existente nos transcritos em relação a cada gene. O programa ainda determina a tradução dos transcritos, locais de poliadenilação e informação sobre expressão em diferentes tecidos ao nível de isoformas e exões individuais. O resultado consiste numa base de dados em MySQL com dez tabelas por genoma, optimizadas de modo a se poder fazer todo o tipo de pesquisas. A base de dados foi chamada ExonMine e, em colaboração com outros elementos do grupo, foi integrada num servidor no Instituto de Medicina Molecular com acesso ao público [<http://www.imm.fm.ul.pt/exonmine/>]. Os updates são gerados de quatro em quatro meses.

### ***Estudo estatístico - Descoberta de novos elementos***

O estudo estatístico dos dados gerados e uma comparação com outras bases de dados, revelou que a ExonMine detecta níveis mais altos de splicing alternativo do que qualquer outra base de dados. Uma análise dos dados revelou que se detectam mais exões novos, que não têm sequência codificante, no humano do que no rato. Sabe-se que transcritos não codificantes estão envolvidos na regulação da expressão de genes e que elementos móveis (como novos exões) estão na base da evolução de genes e de genomas em geral. Este resultado identifica portanto potenciais elementos que poderão ser fundamentais para se decifrar a diferença em termos de regulação genética, entre o humano e outros organismos menos complexos.

### ***Aplicações de grande escala***

#### *Procura de motivos envolvidos na regulação de splicing alternativo*

A base de dados ExonMine foi desenhada de modo a se poderem facilmente fazer procuras de elementos regulatórios envolvidos em splicing alternativo (ou mesmo de qualquer elemento regulatório a nível de sequência primária). O sucesso deste objectivo foi demonstrado num estudo em que, com base num pequeno algoritmo que questiona a base de dados ExonMine, se detectaram, no espaço de umas horas, combinações de elementos regulatórios que foram posteriormente validados experimentalmente e publicados [Pacheco *et al.* 2006] pela Doutora Teresa Pacheco.

#### *Construção de um microarray para estudo da Distrofia Muscular*

Em colaboração com o grupo do Doutor Juan Valcárcel (CRG-Centre de Regulació Genómica, Barcelona, Spain), a base de dados ExonMine foi utilizada para a construção de um microarray especificamente concebido para o estudo de alterações de splicing alternativo em distrofia muscular incluindo informação sobre cerca de 400 genes específicos de musculo e cerca de

400 factores de splicing.

#### *Análise do microarray comercial ExonArray 1.0 ST e de tiling arrays da Affymetrix*

Para aplicações de grande escala a base de dados está a ser utilizada para a análise de microarrays produzidos comercialmente pela Affymetrix com cerca de 5.5 milhões de sondas e que visam cobrir o genoma inteiro ao nível de um exão. Dados de 'tiling arrays' de Affymetrix que cobrem o genoma inteiro com uma resolução de 5 nucleótidos também foram comparados aos dados da ExonMine e são objecto de projectos em curso.

#### **Aplicações de pequena escala**

##### *Análise de variação do splicing na família U2AF*

A diversificação por splicing alternativo da importante família de factores de splicing U2AF foi investigada e publicada [Mollet et al. 2006]. Em todos os sete genes analisados foram detectados, com base no ExonMine, novas formas de splicing alternativo, com ou sem sequência codificante.

##### *Análise de novos elementos na família U2AF35*

Para a família de genes U2AF35 foi feita uma análise ainda mais aprofundada com o objectivo de validar as novas isoformas detectadas pelo ExonMine. Todas as isoformas, escolhidas para uma validação experimental, continham novos exões que se encontravam em humano mas não em rato, e foram confirmadas por RT-PCR por outros elementos do grupo da Professora Doutora Maria Carmo-Fonseca. Para além da validação experimental foi ainda feita uma análise do potencial codificante das novas isoformas e da conservação dos novos exões noutros genomas. Em dois casos de isoformas contendo novos exões, mas sem sequência codificante, foi feito o estudo da estrutura secundária do mRNA e foi detectado potencial para expressão de microRNAs, um dos quais foi confirmado experimentalmente por outro elemento do grupo da Professora Doutora Maria Carmo-Fonseca.

#### **Conclusões**

Neste trabalho não só se desenvolveu e optimizou um método reprodutível e eficaz para gerar informação sobre splicing alternativo para qualquer organismo, como também se realizaram várias análises que mostram claramente que os objectivos pretendidos foram atingidos. Foram detectadas mais variações em splicing alternativo do que em qualquer outra base de dados, incluindo formas de splicing alternativo envolvendo novos exões não codificantes, até agora ignorados. Algumas das novas formas de splicing alternativo foram validadas por elementos do laboratório de Professora Doutora Maria Carmo-Fonseca, e foram aplicados os dados à análise de genes em grande escala, nomeadamente para análise de motivos envolvidos na regulação do splicing alternativo e na análise de microarrays.

Esta base de dados é além disso um projecto activo e com continuidade, tendo vários projectos dependentes dos seus dados em execução, nomeadamente: análise de microarrays independentes no laboratório de Professora Doutora Maria do Carmo-Fonseca e em colaboração com o laboratório do Doutor Juan Valcárcel (CRG-Centre de Regulació Genómica, Barcelona, Spain); análise de splicing alternativo de factores de splicing (projecto em execução); e análise de splicing alternativo em factores de coagulação pelo grupo de Claire Shovlin (Imperial College, London). A base de dados ExonMine será mantida no Instituto de Medicina Molecular da Faculdade de Medicina da Universidade de Lisboa.