

Melissa Nunes Miziara

Análise Comparativa “*in Silico*” de
Sequências não Codificantes de Genes
Ortólogos entre Humano e Bovino





**ANÁLISE COMPARATIVA “*IN SILICO*” DE
SEQÜÊNCIAS NÃO CODIFICANTES DE GENES
ORTÓLOGOS ENTRE HUMANO E BOVINO**

Melissa Nunes Miziara

**Dissertação apresentada para
obtenção do Título de Mestre
em Genética.**

São José do Rio Preto - SP

**ANÁLISE COMPARATIVA "IN SILICO" DE
SEQÜÊNCIAS NÃO CODIFICANTES DE GENES
ORTÓLOGOS ENTRE HUMANO E BOVINO**

Melissa Nunes Miziara



t. 1.283/04

Dissertação apresentada ao Programa
de Pós-Graduação em Genética,
como parte das exigências para
obtenção do Título de Mestre.

**Orientador: Profª Drª M. Elisabete
J. Amaral**

**São José do Rio Preto – SP
Fevereiro de 2004**

577.23
M685a

202 0400 0

Miziara, Melissa Nunes.

Análise comparativa "*in silico*" de seqüências não codificantes de genes ortólogos entre humano e bovino / Melissa Nunes Miziara – São José do Rio Preto : [s.n.], 2004
66 f. : il. ; 30 cm.

Orientador: Maria Elisabete Jorge Amaral

Dissertação (mestrado) – Universidade Estadual Paulista. Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Genética animal. 3. Seqüência de nucleotídios. 4. Homologia. 5. Evolução molecular. 6. Bovino. 7. Bioinformática. 8. Ortologia. I. Amaral, Maria Elisabete Jorge. II. Universidade Estadual Paulista. Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU – 577.21



MELISSA NUNES MIZIARA

**ANÁLISE COMPARATIVA “*IN SILICO*” DE SEQÜÊNCIAS NÃO
CODIFICANTES DE GENES ORTÓLOGOS ENTRE HUMANO E BOVINO**

COMISSÃO JULGADORA

DISSERTAÇÃO PARA OBTENÇÃO DO GRAU DE MESTRE

Presidente e Orientador:

2º Examinador

3º Examinador

São José do Rio Preto, ____/____/____.



“Aos meus pais, que me apoiaram e permitiram a realização de um sonho. Sem vocês, eu não teria chegado até aqui”.

Agradecimentos

À minha orientadora Elisabete, que desde a graduação tem me incentivado a seguir os caminhos da pesquisa científica. Agradeço seu apoio, confiança e conselhos, os quais vou levar com gratidão.

Às minhas colegas Nedenia e Liliane, pelos papos no laboratório e pela ajuda prestada quando mais precisei.

À CNPq, por ter me concedido uma bolsa de estudos para realização deste trabalho.

À minha amiga Paola, que iniciou esta jornada comigo e me consolou em todas as vezes que não pude voltar para Franca por causa do trabalho, ouvindo minhas reclamações e histórias. Obrigada por sua grande amizade.

Ao meu namorado André, pelo amor, amizade e paciência que demonstrou em todos os momentos, principalmente nos mais difíceis, me mostrando que sou mais forte do que sempre imaginei.

Aos meus pais, por acreditarem em mim, pelo carinho e pela vontade de me ouvir dizendo que vou embora no final de semana para vê-los.

À minha querida irmã Maíra, que sempre me emociona quando diz ao telefone que está com saudades. Obrigada pelo companheirismo.

À minha avó Olga que, apesar de estarmos quase sempre distantes, está presente em meus pensamentos.

À minha avó Naninha, que queria estar comigo neste momento de realização. Vó, tenho certeza de que a senhora está me observando e transmitindo seu amor esteja onde a senhora estiver.

Obrigada a todos pelo carinho, amizade e consideração.

“Para realizar grandes conquistas devemos não apenas agir, mas também sonhar; não apenas planejar, mas também acreditar”.

Anatole France

*“Nunca ande pelo caminho
traçado, pois ele conduz somente
até onde os outros foram”.*

Alexandre Graham Bell

Inventor do telefone



SUMÁRIO

Introdução	1
Evidências da participação das seqüências não codificantes na regulação gênica de eucariotos	8
Objetivos	14
Artigo submetido à publicação	16
Discussão	41
Conclusões	51
Referências Bibliográficas	54
Resumo	63
Abstract	65

INTRODUÇÃO



Os genes de eucariotos são constituídos de seqüências de DNA que podem ser classificadas como seqüências codificantes e seqüências não codificantes. As seqüências codificantes são denominadas exons, que consistem de qualquer trecho não-intron da seqüência codificante de um gene. Quando unidos, os exons constituem o mRNA e são traduzidos em proteína (Griffiths, 2002). Da segunda classe de seqüências fazem parte seqüências de DNA que não codificam para proteínas ou RNAs estáveis (Hardison, 2000), incluindo as regiões 5' e 3' flanqueadoras não transcritas (*Flanking Regions* - FLRs), as regiões 5' e 3' transcritas não traduzidas (*Untranslated Regions* - UTRs) e os introns (Duret *et al.*, 1993). As seqüências flanqueadoras são regiões adjacentes à porção codificante do gene que não são transcritas, e que participam do processo de início e término da transcrição. As seqüências 5'-UTRs são definidas como a região que se estende do sítio *cap* até o códon de iniciação, enquanto as seqüências 3'-UTRs são definidas como a região do mRNA que se estende do códon de terminação até o sítio de início da cauda poli-A (Pesole *et al.*, 2000). Entre as seqüências não codificantes, as regiões 5' e 3'-UTR dos mRNAs eucarióticos têm demonstrado conter elementos cruciais para muitos aspectos da regulação gênica e da expressão, incluindo o controle do transporte de mRNAs para fora do núcleo, eficiência da tradução, localização celular e estabilidade do mRNA (Pesole, 1999; Pesole *et al.*, 2000; Larizza *et al.*, 2002; Mignone *et al.*, 2002). Finalmente, os introns são as seqüências presentes entre os exons, e que são removidas quando o transcrito primário é processado para resultar no mRNA maduro (Lewin, 2000). A importância dos introns ainda não é completamente conhecida. Recentes



estimativas da frequência de *splicing* alternativo no genoma humano indicam que mais de 50% dos genes conhecidos possuem dois ou mais transcritos produzidos pela geração de diferentes mRNAs devida a variações no padrão de *splicing* do pré-mRNA (Cooper, 2003; Hare e Palumbi, 2003).

A crescente disponibilidade de seqüências de DNA em bancos de dados públicos e os avanços da bioinformática vêm permitindo o desenvolvimento de novas abordagens e técnicas para a análise destas seqüências através da genômica comparativa, denominada como o estudo das diferenças e similaridades na estrutura e organização dos genomas em diferentes organismos (Primrose e Twyman, 2003).

A análise comparativa de seqüências de nucleotídeos está se mostrando cada vez mais importante como uma ferramenta para identificar regiões reguladoras no DNA genômico, compreender os mecanismos e história da evolução molecular, e deduzir as similaridades e diferenças na organização gênica entre espécies de interesse (Schwartz *et al.*, 2000).

Comparando seqüências de DNA genômico de diversas espécies, elementos funcionais podem ser reconhecidos com base na sua conservação evolutiva, já que importantes seqüências biológicas são conservadas entre espécies devido a restrições funcionais e, portanto, tendem a evoluir a uma taxa menor que as seqüências não funcionais (Batzoglou *et al.*, 2000; Meisler, 2001; Blanchette *et al.*, 2002; Ludwig, 2002; Sidow, 2002; Frazer *et al.*, 2003; Nobrega e Pennacchio, 2004). Os genes ortólogos são os genes mais comparados nessas análises e definem a base dos estudos de evolução molecular, pois, ao contrário de

pseudogenes, genes ortólogos sofrem pressão seletiva para reter seus sítios de ligação funcionais (Lenhard *et al.*, 2003). Ortólogos referem-se a genes homólogos entre diferentes genomas que derivaram de um mesmo gene presente na última espécie ancestral em comum (Koonin *et al.*, 1996; Wei *et al.*, 2002, Frazer *et al.*, 2003).

A principal estratégia utilizada para comparar as seqüências dos genes ortólogos de diferentes organismos é denominada *footprinting* filogenético. Este método consiste em identificar seqüências reguladoras de DNA com base na sua conservação, detectada no alinhamento do DNA genômico de diferentes espécies (Weitzman 2003, Nobrega e Pennacchio, 2004). O conceito baseia-se no fato de que importantes módulos reguladores estão sob pressão seletiva durante a evolução. A comparação de dois ou mais genomas resulta na identificação de seqüências conservadas que podem ser biologicamente relevantes (Weitzman 2003; Ureta-Vidal *et al.*, 2003; Lenhard *et al.*, 2003; Zhang e Gerstein, 2003).

Para que sejam feitas todas essas comparações de genomas, muitos autores têm desenvolvido programas computacionais para que essas análises sejam feitas de modo eficiente e confiável. Esses programas são baseados nos conceitos de alinhamento, identidade, similaridade e conservação. O alinhamento consiste em arranjar duas ou mais seqüências para alcançar níveis máximos de identidade, a fim de se identificar regiões que se mostram conservadas (<http://www.ncbi.nlm.nih.gov/Education>). Existem dois tipos básicos de alinhamento, o local e o global. Os alinhamentos do tipo local produzem valores de identidade ótimos entre sub-regiões das seqüências, enquanto os alinhamentos

do tipo global produzem valores de identidade ótimos para o tamanho total das duas seqüências que estão sendo comparadas (Frazer *et al.*, 2003). Alguns exemplos de programas de alinhamento são: *BlastZ*, que mostra seus resultados através da ferramenta de visualização de alinhamentos *PipMaker* (<http://bio.cse.psu.edu/>) (Schwartz *et al.*, 2000), *BLAST 2 sequences* (<http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>) (Tatusova e Madden, 1999), e *DNA Block Aligner* (DBA) (Jareborg *et al.*, 1999) (<http://www.sanger.ac.uk/Software/Wise2/dbaform.shtml>). Outros exemplos incluem o programa *GLASS* (<http://crossspecies.lcs.mit.edu>) (Batzoglou *et al.*, 2000); *AVID* (Bray *et al.*, 2003), que mostra seus resultados através da ferramenta de visualização de alinhamentos *VISTA* (<http://www-gsd.lbl.gov/vista>) (Mayor *et al.*, 2000), e *ClustalW* (<http://www.ebi.ac.uk/clustalw/>) (Higgins, 1994).

A identidade é medida pelo número de nucleotídeos ou aminoácidos idênticos resultante do alinhamento entre duas seqüências, e a similaridade é o grau de relação entre duas seqüências nucleotídicas ou protéicas, medido pelo percentual de identidade (<http://www.ncbi.nlm.nih.gov/Education>).

A conservação de seqüências nucleotídicas ou protéicas é definida através de critérios de tamanho e porcentagens de identidade para um alinhamento. Segmentos conservados com percentual de identidade X e tamanho Y são definidos como regiões nas quais cada sub-região de tamanho Y foi pelo menos X% idêntico à outra seqüência com a qual foi alinhada (Mayor *et al.*, 2000). O primeiro trabalho a utilizar esse critério para seleção de seqüências de nucleotídeos conservadas foi o de Loots *et al.* (2000), que definiu uma seqüência



não codificante conservada (do inglês, *Conserved Noncoding Sequence* – CNS) como uma sequência que apresentasse um tamanho maior ou igual a 100bp, com no mínimo 70% de identidade com a outra sequência com a qual foi alinhada.

Com base na utilização de ferramentas computacionais, vários trabalhos compararam sequências de DNA não codificantes de diferentes espécies com o objetivo de identificar elementos que possam participar da regulação gênica.

Duret *et al.* (1993) revisaram genes incluídos no banco de dados *GenBank* (<http://www.ncbi.nlm.nih.gov/GenBank>) que contém sequências não codificantes. Seus resultados revelaram a existência de sequências altamente conservadas com mais de 70% de similaridade em regiões não codificantes dos genes de espécies pertencentes a diferentes classes de vertebrados.

Analisando comparações de 1196 sequências de mRNA e proteína de genes ortólogos entre humano e camundongo, Makalowski *et al.* (1996) identificaram 85% de identidade média em sequências protéicas e um valor médio de 68% de identidade para sequências do tipo UTRs. Outros trabalhos também mostram a descoberta de conservação em regiões não codificantes utilizando comparações entre humano e espécies de roedores (Oeltjen *et al.*, 1997; Ansari-Lari *et al.*, 1998; Shabalina *et al.*, 2001; Gottgens *et al.*, 2001; Couronne *et al.*, 2003).

A escolha das espécies a serem comparadas é essencial para a eficácia do *footprinting* filogenético, sendo influenciada pela distância evolutiva das espécies em questão (Duret e Bucher, 1997; Chain, 2001; Thomas e Touchman, 2002). Se as espécies são muito próximas, como por exemplo humano e chimpanzé, a



distinção de elementos reguladores de regiões não funcionais é dificultada, porque não há tempo evolutivo suficiente para o acúmulo de mutações em bases neutras. Mas se as espécies são muito distantes, como um mamífero e um réptil, então detectar elementos reguladores conservados pode também ser difícil, pois os organismos divergiram demais para preservar qualquer similaridade significativa, ou os processos reguladores são diferentes nas duas linhagens estudadas. Portanto, elementos reguladores que foram adquiridos muito recentemente na evolução podem não ser detectáveis por *footprinting* filogenético.

Quando são comparadas seqüências ortólogas de duas espécies relacionadas evolutivamente, é possível identificar seqüências não codificantes conservadas (CNSs). Porém é difícil diferenciar dentre essas CNSs, quais se apresentaram conservadas devido a uma pressão seletiva para função em processos reguladores, daquelas seqüências não funcionais que se mostraram conservadas devido ao pequeno acúmulo de alterações e falta de divergência entre as espécies estudadas. A solução então consiste em comparar mais de duas seqüências ortólogas (Duret e Bucher, 1997). Dessa forma, é possível detectar seqüências não codificantes que apresentam conservação ativa, ou seja, CNSs ativamente conservadas (<http://www-gsd.lbl.gov/vista>).



Evidências da participação das seqüências não codificantes na regulação gênica de eucariotos

Baseado na suposição de que seqüências não codificantes ativamente conservadas entre humano e camundongo estarão presentes em um terceiro mamífero, Dubchak *et al.* (2000) compararam uma seqüência ortóloga de aproximadamente 200kb do cromossomo 5 humano, 11 de camundongo e 4 de cachorro. O trabalho revelou que os alinhamentos dos três organismos possuem padrões significativos de conservação ativa nas seqüências não codificantes.

Loots *et al.* (2000) procuraram por seqüências não codificantes conservadas nas regiões ortólogas dos genes das interleucinas 4, 13 e 5 de várias espécies, incluindo coelho, bovino, humano e camundongo. Estes autores verificaram uma seqüência conservada de 401 nucleotídeos localizada entre os genes da interleucina 4 e 13, denominada CNS-1. Experimentos com camundongos transgênicos mostraram que a deleção dessa seqüência afeta a expressão das interleucinas em células T, diminuindo mas não cessando sua produção.

Realizando comparações entre camundongo e o peixe baiacu, Aparicio *et al.* (1995) utilizou o gene *Hoxb-4*, membro da família das proteínas *homeobox*, para analisar conservação em seqüências não codificantes. Foram encontrados três segmentos com altas porcentagens de identidade em uma seqüência onde previamente haviam sido descritos *enhancers* em camundongo. Ao testar as funções desses segmentos através de transgênese, observou-se que, quando

deletado, o segmento denominado CR1 causava perda de atividade tecido-específica do gene. Um outro segmento, denominado CR3 de baiacu, quando inserido em camundongos transgênicos, mostrou influenciar na coordenação da expressão do gene *Hoxb-4* de maneira conservada entre mamíferos e teleósteos.

Pizzuti *et al.* (2002), utilizando as técnicas de *SSCP* (Polimorfismo de conformação de fita simples) e sequenciamento de DNA, identificaram 3 alelos do gene humano da resistina que variam no número de repetições ATG presentes na região 3'-UTR. Portadores do alelo com 6 repetições ATG possuem alta sensibilidade à insulina, sugerindo que este polimorfismo pode estar exercendo um papel na susceptibilidade individual para resistência à insulina e diabetes tipo 2.

Wedemeyer *et al.* (2000) estudaram a região 3'-UTR do gene *Rab1a*, que codifica uma proteína reguladora essencial para o transporte de vesículas do retículo endoplasmático ao Complexo de Golgi. O trabalho consistiu em comparar a sequência não codificante de 27 espécies de vertebrados, incluindo desde répteis a mamíferos. A conservação de aproximadamente 91% encontrada dentro da região 3'-UTR desse gene, mesmo entre espécies com grande distância filogenética, sugere que essa sequência contém sítio(s) de ligação de proteínas ligantes de RNA, e está envolvida na função celular pós-transcricional.

Cinco elementos não codificantes altamente conservados foram identificados no gene *SCL* (*Stem Cell Leukemia*) de peixes, mamíferos e galinha, sugerindo seu papel essencial para o padrão apropriado de transcrição deste gene,



que codifica um fator de transcrição que participa da hematopoiese e vasculogênese (Göttgens *et al.*, 2001).

Utilizando a metodologia do *footprinting* filogenético, Nobrega *et al.* (2003) compararam a sequência do gene *DACH* (*Dachshund homolog to Drosophila*) humano e suas regiões intergênicas com a porção cromossômica ortóloga de camundongo. Esta comparação resultou na identificação de 1098 CNSs. Adicionando sequências de outros vertebrados à análise, como anfíbio e o baiacu, o número de sequências destacadas decaiu para 32. Desses elementos candidatos, nove foram testados *in vivo* utilizando camundongos transgênicos, sendo que sete deles se mostraram participantes da regulação da expressão do gene *DACH*. Este gene é expresso em vários tecidos e está envolvido no desenvolvimento do cérebro, membros e órgãos sensoriais.

Comparando um segmento do cromossomo 7 humano que contém o gene supressor de tumor *ST7* (*Supression of Tumorigenicity 7*) com seus segmentos homólogos em babuíno, bovino, camundongo e baiacu, Frazer *et al.* (2003) encontraram aproximadamente 90% de identidade em sequências intergênicas e introns entre humano e babuíno. A maioria dos elementos conservados identificados nas comparações humano-bovino e humano-camundongo não representaram exons do gene *ST7*, sugerindo a presença de sequências potencialmente funcionais.

Hare e Palumbi (2003) testaram se o padrão de conservação encontrado em introns de nove genes de baleia, humano e foca seria consistente com a divergência neutral. Foram encontrados 163 blocos altamente conservados entre



as 3 espécies, cujos nucleotídeos diferiram em aproximadamente 14% no alinhamento múltiplo. Essa conservação, que foi encontrada em 34% da sequência humana utilizada, sugeriu que as sequências de introns exercem um papel funcional na regulação gênica dos genes analisados.

Williams *et al.* (2003) encontraram regiões de introns do gene *CFTR* (regulador de condutância transmembrânica da fibrose cística) de humano que se mostraram conservadas em sequências de bovino e porco. Conservação de porções não codificantes do gene *CFTR* também foi obtida por Thomas e Touchman (2002), que compararam a sequência genômica de humano contendo esse gene com sequências ortólogas de babuíno, camundongo, bovino e baiacu. Thomas *et al.* (2003), Ellsworth *et al.* (2000) e Vuillaumier *et al.* (1997) também utilizaram o gene *CFTR* em suas análises comparativas com camundongo e outras espécies de vertebrados, e seus resultados indicaram números substanciais de segmentos não codificantes conservados em sequências flanqueadoras e em introns desse gene.

Em um outro estudo envolvendo genes da família *HOX*, Santini *et al.* (2003), utilizando o programa de alinhamento *PipMaker*, compararam sequências dos *clusters HOXA* de tubarão, humano, camundongo e 4 espécies de peixe. Dentre os vários segmentos conservados identificados, um elemento presente dentro de um intron chamado HB1, que consiste em sítios de ligação homeodomínio, mostrou ser altamente conservado em vários dos genes estudados nessas espécies.



É possível observar que na literatura existem trabalhos recentes envolvendo o estudo de seqüências não codificantes de bovino. Os trabalhos encontrados utilizam as seqüências desta espécie para auxiliar na identificação de elementos reguladores investigados em segmentos cromossômicos humanos, focalizando suas pesquisas no conhecimento da regulação gênica em humano. Somente um trabalho teve como foco de pesquisa a regulação de um gene em bovino (Hering *et al.*, 1995). Comparando a seqüência de cDNA do gene da proteoglicana de bovino com suas seqüências correspondentes em porco, humano, galinha e rato, este trabalho revelou conservação não só na seqüência codificante, mas também nas regiões 3' e 5'-UTRs do mRNA deste gene.

De acordo com Thomas e Touchman (2002), a comparação de seqüências de genes entre humano e bovino revela uma extensiva conservação fora dos exons. O mapeamento comparativo dos genomas de humano e bovino mostra um alto grau de conservação de sintenia entre as duas espécies, sendo essa conservação tão ou mais extensiva que a conservação entre os genomas de humano e camundongo (Womack, 1987 e O'Brien *et al.*, 1988). Utilizando-se a técnica de Zoo-FISH, Solinas-Toldo *et al.* (1995) estabeleceram 50 segmentos de homologia entre os cromossomos humano e bovino. Todos os cromossomos humanos, exceto o cromossomo Y, hibridizaram com um ou mais cromossomos bovinos, indicando grande identidade genômica entre os dois organismos.

No estudo envolvendo a técnica de mapeamento comparativo *in silico*, Band *et al.* (2000) incluíram 1.087 marcadores genéticos no mapa RH comparativo dos genomas de humano e bovino, dos quais 768 representaram



genes conhecidos. Aproximadamente 83% desses genes em bovino puderam ser identificados como idênticos aos genes de humano. Além dessa identidade, foram encontrados entre os genomas dessas duas espécies 105 segmentos cromossômicos conservados sintenicamente contendo mais de dois genes.

Estes dados fornecem uma visão da semelhança genômica entre humano e bovino, o que permite inferir que a comparação de seqüências não codificantes pode revelar importantes segmentos conservados evolutivamente com potencial funcional.

O desenvolvimento da pesquisa genômica em animais de interesse econômico fornece importantes informações tanto para a manipulação de genes de interesse quanto para a compreensão da genética do desenvolvimento, reprodução, doenças e funcionamento dos genes humanos (Hetzl, 1993). Além disso, a base fisiológica de características economicamente importantes e o potencial das interações intergênicas ou genética/ambiental não poderão ser inteiramente compreendidos até que os processos responsáveis pela regulação e expressão desses genes estejam identificados.



OBJETIVOS

Com base na contribuição deste estudo para o conhecimento da participação de seqüências não codificantes no processo de regulação da expressão de genes de humano e bovino, o presente trabalho teve como objetivos:

- identificar a presença de conservação em seqüências não codificantes de pares de genes ortólogos entre humano e bovino;

- indicar a ocorrência de seqüências ativamente conservadas com potencial funcional nos genes estudados.



*ARTIGO SUBMETIDO À
PUBLICAÇÃO*

Revista Mammalian Genome – ISSN: 0938-8990/ New York - USA



Comparative analysis of noncoding sequences of orthologous bovine and human gene pairs.

Melissa Nunes Miziara¹, Penny K. Riggs² and M. Elisabete J. Amaral¹

¹ Departamento de Biologia, IBILCE - UNESP, São José Rio Preto – SP, 15054-000, Brasil.

Fax: 55-17-221-2390

Phone: 55-17-221-2407

Correspondence: M E J Amaral (eamaral@bio.ibilce.unesp.br)

² M.D. Anderson Cancer Center, Science Park - Research Division, Smithville, TX, USA

(penny@krtraining.com)

Fax: 512-237-2444

Phone: 512-237-9442

▪ **Running head:** Comparative analysis of noncoding sequences



Abstract

Genomic sequence comparison across species has enabled the elucidation of important coding and regulatory sequences encoded within DNA. Of particular interest are the noncoding regulatory sequences, which influence gene transcriptional and posttranscriptional processes. A phylogenetic footprinting strategy was employed in the present study to identify noncoding conservation patterns of 39 human and bovine orthologous genes. Seventy-three conserved noncoding sequences were identified that shared greater than 70% identity over at least 100bp. Thirteen of these conserved sequences were also identified in the mouse genome. Evolutionary conservation of noncoding sequences across diverse species may have functional significance, and these conserved sequences may represent candidate regulatory elements.



Introduction

Increasing interest in genome structure and function has resulted in improved tools for comparative genomics. Putative coding and regulatory sequences have been identified by genomic sequence analysis of diverse species. Noncoding sequences have been shown to influence gene transcriptional and posttranscriptional processes (Huang and Gorman 1990; Liu and Redmond 1998; Wedemeyer et al. 2000; Mazumber et al. 2003). Although noncoding DNA exhibits more sequence variation across species than exons, some of these sequences have been retained by selective pressure (Jareborg et al. 1999). These conserved noncoding sequences may represent candidate regulatory elements. Phylogenetic footprinting has been frequently used to identify these important sequences whose function is not clearly understood. This experimental approach is based on the idea that important regulatory modules remain under selective pressure during evolution, and comparison of two (or more) genomes will identify the conserved sequences that are most likely to be biologically relevant (Weitzman, 2003). The comparative sequence analyses generally use orthologous genes to identify similarities and differences between genomes. Orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species (Koonin et al. 1996).

Previous studies have identified both noncoding sequence conservation (Loots et al. 2000, Batzoglou et al. 2000, Dubchak et al. 2000, Frazer et al. 2001, Chapman et al. 2003, Williams et al. 2003, Lenhard et al. 2003, Nobrega et al. 2003) as well as sequence divergence (Larizza et al. 2002, Cooper et al. 2003). These reports demonstrate that a comparative genomics approach can be an efficient tool for identifying functional sequence elements and providing insights into regulation of genomic machinery. Few studies have utilized cattle as a target organism. Comparison of human and rodent sequences comprises the bulk of comparative genomic studies. Human – bovine comparative mapping, however, reveals a high



degree of genetic conservation, as extensive as human – mouse conservation (O'Brien et al. 1988 and Womack 1987). By a Zoo-FISH approach, Solinas-Toldo et al. (1995) showed that each human chromosome, except chromosome Y, hybridizes to one or more bovine chromosomes, indicating a high level of conservation between these two organisms. Despite the similarity between human and bovine genomes, most previous efforts that utilized comparison of bovine sequences with other genomes focused on human genome analysis and studied specific human chromosome segments (Thomas and Touchman 2002; Frazer et al. 2003; Thomas et al. 2003; Williams et al. 2003). Only Hering et al. (1995) focused on bovine genome, which identified conservation on bovine, human, pig, chicken and rat 5' and 3' untranslated regions of the chondrocyte link protein cDNA sequence.

Considering evidence of human – bovine genetic similarity and the economic importance of identifying and understanding regulation of genetic loci associated with desirable traits, the aim of the present investigation was to identify sequence conservation and candidate functional elements in noncoding regions of 39 human and bovine orthologous.



Material and Methods

In order to investigate conservation between human and bovine noncoding sequences, we searched 1573 *Bos taurus* genes from the Bovmap database (September 2002 release) (<http://locus.jouy.inra.fr/>). Genes that had incomplete mRNA and protein sequences and had no noncoding sequences comparable to *Homo sapiens* sequences obtained from the Locuslink database (Pruitt and Maglott 2001) were excluded. Moreover, nucleotide and amino acid sequences that showed lesser than 70% identity between human and bovine were also excluded. A total of 39 genes were identified as orthologous between human and bovine. Comparable noncoding sequences of these genes (Table 1) were obtained from the GenBank® database (Benson et al. 2003) and extracted by SeqVista graphical tool (Hu et al. 2003). When multiple accessions existed for the same sequence, we picked the one with bigger sequence. The pairwise human/bovine alignments were performed by Avid alignment algorithm and displayed by Vista graphical server (Bray et al. 2003), both applying default parameters. A conserved noncoding sequence (CNS) was identified as orthologous regions sharing at least 70% sequence identity across at least 100bp.

Noncoding sequences that showed conservation between human and bovine were analyzed in multispecies comparison using mouse as the third organism. Mouse sequences were obtained from GenBank®. Human/bovine/mouse alignments were performed using Avid alignment algorithm and displayed using Vista graphical server (Bray et al. 2003; Mayor et al. 2000). Cutoff criteria for defining actively conserved noncoding sequences were calculated using intersection/union (I/U) analyses available at the Vista server. The actively conserved noncoding sequences were identified as the sequences that were conserved in all three pairwise alignments - human/bovine, human/mouse and bovine/mouse.



To visualize identical nucleotide positions of the actively conserved noncoding sequences, human/bovine/mouse alignments were displayed using ClustalW (Higgins et al. 1994) and Genedoc (Nicholas et al. 1997) programs.



Results

One hundred sixteen human/bovine alignments corresponding to 95 introns, 15 3'-untranslated regions (3'UTRs) and 6 5'-UTRs were generated. Twenty-eight of the 39 analyzed orthologous genes showed conservation in their noncoding sequences. Seventy-three conserved noncoding sequences (CNSs) with $\geq 70\%$ identity over $\geq 100\text{bp}$ were obtained in 36 introns, 12 3'-UTRs and one 5'-UTR, totaling 49 noncoding sequences. Comparative results are summarized in Table 2. Lengths of CNSs ranged in size from 101 to 485bp for introns, from 102 to 1116bp for 3'-UTRs and 131bp for 5'-UTR. The conservation of these CNSs ranged from 70 to 93.7% for introns, from 70 to 91.7% for 3'-UTRs, and 73.3% for 5'-UTR the CNS.

The 49 human and bovine noncoding sequences in which conservation was detected were submitted to multiple alignment analysis with the corresponding orthologous sequences of mouse, using the Avid algorithm. Vista plots (Figure 1) were produced to determine actively conserved noncoding sequences in the human/bovine/mouse alignments. In the figure, sequence conservation peaks in intron 3 of RPL3. The cutoff criteria established by I/U analyses for each analyzed sequence ranged from 51 % to 100% conservation over 100bp for human/bovine alignments, and from 50% to 99% conservation over 100bp for bovine/mouse alignments.

By Vista multiple analysis, 45 human/bovine conserved noncoding sequences were identified. Of these 45 sequences, 13 exhibited at least 70% conservation in the three pairwise alignments (human/bovine, human/mouse and bovine/mouse) were considered actively conserved noncoding sequences. The genes that showed actively conserved noncoding sequences in 3'-UTRs included CRTL1, FBN1, LIF, LUM, PDE5A and PPP1R8, while ODC1, RPL3, SLC25A11, SLC25A3 and TNF showed active conservation in introns.

The identical nucleotides of these sequences were verified by Genedoc alignments. An example of the conservation of RPL3 intron 3 is shown in Figure 2.

Percentage identities of all actively conserved noncoding sequences identified from the human/bovine/mouse alignments are listed in Table 3. The most conserved sequence identified was a portion of RPL3 intron3, which showed 100, 97.4 and 99% identity across human/bovine, human/mouse and bovine/mouse alignments, respectively. On the other hand, the least conserved sequence identified was a portion of PDE5A 3'-UTR, which shared 79.4, 72.6 and 70.4% identity in human/bovine, human/mouse and bovine/mouse alignments, respectively. The length of the conserved sequences ranged from 92bp to 913bp.



Discussion

The objective of this analysis was to identify conservation between human and bovine orthologous noncoding sequences. Our human/bovine comparison is based on phylogenetic footprinting, the principle that important regulatory modules are retained via selective pressure during evolution, and that comparison of at least two divergent genomes can reveal conserved sequences that are most likely to be biologically relevant (Weitzman 2003).

Most of analyzed 3'-UTRs possessed CNSs, while only 36 of 95 introns shared CNSs. The results indicate that there was much more CNSs in 3'-UTRs than in introns, and are according to a previous study that identified highly conserved regions in a set of genes from mammals, birds, amphibians and bony fishes (Duret and Bucher 1997). These found conserved sequences were three times more frequent in 3'-noncoding regions than in introns.

The percentages of identities observed in these data showed to be higher in 3'-UTRs (average identity = 77.2%) than in introns (average identity = 73.2%). This result agrees with the expectation that untranslated regions are more conserved than introns because of their crucial role in post-transcriptional and post-translational processes. Similar results were obtained by Jareborg et al. (1999), which compared 77 orthologous mouse and human gene pairs and revealed almost ten times more 90% identical regions in 3' and 5' untranslated regions than in introns. Only 1% of the analyzed introns showed this high identity. From the present work, the most identical CNSs were those found in FBN1 3'-UTR, which showed 303bp and 91.7% identity, and in PPP1R8 3'-UTR, which showed 81.2% identity over 1116bp.

One of the six analyzed 5'-UTRs possessed a conserved sequence, which was less identical than 3'-UTRs and introns CNSs. The lack of identity of 5'-UTR conserved sequence is a surprising result because one might have expected a strong selective pressure on 5'-UTRs, as these sequences contain elements involved in regulation of transcription (Duret and Bucher



1997). However, because only one 5'-UTR sequence was analyzed, we cannot generalize about 5'-UTR conservation across the genome.

With presumably fewer functional constraints, noncoding sequences can accumulate neutral mutations and evolve more rapidly than coding portions of the genomes. When a noncoding sequence encodes some regulatory function, selective pressure may drive sequence conservation within the region (Jareborg et al. 1999; Nobrega and Pennacchio 2004). It is possible that some of the 73 noncoding sequences examined in the present study showed conservation because of lack of divergence time due to intermediate evolutionary distance (70-100 Myrs) between human and bovine. In this case, the lack of divergence would interfere with detection of functional noncoding elements. Thus, it was necessary to compare more than two evolutionary related species. In this report, human/bovine sequence alignments were further examined by comparison to mouse orthologous sequences to verify CNS region. By the use of multiple analyses, it was possible to identify sequences that shared similarities because of functional constraints.

As a result of human/bovine/mouse alignments, active conservation was detected in six 3'-UTRs and seven introns, totalizing thirteen noncoding sequences from eleven genes. Assuming that noncoding sequences that have remained similar during evolution might play roles in gene regulation, the high levels of conservation identified in the assessed sequences may infer presence of candidate functional elements. The results of this study indicate that significant similarity exists between human and bovine noncoding gene sequences, and that comparative analysis of noncoding sequences between these two genomes provides putative regulatory sequences that can be experimentally tested to confirm if they participate in these genes regulation. Thus, these results will help to elucidate the regulation process of these genes in human and bovine genomes.



Acknowledgements

We are grateful to CNPq – Brazil, Grant number 130541/2002-8 which has supported MN Miziara, and to FAPESP, Grant 97/13403-1 to MEJ Amaral. We further wish to thank Dr. JE Womack for many suggestions and critical reading of the manuscript.



References

Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* 10, 950-958

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) Genbank. *Nucleic Acids Res* 31, 23-7

Bray N, Dubchak I, Pachter L (2003) AVID: A global alignment program. *Genome Res* 13, 97-102

Chapman MA, Charchar FJ, Kinston S, Bird CP, Grafham D, Rogers J, Grützner F, Graves JAM, Green AR, Göttgens B (2003) Comparative and functional analyses of *LYLI* loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics* 81, 249-259

Cooper GM, Brudno M, NISC Comparative Program, Green ED, Batzoglou S, Sidow A (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* 13, 813-820

Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* 10, 1304-1306

Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* 7, 399-406



Frazer KA, Sheehan JB, Stokowski RP, Chen X, Hosseini R, Cheng JF, Fodor SP, Cox DR, Patil N (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res* 11, 1651-1659

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* 13, 1-12

Hering TM, Kollar J, Huynh TD, Sandell LJ (1995) Bovine chondrocyte link protein cDNA sequence: interspecies conservation of primary structure and mRNA untranslated regions. *Comp Biochem Physiol B Biochem Mol Biol* 112, 197-203

Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (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, 4673-4680

Hu Z, Frith M, Niu T, Weng Z (2003) SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC Bioinformatics* 4, 1-8

Huang MTF, Gorman CM (1990) Intervening sequences increase efficiency of RNA 3' processing and accumulation of cytoplasmic RNA. *Nucleic Acids Res* 18, 937-947

Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* 9, 815-824



Koonin EV, Mushegian AR, Borr P (1996) Non-orthologous gene displacement. *Trends Genet* 12, 334-336

Larizza A, Makalowski W, Pesole G, Saccone C (2002) Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyls and rodent gene pairs. *Comput Chem* 26, 479-490

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW (2003) Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology* 2, 13-13.11

Liu S-Y, Redmond M (1998) Role of the 3'-untranslated region of RPE65 mRNA in the translational regulation of the RPE65 gene: identification of a specific translation inhibitory element. *Arch Biochem Biophys* 357, 37-44

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons. *Science* 288, 136-140

Makalowski W, Zhang J, Boguski MS (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* 6, 8456-857



Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. APPLICATIONS NOTE 16, 1046-1047

Mazumber B, Seshadri V, Fox PL (2003) Translational control by the 3'UTR: the ends specify the means. Trends in Biochemical Sciences 28, 91-98

Nicholas KB, Nicholas HB Jr, and Deerfield, DW II (1997) GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS 4, 14

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302, 413

Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological Discovery. J Physiol 554, 31-39

O'Brien SJ, Seuánez HN, Womack JE (1988) Mammalian genome organization: An evolutionary overview. Annu Rev Genet 22, 323-351

Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29, 137-140

Solinas-Toldo S, Lengauer C, Fries R (1995) Comparative Genome Map of Human and Cattle. Genomics 27, 489-496



Thomas JW, Touchman JW (2002) Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet* 18, 104-108

Thomas JW et al (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788-793

Wedemeyer N, Schmitt-John T, Evers D, Thiel C, Eberhard D, Jockusch HC (2000) Conservation of the 3'- untranslated region of the *Rab1a* gene in amniote vertebrates: exceptional structure in marsupials and possible role for posttranscriptional regulation. *FEBS Letters* 477, 49-54

Weitzman JB (2003) Tracking evolution's footprints in the genome. *Journal of Biology* 2, 9-9.4

Williams SH, Mouchel N, Harris A (2003) A comparative genomic analysis of the cow, pig, and human *CFTR* genes identifies potential intronic regulatory elements. *Genomics* 81, 628-639

Womack JE (1987) Genetic engineering in agriculture: Animal genetics and development. *Trends in Genetics* 3, 65-68

Internet Homepages

Bovmap database - <http://locus.jouy.inra.fr/>

Locuslink - <http://www.ncbi.nlm.nih.gov/LocusLink/>



Vista graphical server - <http://www-gsd.lbl.gov/vista/>

ClustalW - <http://www.ebi.ac.uk/clustalw/>

Genedoc - <http://www.psc.edu/biomed/genedoc/>

GenBank database- <http://www.ncbi.nlm.nih.gov/GenBank/>



Table 1. GenBank accession numbers of noncoding sequences of 39 orthologous genes between human and bovine.

gene symbol	analyzed noncoding sequence	human accession number	bovine accession number
ACTA1	introns	M20543	U02285
AVP	introns	X62890	X00503
CYP21A2	introns	M26856	M11267
FDX1	introns	M23665 and M23668	M19474 and D00470
FSHB	introns	M54914	M83753
HBB	introns	L26462	X00376
INHBA	introns	X57578	U16239
KRT10	introns	X14487	X02870
NPPA	introns	K02043	M13145
ODC1	introns	M81740	U36394
OXT	introns	M11186	X00502
PNMT	introns	X52730	M36706
PPY	introns	M11726	L33970
PTH	introns	J00301	K01938
RHO	introns	U49742	U89240
RPL3	introns	AJ238852	AJ238851
SLC11A1	introns	D50402	AF064488
SLC25A11	introns	X66114	X66115
SLC25A3	introns	X77337	X77338
STAT5A	introns	AJ412877, AJ412878, AJ412879, AJ412880, AJ412881, AJ412882, AJ412883, AJ412884 and AJ412885	AJ242522 and AJ237937
STAT5B	introns	AJ412888, AJ412890, AJ412891, AJ412892, AJ412893, AJ412894, AJ412896 and AJ412898	AJ237934, AJ237935 and AJ237936
TNF	introns	X02910	AF011926
UCP1	introns	X51952	AF139921
BGN	3'-UTR	U11686	U68143
CHAD	3'-UTR	U96769	U08018
FBN1	3'-UTR	X63556	L28748
LUM	3'-UTR	U18728	L11063
NRG1	3'-UTR	L41827	L12259
PDE5A	3'-UTR	AJ004865	L16545
PPP1R8	3'-UTR	U14575	Z50748
PRSS7	3'-UTR	U09860	U09859
RPE65	3'-UTR	U18991	L11356
SERPINF1	3'-UTR	U29953	AF017058

CYB5	5'-UTR	AH003052	M63326
SCD	5'-UTR	AF320307	AB075020
CRTL1	3' and 5'-UTR	M63579 and X17405	U02292
PTGS2	3' and 5'-UTR	U04636	AF031698
COX4I1	introns and 3'-UTR	AF017115	AH003440, U11068, U11069 and U11070
LIF	introns and 3'-UTR	M27053	D50337 and U65394

Table 2. Results of noncoding sequences comparison obtained from Avid/Vista analysis.
(CNS) conserved noncoding sequence.

noncoding sequence	number of analysed noncoding sequences	number of noncoding sequences containing CNSs	number of CNSs
Introns	95	36	47
3'-UTRs	15	12	25
5'-UTRs	6	1	1

Table 3. Percentage identities of actively conserved noncoding sequences determined using Vista multiple alignments. (CNS) conserved noncoding sequence.

gene symbol	noncoding sequence	% identity of human/bovine CNS	% identity of human/mouse CNS	% identity of bovine/mouse CNS
ODC	INTRON 1	85.9	83.6	85.9
RPL3	INTRON 3	100	97.4	99
	INTRON 7	89	92.2	88.5
SLC25A11	INTRON 1	86.5	78.8	79.4
SLC25A3	INTRON 3	85.8	84.4	94.1
	INTRON 5	92.4	91.1	89.8
TNF	INTRON 3	86.9	74.7	76.2
CRTL1	3'UTR	98.1	89.2	90.2
FBN1	3'UTR	97.5	78.8	83.1
LIF	3'UTR	81.9	75.5	72.8
LUM	3'UTR	92.3	82.4	78.8
PDE5A	3'UTR	79.4	72.6	70.4
PPP1R8	3'UTR	98.2	95.4	95.4

Figure 1. Vista plot corresponding to intron 3 of RPL3 gene. The positions of the sequences are shown in the horizontal axis and the percent identities (50% - 100%) are indicated in the vertical axis. The pink peak represents an actively conserved noncoding sequence identified in the three pairwise alignments – human/bovine, human/mouse and bovine/mouse.

Figure 2. Alignment of human, bovine and mouse RPL3 intron 3 using Genedoc multiple alignment program. Note the identical positions of nucleotides.

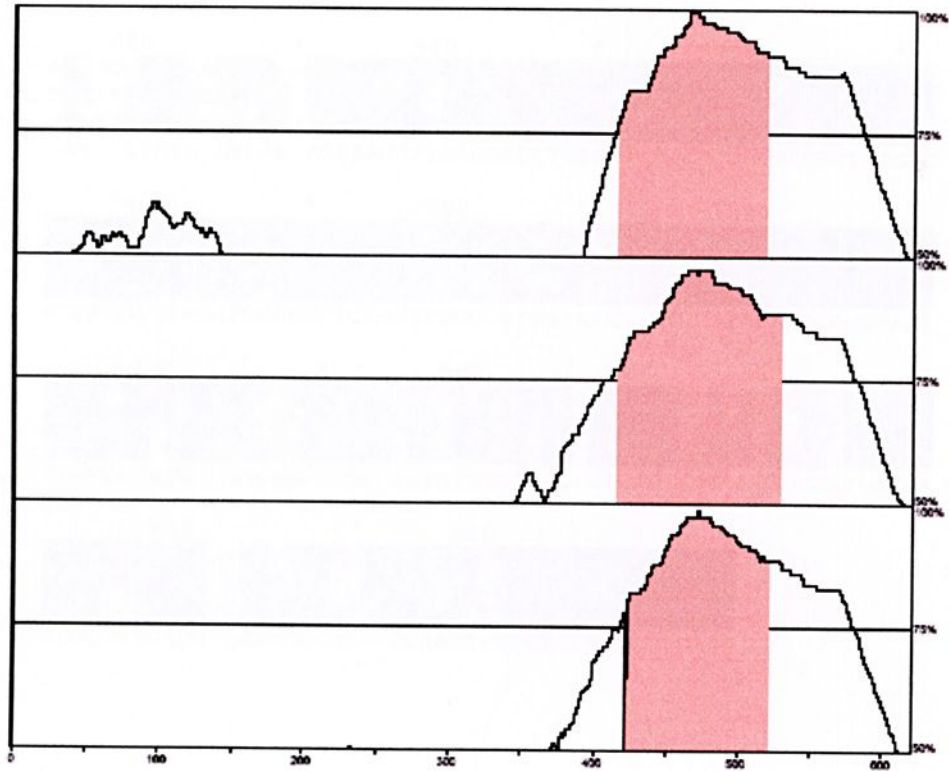
Alignment 1
Seqs: heterobas
C-Rev: 100%, 100 bp
Region: 1

Alignment 2
Seqs: heterobas
C-Rev: 90%, 100 bp
Region: 1

Alignment 3
Seqs: revusica
C-Rev: 90%, 100 bp
Region: 1

X-axis: homp
Resolution: 1
Window size: 100 bp

■ gene
■ exon
■ UTR
■ CDS
■ intron



420 * 440 * 460 *
 Hs-RPL3-I3 : AATGCGTGTAACTTA-ATTAACTTGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 456
 Mm-RPL3-I3 : TATATTGTGGCTTAT-ATTAACTTGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 420
 Bt-RPL3-I3 : GAATGTTGTGTTTCATTATTAACTTGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 391
 At tGTg cTtAt ATTAACCTtGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC

480 * 500 * 520 *
 Hs-RPL3-I3 : CGATGAGCTTCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 515
 Mm-RPL3-I3 : CGATGAGCTTCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 479
 Bt-RPL3-I3 : CGATGAGCTTCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 450
 CGATGAGCTcCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT

540 * 560 * 580 *
 Hs-RPL3-I3 : GAGCTTATTCGACGCGGTGCTAATGTGCTTCTGCTCAGCTCGGAGGGTGATGA : 574
 Mm-RPL3-I3 : GAGCTTATTCGACGCGGTGCTAATGTGCTTCTGCTCAGCTCGGAGGGTGATGA : 538
 Bt-RPL3-I3 : GAGCTTATTCGACGCGGTGCTAATGTGCTTCTGCTCAGCTCGGAGGGTGATGA : 509
 GAGcCCaTCCCCAcacGGTGCCaGTgTGCCTtCC CTCACCgTtGgAgGGgTGATGA

600 * 620 *
 Hs-RPL3-I3 : AGGCTGCACC-TGCTCCCTCCCCAACTCTGCTCTGCTCCTGAAG : 619
 Mm-RPL3-I3 : AGGCTGCACC-GGCCCCCTCCCCAACTCTGCTCTGCTCCTGAAG : 583
 Bt-RPL3-I3 : AGGCTGCACC-TGCTCCCTCCCCAACTCTGCTCTGCTCCTGAAG : 555
 AGGCctGCACC gGGcCCCctCCCCAACTcTGCTCTGCTCCTGAAG

DISCUSSÃO



A busca de conservação envolvendo 116 seqüências não codificantes de 39 pares de genes ortólogos entre humano e bovino revelou 73 regiões conservadas definidas na literatura como seqüências com mais de 100bp e identidade de nucleotídeos igual ou superior a 70%. Este critério foi atribuído em vários trabalhos para identificação de seqüências não codificantes conservadas (CNSs) (Loots, 2000; Kent e Zahler, 2000; Frazer *et al.*, 2003; Nobrega *et al.*, 2003; Couronne *et al.*, 2003). Das CNSs identificadas, 47 se apresentaram dentro de 36 dos 95 introns analisados; 25 foram identificadas em 12 das 15 regiões 3'-UTRs, e apenas uma CNS foi encontrada em uma das 6 regiões 5'-UTRs analisadas. Os resultados obtidos indicam que neste conjunto de genes, há um número maior de CNSs em seqüências do tipo 3'-UTR que em introns, concordando com os dados obtidos na literatura de que regiões que mostram alta conservação em genes de vertebrados são três vezes mais encontradas em regiões 3'-UTRs que em introns (Duret e Bucher, 1997).

Além da detecção de regiões conservadas ser maior em regiões 3'-UTRs que em introns, o percentual de identidade de nucleotídeos nessas regiões também foi maior que nas seqüências encontradas nos introns. Analisando as conservações obtidas, observou-se que as CNSs dos introns, com valor médio de identidade de 73,2%, se mostraram menos idênticas que as CNSs das regiões 3'-UTRs, que apresentaram identidade média de 77,2%. Resultados similares foram registrados na literatura, onde a comparação de 77 pares de genes ortólogos entre humano e camundongo revelou quase 10 vezes mais regiões com identidade de nucleotídeos maior que 90% em regiões 3' e 5'-UTRs que em introns (Jareborg *et al.*, 1999).



De acordo com Pesole *et al.* (2000) e Mignone *et al.* (2002), entre as regiões não codificantes, as regiões 5' e 3'-UTR dos mRNAs eucarióticos apresentam elementos cruciais para muitos aspectos da regulação e expressão gênica. Estes aspectos incluem o controle do transporte de mRNAs para fora do núcleo, eficiência da tradução, localização celular e estabilidade. Outro exemplo na literatura que encontrou um alto grau de conservação nessas regiões é o trabalho de Makalowski e Boguski (1998), que analisaram alinhamentos de 2.820 pares de seqüências protéicas e de mRNA de genes ortólogos entre rato, camundongo e humano. As identidades encontradas nos alinhamentos humano-camundongo, humano-rato e camundongo-rato foram de 71, 70,1 e 86,3%, respectivamente.

Nas análises do presente trabalho, destacam-se as CNSs encontradas nas regiões 3'-UTRs dos genes FBN1 e PPP1R8. A CNS do primeiro gene apresentou 303bp e 91,7% de identidade e a CNS do gene PPP1R8 se mostrou 81,2% idêntica em 1116 nucleotídeos. A CNS que apresentou a porcentagem de identidade mínima para a classificação de conservada (70%) foi a registrada no intron 1 do gene ACTA1.

Os resultados obtidos mostram ainda que também ocorre uma maior conservação em seqüências do tipo UTRs que em introns entre os ortólogos de humano e bovino, muito embora nossos dados contenham na sua quase totalidade regiões 3'-UTRs.

Seqüências funcionais tendem a evoluir a uma taxa mais baixa que seqüências não funcionais, e a ser altamente conservadas entre os mamíferos, sugerindo a análise comparativa de seqüências como uma estratégia para sua

identificação (Loots *et al.*, 2000; Frazer *et al.*, 2003). Entretanto, comparações entre somente duas espécies de mamíferos tornam difícil estimar a proporção de seqüências conservadas que realmente apresentam alguma função reguladora (Meisler, 2001). Se as espécies são muito próximas, pode ser complicado selecionar seqüências candidatas a funções na regulação gênica. Por outro lado, se as espécies são muito distantes, algoritmos de alinhamento podem não identificar elementos importantes potencialmente envolvidos na regulação (Chapman *et al.*, 2003).

Existem vários trabalhos da literatura que comparam seqüências não codificantes de humano com seqüências de espécies distantes evolutivamente em até 400 milhões de anos (Aparicio *et al.*, 1995; Bagheri-Fam *et al.*, 2001; Gilligan *et al.*, 2002). Apesar dessas comparações serem úteis na identificação de seqüências de importância funcional que se mantiveram por um longo tempo evolutivo, a detecção de elementos reguladores adquiridos recentemente não pode ser realizada. Por outro lado, comparações entre espécies muito próximas, como por exemplo, humano e outros primatas (Boffelli *et al.*, 2003), que divergiram há aproximadamente 6 milhões de anos, só conseguem identificar elementos que diferem de uma espécie para outra, e são úteis para determinar a base genética da variação fenotípica, e não as semelhanças entre espécies (Chain, 2001).

Considerando o tempo de divergência primatas-artiodáctilos de 68-102 milhões de anos (Thomas e Touchman, 2002), é possível que parte das CNSs identificadas nas comparações humano/bovino reflitam a falta de divergência entre as duas espécies e representem seqüências não funcionais que não

acumularam mutações ao longo da evolução dessas duas espécies, superestimando o número real de seqüências com importância funcional. A solução para essa problemática consiste em alinhar seqüências de mais de duas espécies. A utilização de mais de duas espécies na análise comparativa de genomas é baseada na hipótese de que seqüências não codificantes ativamente conservadas entre humano e camundongo, por exemplo, estarão presentes em um terceiro mamífero, enquanto que seqüências não codificantes que são similares entre humano e camundongo somente devido ao acúmulo insuficiente de mutações casuais estarão ausentes no terceiro mamífero (Dubchak *et al.*, 2000; Frazer *et al.*, 2001). Assim, se as CNSs registradas nas comparações humano/bovino também forem identificadas em alinhamentos múltiplos utilizando-se um terceiro organismo, a probabilidade de essas seqüências apresentarem um potencial funcional é maior.

Frazer *et al.* (2003) afirmam que comparações feitas com múltiplas espécies que divergiram há aproximadamente 40-80 milhões de anos, tais como humano/camundongo e humano-bovino, permitem determinar quais seqüências não codificantes se mostram conservadas nessas três espécies e quais poderiam ser consideradas ativas. De acordo com Li *et al.* (1990), Elzirik *et al.* (2001) e Ureta-Vidal *et al.* (2003), a distância evolutiva humano-bovino/camundongo é de aproximadamente 80-100 milhões de anos.

Portanto, em nosso estudo, além da detecção de conservação entre seqüências não codificantes de humano e bovino, análises múltiplas envolvendo seqüências de um terceiro organismo foram realizadas, a fim de se “filtrar” as CNSs para que sejam identificadas somente aquelas seqüências que

potencialmente apresentam importância biológica, também denominadas CNSs ativamente conservadas.

Para excluir as CNSs identificadas devido a falta de divergência entre humano e bovino, foram realizados alinhamentos múltiplos, incluindo-se as seqüências do genoma de camundongo. Utilizando-se o critério de Dubchak *et al.* (2000) de que a presença de seqüências conservadas em três espécies é devida à conservação ativa e não à ancestralidade, é possível que, se uma seqüência que se mostrou conservada entre humano e bovino também estiver presente com altos percentuais de identidade numa seqüência de camundongo, é possível que esta seqüência seja funcional.

Nas análises dos alinhamentos múltiplos, foram identificadas 13 CNSs ativamente conservadas em 11 genes, com tamanho variando de 92 a 913 nucleotídeos com identidade igual ou superior a 70% nos três alinhamentos, humano/bovino, humano/camundongo e bovino/camundongo. Foram identificadas 7 CNSs ativamente conservadas nos introns e 6 nas regiões 3'-UTRs. Proporcionalmente, foram registradas mais CNSs ativamente conservadas em regiões 3'-UTRs que em introns. Estes resultados confirmam as observações já citadas anteriormente de que a probabilidade de se encontrar elementos reguladores em seqüências do tipo UTRs é maior que em introns, embora a CNS proveniente da 5'-UTR não tenha sido indicada como ativamente conservada. A maior CNS ativamente conservada foi registrada na região 3'-UTR do gene *FBN1*, apresentando tamanho médio nos três alinhamentos de 383bp, e a menor CNS ativa foi encontrada no intron 1 do gene *ODC1*, com 100bp de tamanho

médio nos três alinhamentos. As CNSs com maior percentual de identidade nos três alinhamentos estão registradas no intron 3 do gene RPL3, com valor médio de identidade de 98,8%, e na região 3'-UTR do gene PPP1R8, com valor médio de identidade de 96,3%. Entre as seqüências com menor percentual de identidade de nucleotídeos, estão a CNS da região 3'-UTR do gene LIF, com valor médio de identidade nos três alinhamentos de 76,7%, e a CNS da região 3'-UTR do gene PDE5A, com 74,1% de identidade média. Observando as altas porcentagens de identidade registradas nas CNSs ativamente conservadas dos genes RPL3 e PPP1R8, é possível inferir que existem elementos dentro dessas seqüências com alguma função conservada na regulação desses genes.

Não existem até o momento registros na literatura que descrevam as seqüências conservadas identificadas neste estudo como elementos reguladores, exceto para a região 3'-UTR do gene CRTL1. Hering *et al.* (1995) analisaram a seqüência de cDNA do gene CRTL1 de bovino, obtida por PCR e seqüenciamento, através da comparação com seqüências correspondentes em porco, rato, humano e galinha. Foram registradas conservações tanto na região 5' quanto na região 3'-UTR. Em particular, um motivo conservado do tipo A(T_n)A foi encontrado dentro da região 3'-UTR entre as 4 espécies estudadas, sugerindo sua participação na regulação da estabilidade do mRNA deste gene. Este motivo não foi identificado dentro da CNS obtida pelo presente estudo.

É possível observar que as CNSs ativamente conservadas encontradas no presente trabalho se apresentaram mais conservadas entre humano e bovino que entre humano e camundongo ou bovino e camundongo. Resultados similares



foram obtidos por Thomas e Touchman, que comparou uma porção do gene *CFTR* com as seqüências ortólogas de babuíno, bovino, camundongo e o baiacu. De acordo com esses autores, o menor grau de conservação observado entre as seqüências de humano e camundongo é consistente com a alta taxa de mutação observada na linhagem de camundongo, devido a um maior número de gerações, referido como “hipótese do efeito do tempo de geração” (Li *et al.*, 1990; Ohta, 1993), ou seja, as altas taxas ocorrem porque camundongos possuem um tempo menor de geração. Os resultados de Thomas *et al.* (2003) confirmam que os genomas de camundongos estão sofrendo modificações mais rápido que os de primatas, carnívoros e artiodáctilos, concordando com o *Mouse Genome Sequencing Consortium* (2001) e Li *et al.* (1996). Como as linhagens de primatas e artiodáctilos evoluem mais ou menos na mesma taxa, enquanto que a linhagem de camundongo evolui 1.5 vezes mais rápido (Wu e Li, 1985; Li *et al.*, 1990), as altas taxas de mutação em seqüências de camundongo poderiam explicar a menor identidade observada nas comparações humano/camundongo e bovino/camundongo.

As razões para a conservação de seqüências não codificantes incluem principalmente sua relevância na regulação da expressão gênica e manutenção da organização do genoma (Jareborg *et al.*, 1999). Porém, somente evidências de uma conservação evolutiva do DNA não pode, diretamente, indicar função. A identificação da conservação fornece uma estratégia para revelar e priorizar seqüências para experimentos biológicos posteriores (Pennacchio e Rubin, 2003). Atualmente, estudos do envolvimento das seqüências ativamente conservadas na



regulação estão sendo realizados principalmente através de experimentos com camundongos transgênicos, que podem fornecer testes definitivos de função para elementos reguladores candidatos de transcrição (Meisler, 2001).

As análises *in silico* representam uma primeira varredura de um genoma de interesse, e vem mostrando ser eficiente em revelar seqüências candidatas a alguma função no genoma. Um exemplo registrado na literatura da utilidade desse tipo de abordagem técnica para se estudar elementos reguladores é o trabalho de Nobrega *et al.* (2003), que utilizou a metodologia do *footprinting* filogenético para comparar o gene *DACH* e suas regiões espaçadoras entre humano e camundongo. Suas análises identificaram 1098 CNSs. Adicionando seqüências de outros vertebrados à análise, 32 seqüências mostraram ser CNSs ativamente conservadas, das quais nove foram testadas *in vivo* utilizando camundongos transgênicos. Sete delas foram identificadas como elementos participantes da regulação da expressão do gene *DACH*.

Considerando-se que o genoma de bovino apresenta um tamanho estimado de 2.990-3532 cM (O'Brien *et al.*, 1999), estudos *in silico* são de grande importância para orientar a seleção das seqüências candidatas aos testes experimentais, tendo em vista a elaboração e complexidade exigidas nestes experimentos.

As análises *in silico* de seqüências de DNA não codificantes podem contribuir grandemente para o nosso entendimento da estrutura e função dos genes e do genoma bovino. O presente trabalho identificou nos 39 pares de genes ortólogos entre humano e bovino, 13 seqüências candidatas a potenciais funções



na regulação dos seus respectivos genes. O valor dessas CNSs ativamente conservadas será intensificado pela determinação acurada da sua função biológica e/ou dos tipos de elementos reguladores presentes.



CONCLUSÕES

O presente estudo apresentou as seguintes conclusões descritas abaixo:

- Das 116 seqüências não codificantes provenientes de 39 pares de genes ortólogos human/bovino, 49 (em 28 genes) apresentaram conservação entre os genomas de humano e bovino. Essa conservação foi evidenciada em 73 seqüências conservadas (CNSs);
- As CNSs encontradas nas regiões 3' transcritas não traduzidas (3'-UTRs) se mostraram mais conservadas e foram identificadas em um número três vezes maior que nos introns;
- A CNS com maior percentual de identidade de nucleotídeos foi registrada na região 3'-UTRs do gene FBN1, com 91,7% de identidade. A CNS com menor percentual de identidade de nucleotídeos foi registrada no intron 1 do gene ACTA1, com 70% de identidade;
- Foram identificadas 13 CNSs ativamente conservadas, sendo 6 provenientes de regiões 3'-UTRs e 7 provenientes de introns. Proporcionalmente, o número de conservações ativas registrado foi maior para 3'-UTRs que para introns;
- As CNSs ativamente conservadas com maiores percentuais de identidade nos três alinhamentos são a do intron 3 do gene RPL3, com valor médio de identidade de 98,8%, e da região 3'-UTR do gene PPP1R8, com valor médio de identidade de 96,3%. As seqüências com menor percentual de identidade de nucleotídeos, são a CNS da região 3'-UTR do gene LIF, com valor médio de



identidade nos três alinhamentos de 76,6%, e a CNS da região 3'-UTR do gene PDE5A, com 74,1% de identidade média;

- As CNSs ativamente conservadas encontradas no presente trabalho apresentaram maior porcentagem de identidade entre humano e bovino que entre humano e camundongo ou bovino e camundongo;

- O genoma de camundongo se mostrou adequado para a seleção de CNSs ativamente conservadas;

- As 13 CNSs ativamente conservadas representam regiões candidatas a testes experimentais, a fim de se estabelecer o papel funcional dessas seqüências na regulação gênica dos seus genes.



ANSARI-LARI, M. A., et al. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. **Genome Research**, v. 8, p. 29-40. 1998.

APARICIO, S., et al. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. **Proc. Natl. Acad. Sci. USA**, v. 92, p. 1684-1688. 1995.

BAGHERI-FAM, S., et al. Comparative genomics of the SOX9 region in human and *Fugu rubripes*: conservation of short regulatory sequence elements within large intergenic regions. **Genomics**, v. 78, p. 73-82. 2001.

BAND, M. R., et al. An ordered comparative map of the cattle and human genomes. **Genome Research**, v. 10, p. 1359-1368. 2000.

BATZOGLOU, S., et al. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. **Genome Research**, v. 10, p. 950-958. 2000.

BLANCHETTE, M.; SCHWIKOWSKI, B.; TOMPA, M. Algorithms for phylogenetic footprinting. **Journal of Computational Biology**, v. 9, n. 2, p. 211-223. 2002.

BOFFELLI, D., et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. **Science**, v. 299, p. 1391-1394. 2003.

BRAY, N.; DUBCHAK, I.; PACHTER, L. AVID: A Global Alignment Program. **Genome Research**, v. 13, p. 97-102. 2003.

CHAIN, P. **Examining the current problems of whole genome comparison: a review.** 2001. 26 p. Disponível em: <<http://cmgm.stanford.edu/biochem218/>>. Acesso em: 20 fev. 2001.

CHAPMAN, M. A., et al. Comparative and functional analyses of *LYLI* loci establish marsupial sequences as a model for phylogenetic footprinting. **Genomics**, v. 81, p. 249-259. 2003.



COOPER, G. M., et al. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. **Genome Research**, v. 13, p. 813-820. 2003.

COURONNE, O., et al. Strategies and Tools for Whole-Genome Alignments. **Genome Research**, v. 13, p. 73-80. 2003.

DUBCHAK, I., et al. Active Conservation of Noncoding Sequences Revealed by Three-Way Species Comparisons. **Genome Research**, v. 10, p. 1304-1306. 2000.

DURET, L.; DORKELD, F.; GAUTIER, C. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. **Nucleic Acids Research**, v. 21, n. 10, p. 2315-2322. 1993.

DURET, L.; BUCHER, P. Searching for regulatory elements in human noncoding sequences. **Current Opinion in Structural Biology**, v. 7, p. 399-406. 1997.

ELLSWORTH, R. E., et al. Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. **Proc. Natl. Acad. Sci. USA**, v. 97, n. 3, p. 1172-1177. 2000.

ELZIRIK, E., et al. Molecular dating and biogeography of the early placental mammal radiation. **Journal of Heredity**, v. 92, p. 212-219. 2001.

FRAZER, K. A., et al. Evolutionarily conserved sequences on human chromosome 21. **Genome Research**, v. 11, p. 1651-1659. 2001.

FRAZER, K. A., et al. Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. **Genome Research**, v. 13, p. 1-12. 2003.

GILLIGAN, P.; BRENNER, S.; VENKATESH, B. *Fugu* and human sequence comparison identifies novel human genes and conserved non-coding sequences. **Gene**, v. 294, p. 35-44. 2002.

GÖTTGENS, B., et al. Long-range comparison of human and mouse *SCL* loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. **Genome Research**, v. 11, p. 87-97. 2001.

GRIFFITHS, A. J. F., et al. **Introdução à Genética**. 7. ed. Rio de Janeiro: Guanabara Koogan, 2002. 794 p.

HARDISON, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. **Trends in Genetics**, v. 16, n. 9, p. 369-372. 2000.

HARE, M. P.; PALUMBI, S. R. High intron conservation across three mammalian orders suggests functional constraints. **Molecular Biology and Evolution**, v. 20, n. 6, p. 969-978. 2003.

HERING, T. M., et al. Bovine chondrocyte link protein cDNA sequence: interspecies conservation of primary structure and mRNA untranslated regions. **Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology**, v. 112, n. 2, p. 197-203. 1995.

HETZEL, J. Livestock genome research on the march. **Nature Genetics**, v. 4, p. 327-328. 1993.

HIGGINS, D., et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, v. 22, p. 4673-4680. 1994.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921. 2001.

JAREBORG, N.; BIRNEY, E.; DURBIN, R. Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs. **Genome Research**, v. 9, p. 815-824. 1999.

KENT, W. J; ZAHLER, A. M. Conservation, regulation, synten, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. **Genome Research**, v. 10, p. 1115-1125. 2000.

KOONIN, E. V.; MUSHEGIAN, A. R.; BORR, P. Non-orthologous gene displacement. **Trends in Genetics**, v. 12, p. 334-336. 1996.

LARIZZA, A., et al. Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyls and rodent gene pairs. **Computers and Chemistry**, v. 26, p. 479-490. 2002.

LENHARD, B., et al. Identification of conserved regulatory elements by comparative genome analysis. **Journal of Biology**, v. 2, p. 13-13.11. 2003.

LEWIN, B. **Genes VII**. 7. ed. New York: Oxford University Press, 2000. 990 p.

LI, W-H., et al. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. **Proc. Natl. Acad. Sci. USA**, v. 87, p. 6703-6707. 1990.

LI, W-H., et al. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. **Molecular phylogenetics and evolution**, v. 5, p. 182-187. 1996.

LOOTS, G. G., et al. Identification of Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons. **Science**, v. 288, p. 136-140. 2000.

LUDWIG, M. Z. Functional evolution of noncoding DNA. **Current Opinion in Genetics & Development**, v. 12, p. 634-639. 2002.

MAKALOWSKI, W.; ZHANG, J.; BOGUSKI, M. S. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. **Genome Research**, v. 6, p. 8456-857. 1996.



MAKALOWSKI, W.; BOGUSKI, M. S. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. **Proc. Natl. Acad. Sci. USA**, v. 95, p. 9407-9412. 1998.

MAYOR, C., et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. **APPLICATIONS NOTE**, v. 16, n. 11, p. 1046-1047. 2000.

MEISLER, M. H. Evolutionarily Conserved Noncoding DNA in the Human Genome: How Much and What For? **Genome Research**, v. 11, p. 1617-1618. 2001.

MIGNONE, F., et al. Untranslated regions of mRNAs. **Genome Biology**, v. 3, n. 3, p. 0004.1-0004.10. 2002.

NOBREGA, M. A., et al. Scanning human gene deserts for long-range enhancers. **Science**, v. 302, p. 413. 2003.

NOBREGA, M. A.; PENNACCHIO, L. A. Comparative genomic analysis as a tool for biological discovery. **Journal of Physiology**, v. 554.1, p. 31-39. 2004.

O'BRIEN, S. J.; SEUÁNEZ, H. N.; WOMACK, J. E. Mammalian genome organization: An evolutionary overview. **Annual Review of Genetics**, v. 22, p. 323-351. 1988.

O'BRIEN, S. J., et al. The promise of comparative genomics in mammals. **Science**, v. 286, p. 458-481. 1999.

OELTJEN, J. C., et al. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. **Genome Research**, v. 7, p. 315-329. 1997.

OHTA, T. An examination of the generation-time on molecular evolution. **Proc. Natl. Acad. Sci. USA**, v. 90, p. 10676-10680. 1993.



PENNACCHIO, L. A.; RUBIN, E. M. Comparative genomic tools and databases: providing insights into the human genome. **Journal of Clinical Investigation**, v. 111, n. 8, p. 1099-1106. 2003.

PESOLE, G. Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. **Trends in Genetics**, v. 15, n. 9, p. 378. 1999.

PESOLE, G., et al. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. **Nucleic Acids Research**, v. 28, n. 1, p. 193-196. 2000.

PIZZUTI, A., et al. An ATG repeat in the 3'-untranslated region of the human resistin gene is associated with a decreased risk of insulin resistance. **The Journal of Clinical Endocrinology & Metabolism**, v. 87, n. 9, p. 4403-4406. 2002.

PRIMROSE, S. B.; TWYMAN, R. M. **Principles of Genome Analysis and Genomics**. 3. ed. Turin: Blackwell Publishing Company, 2003. 263 p.

SANTINI, S.; BOORE, J. L.; MEYER, A. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. **Genome Research**, v. 13, p. 1111-1122. 2003.

SCHWARTZ, S., et al. PipMaker – A Web Server for Aligning Two Genomic DNA Sequences. **Genome Research**, v. 10, p. 577-586. 2000.

SHABALINA, S. A., et al. Selective constraint in intergenic regions of human and mouse genomes. **Trends in Genetics**, v. 17, n. 7, p. 373-376. 2001.

SIDOW, A. Sequence first. Ask questions later. **Cell**, v. 111, p. 13-16. 2002.

SOLINAS-TOLDO, S.; LENGAUER, C.; FRIES, R. Comparative Genome Map of Human and Cattle. **Genomics**, v. 27, p. 489-496. 1995.



TATUSOVA, T. A.; MADDEN, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. **FEMS Microbiology Letters**, v. 174, p. 247-250. 1999.

THOMAS, J. W.; TOUCHMAN, J. W. Vertebrate genome sequencing: building a backbone for comparative genomics. **Trends in Genetics**, v. 18, n. 2, p. 104-108. 2002.

THOMAS, J. W., et al. Comparative analyses of multi-species sequences from targeted genomic regions. **Nature**, v. 424, p. 788-793. 2003.

URETA-VIDAL, A.; ETTWILLER, L.; BIRNEY, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. **Nature Reviews in Genetics**, v. 4, p. 251-262. 2003.

VUILLAUMIER, S., et al. Cross-species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. **Biochemistry Journal**, v. 327, p. 651-62. 1997.

WEDEMEYER, N., et al. Conservation of the 3'- untranslated region of the *Rab1a* gene in amniote vertebrates: exceptional structure in marsupials and possible role for posttranscriptional regulation. **FEBS Letters**, v. 477, p. 49-54. 2000.

WEI, L., et al. Comparative genomics approaches to study organism similarities and differences. **Journal of Biomedical Informatics**, v. 35, p. 142-150. 2002.

WEITZMAN, J. B. Tracking evolution's footprints in the genome. **Journal of Biology**, v. 2, p. 9-9.4. 2003.

WILLIAMS, S. H.; MOUCHEL, N.; HARRIS, A. A comparative genomic analysis of the cow, pig, and human *CFTR* genes identifies potential intronic regulatory elements. **Genomics**, v. 81, p. 628-639. 2003.

WOMACK, J. E. Genetic engineering in agriculture: Animal genetics and development. **Trends in Genetics**, v. 3p. 65-68. 1987.



WU, C-I.; LI, W-H. Evidence for higher rates of nucleotide substitution in rodents than in man. **Proc. Natl. Acad. Sci. USA**, v. 82, p. 1741-1745. 1985.

ZHANG, Z.; GERSTEIN, M. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. **Journal of Biology**, v. 2, p. 11-11.4. 2003.

Homepages da Internet

Banco de dados GenBank. Disponível em:
<<http://www.ncbi.nlm.nih.gov/GenBank>>. Acesso em: 2002, 2003.

Glossário de termos genéticos do NCBI. Disponível em:
<<http://www.ncbi.nlm.nih.gov/Education>>. Acesso em: 2002, 2003.

Programa *Blast 2 Sequences*. Disponível em:
<<http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>>. Acesso em: 2002.

Programa *PipMaker*. Disponível em: <<http://bio.cse.psu.edu>>. Acesso em: 2002.

Programa *GLASS*. Disponível em: <<http://crossspecies.lcs.mit.edu>>. Acesso em: 2002.

Programa *VISTA*. Disponível em: <<http://www-gsd.lbl.gov/vista>>. Acesso em: 2002, 2003.

Programa *ClustalW*. Disponível em: <<http://www.ebi.ac.uk/clustalw>>. Acesso em: 2002.

Programa *DBA*. Disponível em: <<http://www.ebi.ac.uk/Wise2/dbaform.html>>. Acesso em: 2002.



RESUMO

MIZIARA, M. N. Análise comparativa “*in silico*” de seqüências não codificantes de genes ortólogos entre humano e bovino. São José do Rio Preto, 2004. 66p. Dissertação (Mestrado em Genética) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista “Júlio de Mesquita Filho”.

As análises comparativas *in silico* de seqüências de DNA representam uma primeira varredura de um genoma de interesse, e vêm mostrando eficiência na revelação de porções do genoma com relevância biológica. Utilizando-se ferramentas de alinhamento e análise de identidade, foram analisadas seqüências não codificantes de 39 pares de genes ortólogos entre humano e bovino, a fim de identificar padrões de conservação que pudessem indicar regiões candidatas relacionadas à regulação desses genes. Setenta e três seqüências não codificantes conservadas em introns, regiões 3’ e 5’-UTRs foram identificadas em 28 genes. As seqüências encontradas nas regiões 3’-UTRs mostraram mais conservação e foram identificadas em um número três vezes maior que nos introns. A inclusão do genoma de camundongo nas análises possibilitou a identificação de 13 seqüências não codificantes ativamente conservadas ao longo da evolução. A seqüência ativamente conservada com a maior porcentagem de identidade dos nucleotídeos foi a do intron 3 do gene RPL3, com 98,8% de identidade, e a seqüência ativamente conservada com menor porcentagem de identidade dos nucleotídeos foi a do PDE5A, com 74,1% de identidade. Considerando a similaridade entre os genomas humano e bovino, os resultados deste trabalho contribuíram para a seleção de seqüências candidatas à regulação, e podem ser submetidas a testes experimentais envolvendo a regulação dos seus respectivos genes. O conhecimento da regulação permite a manipulação de genes de interesse econômico, no caso do genoma bovino, e o entendimento dos mecanismos envolvidos principalmente em doenças humanas.

Palavras-chave: genes, ortólogos, humano, bovino, seqüências não codificantes, conservação.



ABSTRACT

MIZIARA, M. N. *In silico* comparative analysis of noncoding sequences of orthologous genes between human and bovine. São José do Rio Preto, 2004. 66p. Dissertação (Mestrado em Genética) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista “Júlio de Mesquita Filho”.

In silico comparative analyses of DNA sequences represent the first screening of a target genome, and have been efficient to reveal genome portions related with biological relevance. Using alignment and identity analyses, we analyzed noncoding sequences from 39 orthologous gene pairs between human and bovine, with the goal to identify conservation patterns that could indicate putative regions related to gene regulation. Seventy-three conserved noncoding sequences were shown in introns, 3' and 5'-UTRs from 28 genes. The sequences found in the 3'-UTRs showed more conservation and were identified in a number three times greater than in the introns. The inclusion of mouse genome in the analyses allowed the identification of 13 actively conserved noncoding sequences during evolution. The active conserved sequence with the highest nucleotide identity was of RPL3 intron 3, with 98,8% identity, and the smallest nucleotide identity was of PDE5A, with 74,1% identity. Considering the human and bovine genome similarity, the results of this study provides the selection of putative regulatory sequences that can be experimentally tested to confirm their participation in the regulation of the respective genes. The knowledge of gene regulation allows the manipulation of genes with economic interest from the bovine genome, and the understanding of the mechanisms mainly involved in human diseases.

Key-words: genes, orthologous, human, bovine, noncoding sequences, conservation.



Autorizo a reprodução deste trabalho
São José do Rio Preto, 13/02/2004

Melissa Nunes Miziara

