UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

MARCELO AUGUSTO PORTUGAL MATTIOLI

"Estudo do padrão de utilização de códons em seqüências gênicas de eucalipto visando a maximização da produção de proteínas através da otimização das seqüências de DNA"

Dissertação apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Genética e Biologia Molecular, na área de Genética Vegetal e Melhoramento

Orientador: Prof. Dr. Marcelo Menossi Teixeira

Campinas, 2008

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

M434m

Mattioli, Marcelo Augusto Portugal
Análise da utilização de códons em eucalipto visando a maximização da produção de proteínas através da otimização das seqüências de DNA / Marcelo Augusto Portugal Mattioli. – Campinas, SP: [s.n.], 2008.

Orientador: Marcelo Menossi Teixeira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Eucalipto. 2. Codons preferenciais. 3. Genes sintéticos. 4. Otimização. I. Teixeira, Marcelo Menossi. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

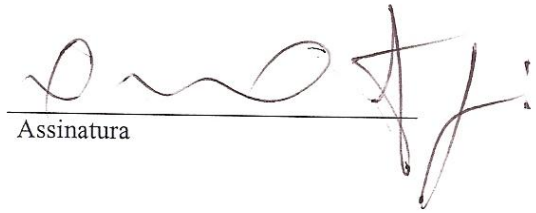
(scs/ib)

Título em inglês: Analysis of codon usage patterns of eucalyptus genomes for DNA optimization and enhanced protein expression.
Palavras-chave em inglês: Eucalyptus; Codon bias; Genes, Synthetic; Optimization.
Área de concentração: Genética Vegetal e Melhoramento.
Titulação: Mestre em Genética e Biologia Molecular.
Banca examinadora: Marcelo Menossi Teixeira, Luis Eduardo Aranha Camargo, Helaine Carrer.
Data da defesa: 09/06/2008.
Programa de Pós-Graduação: Genética e Biologia Molecular.

Data da defesa: 9 de junho de 2008

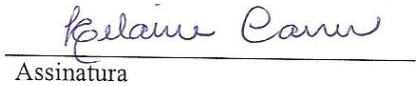
BANCA EXAMINADORA

Prof. Dr. Marcelo Menossi Teixeira (Orientador)



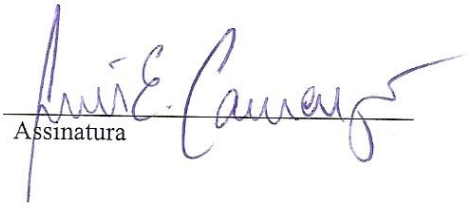
Assinatura

Profa. Dra. Helaine Carrer



Assinatura

Prof. Dr. Luis Eduardo Aranha Camargo



Assinatura

Alice and the Cheshire Cat

“Would you tell me, please, which way I ought to go from here?”

“That depends a good deal on where you want to get to,” said the Cat

“I don’t much care where—” said Alice.

“Then it doesn’t matter which way you go,” said the Cat.

“As long as I get somewhere,” Alice added as an explanation.

“Oh, you’re sure to do that,” said the Cat, “if you only walk long enough”

Lewis Carroll, Alice’s Adventures in Wonderland

DECLARAÇÃO

Declaro para os devidos fins que o conteúdo de minha dissertação/ tese de mestrado/doutorado intitulada Maximização da produção de proteínas em eucalipto através da otimização das seqüências de DNA:

( ) não se enquadra no Artigo 1°, § 3° da Informação CCPG 002/06, referente a bioética e biossegurança.

(x) está inserido no Projeto CIBio (Protocolo n°04/2006), intitulado "Utilização de organismos geneticamente modificados como biorreatores"

( ) tem autorização da Comissão de Ética em Experimentação Animal (Protocolo n°_____).

( ) tem autorização do Comitê de Ética para Pesquisa com Seres Humanos (?) (Protocolo n°_____).

Marcelo Augusto Portugal Mattioli

Marcelo Menossi Teixeira

Para uso da Comissão ou Comitê pertinente:

(X) Deferido ( ) Indeferido

Nome: Marcelo Menossi
Função: Presidente CIBIO/CBMEG

AGRADECIMENTOS

Ao professor Menossi pela oportunidade, incentivo e orientação neste trabalho. E mais do que isso, pelo empenho em evidenciar o imenso campo de atuação que existe além dos limites da universidade e onde um biólogo pode se realizar.

Aos pesquisadores Andrés Yunes, Márcio José, Luis Eduardo Aranha e Helaine Carrer, membros da pré-banca e banca, pelo tempo dedicado ao estudo deste manuscrito e a toda realização deste trabalho.

Aos participantes do projeto FORESTS – Fapesp, Suzano, VCP, Duratex e Ripasa - tanto pelo apoio financeiro quanto por acreditar que as parcerias entre ciência e mercado podem ser profícuas para todos participantes. Agradecimento especial a Esteban Roberto González, pesquisador da Suzano, pela atenção e conselhos ao longo deste trabalho.

Aos amigos do LGF, Renato, Boi, Sandra, Cris, Jú, Ariane, Dona Edna, Dudu, Wilson, Grazi, Lúcia, Jaú, Rafa, Vanúzia, Renata, Agustina, Eduardo, por toda ajuda e tempo dedicado em explicações, discussões, trocas de aprendizado e até mesmo nos trabalhos braçais!

A todo pessoal do CBMEG, em especial ao Professor Michel, Amanda, Renato, Mario e Sílvia.

Aos novos amigos do Instituto Inovação, Bruno, Lívia, Guilherme, Denise, Daniel e Eduardo pelo apoio nesta reta final.

Aos amigos e colegas desta jornada de Unicamp, impossível citar todos os nomes, impossível mesmo, seriam mais páginas de agradecimento de que de tese.

A minha família, Dê, Mãe, Rodrigo, Mari e agora Vê e Alice, pelo apoio infinito em todos os momentos, pela confiança, suporte e incentivo.

Finalmente a Deus, ao acaso, sorte, ou o nome que cada um prefere dar à essas forças que permitiram que todas essas pessoas cruzassem meu caminho e que agora, de uma forma ou de outra, façam parte da minha vida. Seja por serem presenças constantes, seja pelos pedacinhos que roubei delas e agora fazem parte de mim.

ÍNDICE GERAL

ÍNDICE DE ILUSTRAÇÕES E TABELAS

1. RESUMO

O código genético é degenerado, o que permite que um mesmo aminoácido seja codificado por trincas de nucleotídeos distintas. Trincas de um mesmo aminoácido não ocorrem com a mesma freqüência, sugerindo que existe utilização preferencial de codons (*codon usage bias*). Diferentes organismos geralmente apresentam *codon usage bias* distintos e estas diferenças podem ser fatores limitantes na expressão heteróloga. Além do *codon usage bias* existem outros fatores presentes em um gene que podem influenciar a intensidade com que este é traduzido, dentre eles destacam-se: a constituição das regiões gênicas próximas ao códon de iniciação (contexto de iniciação) e a região terminadora da tradução (contexto de terminação).

Nosso objetivo foi identificar a presença destes padrões em seqüências ESTs de eucalipto e gerar informações que permitam aumentar os níveis de produção de proteínas de interesse em eucalipto geneticamente modificado.

Em concreto, foram alcançados os seguintes objetivos:

- Identificados os códons mais utilizados em eucalipto;
- Avaliada as variações na utilização do códon entre genes mais expressos e menos expressos e também entre genes expressos em diferentes tecidos;
- Identificados os padrões presentes nas regiões iniciadora e terminadora da tradução;
- Construído um software que indica as alterações necessárias na seqüência de DNA de qualquer gene que se tenha interesse em expressar em eucalipto;
- Desenhados e construídos genes sintéticos para futura validação *in vivo* da metodologia de otimização da tradução de proteínas de eucalipto desenvolvida e implementada no software.

2. ABSTRACT

Most of the amino acids are represented by more than one codon because of that the genetic code is said to be degenerated. Codons from the same amino acid does not occur with the same frequency, suggesting that there is preferential use of codons (codon usage bias). Different organisms often have different codon usage bias and these differences can be limiting factors in heterologous expression. Besides the codon usage bias there are other factors present in a gene that may influence the intensity with which this is translated, among them are: the DNA bases surrounding the initiation codon (initiation context) and the stop codon (termination context). Our goal is to identify the presence of these patterns - codon usage bias, initiation and termination contexts - in eucalyptus EST sets and generate information to increase protein expression in genetically modified eucalyptus. Specifically, the following objectives have been achieved:

- Identified the set of codons most used in the eucalyptus;
- Verified changes in the use of codons between genes expressed at different levels and also between genes expressed in different tissues;
- Identified patterns surrounding the translation initiation and stop sites;
- Development of a software that indicates the necessary changes in the DNA sequence of any gene to be expressed in eucalyptus
- Designed and constructed synthetic genes for further validation *in vivo* of the optimization methodology developed and implemented in the software.

3. INTRODUÇÃO

3.1. A Síntese Protéica

Francis Crick enunciou o Dogma Central da Biologia Molecular pela primeira vez em 1958 (Crick, 1958) e posteriormente, em 1970, o publicou na Revista Nature (Crick, 1970). O Dogma Central afirma que a especificidade de um fragmento de DNA depende apenas da seqüência de seus nucleotídeos[1] e que essa seqüência é a chave para a disposição dos aminoácidos em uma proteína particular. O dogma central pode ser resumido a uma visão esquemática (Figura 1) que permite compreender de que forma a informação presente em biopolímeros – DNA, RNA e proteínas -, é transferida através das gerações celulares e como essa informação é utilizada por todos organismos vivos no desempenho de suas funções.

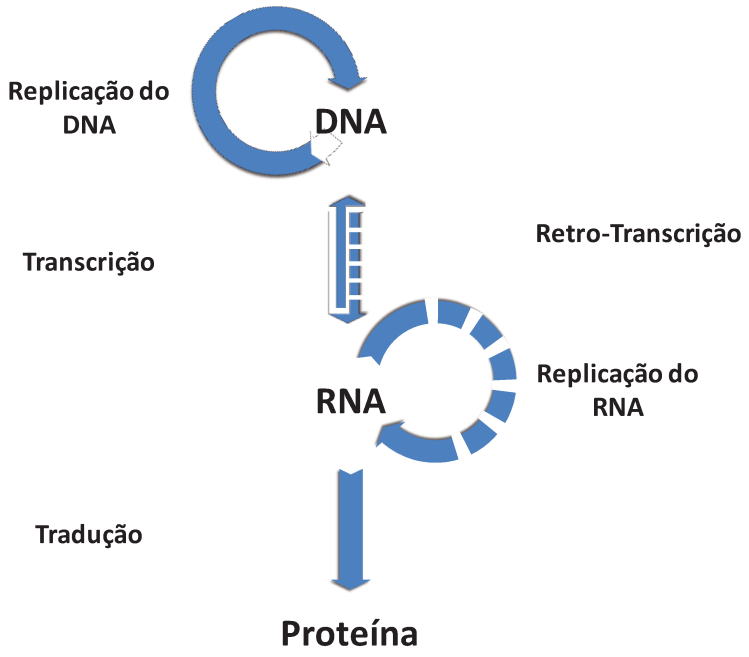


Figura 1. O Dogma Central da Biologia Molecular (adaptado de: Crick, 1970)

[1] A unidade básica que forma as moléculas de DNA e RNA, composta por uma base nitrogenada, uma pentose e um grupo fosfato. Ocorrem 4 tipos de bases nas moléculas de DNA: adenina (A), guanina (G), timina (T) e citosina (C). Na molécula de RNA ocorre Uracila (U) no lugar da timina.

Quando se reproduzem, os organismos vivos transmitem seu material genético para sua progênie. Uma célula individual é a menor unidade reprodutiva possível e é o veículo de transmissão da informação genética em todas as espécies de seres vivos.

As células vivas armazenam a informação genética da mesma forma, através da dupla fita de DNA. A célula replica sua informação individualizando as fitas de DNA. Cada fita individualizada é utilizada como modelo para a polimerização de uma nova fita, com uma seqüência de nucleotídeos complementar[2] a fita que serviu como modelo. Esse processo é a replicação do DNA. Este mesmo sistema de polimerização a partir de uma fita modelo é utilizado na transmissão da informação presente no DNA para moléculas de RNA, em um processo chamado de transcrição. Estas moléculas atuam em conjunto com a maquinaria de tradução[3] e guiam a síntese de proteínas. Este processo conta com a participação dos ribossomos, que são formados por RNA ribossomal e proteínas (Alberts et al. 2004). As proteínas são os principais catalisadores e efetuadores das reações químicas que acontecem nos organismos, além disso tem muitas outras funções, tais como:

- Estrutural ou plástica - as proteínas participam da formação de tecidos, dando as propriedades de rigidez, consistência e elasticidade necessárias a cada diferente tecido do organismo (Alberts et al. 2004).
- Hormonal – muitas proteínas são capazes de desencadear respostas específicas sobre células, órgãos ou estruturas de um organismo como, por exemplo, as citocinas que promovem um retardamento na senescência foliar em plantas (Oliveira et al. 2007)
- Defesa - proteínas podem estar envolvidas em diversos mecanismos de defesa dos organismos. Proteínas produzidas pelas plantas conhecidas

[2] O termo complementar refere-se a seqüências de DNA ou RNA em que cada base consecutiva, a partir da extremidade 5' da primeira fita, forma pontes de hidrogênio com as bases da segunda fita a partir da sua extremidade 3', de acordo com as afinidades de pareamento, ou seja, A com T ou A com U, e C com G.

[3] Todas os componentes celulares envolvidos no processo se síntese protéica: Ribossomos, RNA transportadores (tRNA), enzimas e ATP.

como ureases, por exemplo, apresentam atividade inseticida contra caruncho, percevejos e pulgões (Mulinari et al. 2007)

- Energética – Algumas proteínas podem servir como fonte de reserva energética. As proteínas de reserva correspondem a cerca de 5 a 8% do peso seco de grãos de cereais. (Bernades et al. 1998)
- Enzimática – A enzima RuBisCO (ribulose-bisfosfato carboxilase oxigenase, EC 4.1.1.39), capaz de se ligar tanto a oxigênio quanto gás carbônico, é responsável pela fixação do carbono atmosférico em moléculas orgânicas. A RuBisCo é a enzima mais abundante nas plantas e conseqüentemente a proteína mais abundante no planeta.(Alberts et al. 2004)

Por mais divergente e específica que seja a função de cada proteína, sua estrutura básica é sempre a mesma: um conjunto de aminoácidos. A especificidade de sua função, no entanto, depende da seqüência, da forma, como os aminoácidos são arranjados, que por sua vez é determinada pela seqüência de DNA correspondente ao gene que codifica esta proteína. (Crick, 1970)

Este processo que começa com a leitura de DNA e termina com sua decodificação em proteínas específicas é denominado de expressão gênica. O controle da expressão gênica pode acontecer em diversos pontos do processo e em diferentes compartimentos celulares (figura 2).

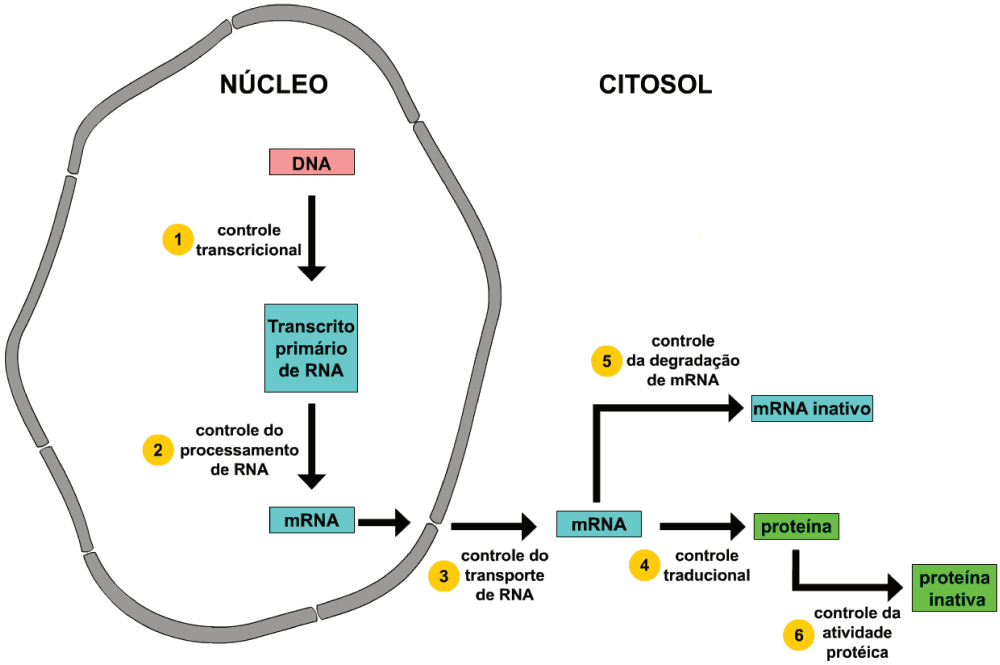


Figura 2. Etapas da expressão gênica (adaptado de: Alberts et at. 2004)

A proposta original do dogma central foi ampliada com a descoberta, em 1970, da enzima transcriptase reversa (Baltimore, 1992) capaz de sintetizar DNA utilizando como molde uma fita de RNA. Em 1963, Haruna e colaboradores haviam demonstrado que o RNA também podia servir de molde para a síntese de outras moléculas de RNA, processo realizado pela enzima replicase codificada por um vírus infeccioso cuja informação genética está contida numa molécula simples de RNA. Estes estudos permitiram que o Dogma Central se ampliasse sem, contudo, perder a unidirecionalidade, ou seja, do DNA à proteína.

O nível de expressão protéica – a quantidade de proteína que é produzida em uma célula - é afetado por muitos fatores que agem sobre diferentes estágios da expressão gênica, incluindo, principalmente, a transcrição e a tradução. A tradução envolve três passos principais e sua eficiência pode ser influenciada em cada um destes passos (Lithwick e Margalit 2003). As três etapas da tradução são:

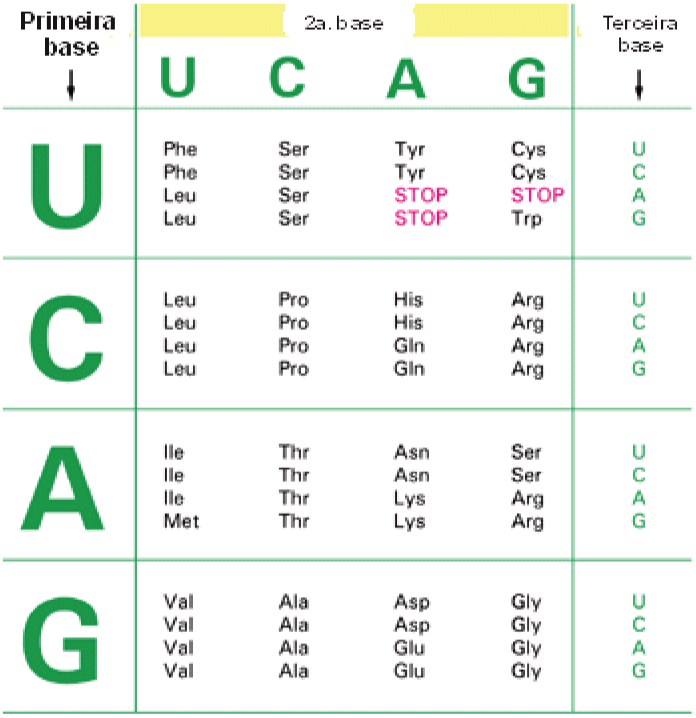
- Iniciação - A subunidade menor do ribossomo liga-se à extremidade 5' do RNA mensageiro (mRNA), e desliza ao longo da molécula até

encontrar o códon de iniciação (AUG), onde ocorre a aproximação do tRNA transportador do aminoácido metionina, que se liga ao códon de iniciação por complementaridade. Em seguida a subunidade maior liga-se à subunidade menor do ribossomo. Também participam desta etapa outras moléculas, dentre elas diversos fatores de transcrição.

- Alongamento – Neste etapa um 2º tRNA transporta o aminoácido correspondente ao segundo códon da seqüência gênica. A maquinaria ribossomal, através de um processo que consome energia, realiza a ligação entre a metionina inicial e o segundo aminoácido. Ligações entre aminoácidos são chamadas de ligações peptídicas. O complexo ribossomal avança três bases ao longo do mRNA no sentido 5' -> 3', o tRNA correspondente a metionina é liberado e o sítio antes ocupado por ele passa a ser ocupado pelo tRNA correspondente ao segundo aminoácido da seqüência. O tRNA correspondente ao terceiro aminoácido ocupa o segundo sítio do complexo ribossomal e é realizada outra ligação peptídica. Esse processo vai se repetindo até a finalização da síntese protéica.
- Finalização: O término da síntese é desencadeado pela presença de um códon de finalização (UAA, UAG ou UGA) na seqüência gênica. Quando o ribossomo chega ao códon de terminação as subunidades do ribossomo separam-se e o peptídeo é libertado.

O código genético usa 61 trincas de nucleotídeos (códons) para codificar 20 aminoácidos e três códons para sinalizar a terminação, fazendo com que um mesmo aminoácido possa ser codificado por códons sinônimos, isto é: trincas diferentes de nucleotídeos que correspondem ao mesmo aminoácido (Figura 3). Esses códons são "lidos" no ribossomo e pareiam com tRNAs complementares que "carregam" o aminoácido apropriado, como explicado anteriormente. A degeneração do código genético permite que seqüências de ácidos nucléicos alternativas codifiquem o mesmo peptídeo. As freqüências com que códons diferentes são utilizados variam significativamente entre organismos, entre proteínas expressas em níveis altos ou baixos dentro do

mesmo organismo, e às vezes até mesmo dentro do mesmo operon (Gouy e Gautier, 1982). Este fenômeno é conhecido como *codon usage bias* ou utilização preferencial de códons.



Primeira base	2a. base U	2a. base C	2a. base A	2a. base G	Terceira base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figura 3. Possíveis trincas codificantes para cada aminoácido

A taxa de alongamento – a velocidade com que a seqüência de peptídeos é sintetizada - durante a tradução parece ser influenciada pela utilização preferencial de códons. Esta suposição está baseada em dois resultados:

- Os níveis de espécies diferentes de tRNA intracelular estão correlacionados com a abundância dos códons correspondentes (Ikemura 1985).
- Genes altamente expressos demonstraram utilizar preferencialmente códons mais freqüentes. (Kanaya et al. 1999).

Assim, é esperado que genes altamente expressos utilizem códons que têm níveis mais altos de tRNAs correspondente, aumentando a taxa de progressão da tradução (Lithwick e Margalit 2003).

Lithwick e Margalit (2003) analisaram três caracteres envolvidos na tradução em procariotos: o potencial de pareamento entre rRNA e a seqüência de Shine-Delgarno presente no mRNA, identidade do códon de parada e codon usage bias, sendo este último fator o que mostrou maior correlação com níveis de expressão protéica.

A incompatibilidade na utilização preferencial de códons pode prejudicar a expressão de genes heterólogos. Uma solução plausível é ampliar o "pool" de tRNA intracelular do hospedeiro. Isso pode ser feito super expressando genes de tRNAs raros. Embora a super-expressão de tRNA apareça inicialmente como uma solução atraente, há alguns empecilhos. Diferentes tRNAs precisariam ser super-expressados para genes de organismos diferentes e essa estratégia não é tão atrativa para hospedeiros que sejam mais difíceis de serem manipulados do que *E. coli* (Kane 1995). Uma alternativa é modificar o gene a ser expressado (Gustafsson et al. 2004).

Em geral, quanto maior o número de códons pouco freqüentes que um gene contém, maiores são as chances da expressão protéica atingir níveis insatisfatórios (Kane 1995). Ademais, níveis baixos de expressão são intensificados se os códons raros ocorrem em grupos ou na região N-terminal da proteína. Uma estratégia comum para aumentar a expressão é alterar os códons do gene alvo de forma que eles se assemelhem mais ao padrão de utilização de códons do hospedeiro. Para este objetivo são utilizadas técnicas que vão desde mutações sitio direcionadas (Kink et al. 1991) até a síntese de genes completos (Nambiar et al. 1984).

Com o conhecimento do padrão de uso de códon de um organismo é possível realizar a otimização dos códons do transgene que se deseja expressar no organismo receptor. Esse ajuste ocorre por uma substituição de um códon de baixa freqüência por outro de alta freqüência, conforme a distribuição de freqüência da espécie desejada. Esse procedimento não altera a seqüência da proteína e permite maximizar a sua síntese, pois aumenta a eficiência da tradução do seu mRNA. A utilidade dessa estratégia é amplamente conhecida, como no caso da humanização da proteína do gene GFP (Zolotukhin et al.

1996); vacinas plasmidiais (Uchijima et al. 1998), e no uso do gene *cry* que produz cristais da toxina Bt do *Bacillus thuringiensis* (Loguercio et al. 2002).

Existe muita especulação relativa às forças evolutivas que produziram estas diferenças na preferência por determinados códons (Grosjean e Fiers 1982). A distribuição de códons responde ao conteúdo de GC genômico e as mudanças na utilização de códons são explicadas, pelo menos em parte, por um equilíbrio mutação-seleção entre os códons sinônimos (Knight et al. 2001). Alguns pesquisadores hipotetizaram que a utilização preferencial de códons tende a diminuir a diversidade de tRNAs isoaceptores, reduzindo a carga metabólica e sendo então benéfica a organismos que gastam parte de suas vidas sobre condições de crescimento rápido (Andersson e Kurland 1991). Quaisquer que sejam as razões para existência de codon usage bias, está claro que este uso preferencial pode ter um impacto profundo na expressão de proteínas heterólogas (Kane 1995).

Tabela 1. Exemplos de otimização e resultados obtidos

(Gustafsson,2004)

Gene origin	Protein name	Host	Improvement
Homo sapiens	IL-2	*Escherichia coli*	16-fold
Clostridium tetani	Fragment C	*E. coli*	4-fold
Bacillus thuringiensis	CrylA(b), CrylA(c)	*Lycopersicon esculentum*	100-fold
B. thuringiensis	CrylA(b), CrylA(c)	*Nicotiana tabacum*	Below detection versus >0.1% of total protein
Mus musculus	Ig κ chain	*Saccharomyces cerevisiae*	>50-fold
Bacillus hybrid	(1,3-1,4)-β-glucanase	*Hordeum vulgare*	Below detection versus 40 ng per 2×10^5 protoplasts
H. sapiens	TnT	*E. coli*	10- and 40-fold (two different constructs)
HIV	Gp120	*H. sapiens*	>40-fold
Aequorea victoria	GFP	*H. sapiens*	Below detection versus substantial signal
A. victoria	GFP	*H. sapiens*	22-fold
A. victoria	Mutated GFP	*Candida albicans*	Below detection versus strong band in western
M. musculus	c-Fos	*E. coli*	Below detection versus 20% of soluble protein
Spinacia oleracea	Plastocyanin	*E. coli*	1.2-fold
H. sapiens	Neurofibromin	*E. coli*	3-fold
Listeria monocytogenes	LLO	*M. musculus*	100-fold
H. sapiens	M2-2	*E. coli*	140-fold
Rickettsia prowazekii	Tlc	*E. coli*	No effect
BPV1	L1 and L2	Mammalian	>1000-fold
H. sapiens	PC-TP	*E. coli*	Trace levels versus 10% of cytosolic protein
H. sapiens	hCG-β	*Dictyostelium*	4–5 fold
Triticum aestivum	CYP73A17	*S. cerevisiae*	4-, 7- and 13-fold (three different constructs)
T. aestivum	CYP73A17	*N. tabacum*	5-fold
HIV	Gag	*H. sapiens*	>322-fold
Dermatophagoides	ProDer p1	Mammalian	5–10-fold
HIV	Gag	*H. sapiens*	1.5–2-fold
Plasmodium	EBA-175 region II and MSP-1	*M. musculus*	4-fold
Tn10/herpes simplex virus	rtTA	*M. musculus*	>20-fold
HPV	L1	*H. sapiens*	1×10^4–1×10^5-fold
Corynebacterium diphtheriae –mammal hybrid	DT	*Pichia pastoris*	0 versus 10 mg L^{-1}
P1 phage	Cre	Mammalian	1.6-fold
Actina equina	Equistatin	*P. pastoris*	2-fold
H. sapiens	IL-6	*E. coli*	3-fold
H. sapiens	Glucocerebrosidase	*P. pastoris*	8- and 10-fold (two different constructs)
Schistosoma mansoni	SmGPCR	*H. sapiens*	Barely detectable versus strong band in western
Caenorhabditis elegans	GluClα1, GluClβ	*Rattus norvegicus*	6–9-fold
Herpesvirus	U51	Mammalian	10–100-fold
HIV	Gag, Pol, Env, Nef	*H. sapiens*	>250×, >250×, >45×, >20×, respectively
H. sapiens	IL-18	*E. coli*	5-fold
HPV	E5	Mammalian	6–9-fold
HPV	E7	Mammalian	20–100-fold
Plasmodium	F2 domain of EBA175	*E. coli, P. pastoris*	4-fold and 9-fold

Embora o *codon usage bias* de um gene possua um papel grande em sua expressão, seria incorreto sugerir que este é o único fator envolvido. A escolha de vetores de expressão e de promotores também é importante (Higgins e Hames 1999). As seqüências de nucleotídeos que cercam a região N-terminal da proteína parecem particularmente sensíveis, tanto para a presença de códons raros (Hoekema et al. 1987, Deana et al. 1998) quanto para a identidade dos códons imediatamente adjacentes ao AUG iniciador (Stenstrom

e Isaksson 2002). Também existem certas interações entre nível de tradução e estabilidade de mRNA que não foram completamente compreendidas (Wu et al. 2004), sendo que a estrutura da região 5` do mRNA também tem um efeito significativo (Griswold et al. 2003).

Analises estatísticas dos nucleotídeos presentes ao redor do AUG, códon de início da tradução, em diferentes organismos mostraram a preferência por certos nucleotídeos em determinadas posições (Lukaszewicz et al. 2000). Em eucariotos, a tradução por ribossomos citosólicos geralmente começa a partir do primeiro AUG transcrito (Kozak 1991).

Entretanto o reconhecimento de um AUG como ponto de início da tradução depende de diversos fatores, tais como: distância da extremidade 5`, a estrutura secundária ao redor do AUG e a seqüência de nucleotídeos que flanqueia o sitio de iniciação, chamada de contexto de iniciação (Boyd e Thummel 1993, Kozak 1991, Roy et al. 1990).

A composição da região vizinha ao AUG iniciador de genes disponíveis nos bancos de dados públicos tem sido extensivamente analisada e a análise de 5074 seqüências de plantas permitiu a identificação de um consenso para monocotiledôneas (a:c A:G A:C c AUG G C) e dicotiledôneas (a A A:C a AUG G C) onde as bases -3 e +4 são as mais importantes, considerando o A do códon AUG como posição +1 (Joshi et al. 1997). Estes consensos foram determinados através da regra 50/75 descrita por Cavener (1987). Uma base singular recebe o status de consenso e é escrita com letra maiúscula caso a freqüência relativa de um único nucleotídeo em uma determinada posição seja maior que 50% e duas vezes maior que a freqüência do segundo nucleotídeo mais freqüente. Quando apenas um nucleotídeo não preenche estes critérios, um par de bases pode ser determinado como consenso caso a soma das freqüências relativas destes dois nucleotídeos na mesma posição seja maior que 75%. Quando nenhum destes critérios é estabelecido em uma posição, é indicado o nucleotídeo mais freqüente em letras minúsculas.

Estudos *in vitro* demonstraram uma menor importância do contexto de iniciação no processo traducional de plantas do que em outros organismos

(Lutcke et al. 1987). Esse fato diverge dos dados de ensaios em vivo onde é demonstrada a importância das posições -3, -2, +4 e +5 (Lukaszewicz et al. 2000).

A observação de que contextos de iniciação sub-ótimos estão presentes em muitos genes sugere a hipótese que este contexto pode estar envolvido com a modulação da expressão gênica. Esse pode ser o caso para transcritos que codificam duas proteínas que diferem em suas regiões N-terminal. Alguns transcritos codificam proteínas que podem operar tanto no citosol quanto em alguma organela. O peptídeo maior, iniciado no primeiro AUG em um contexto sub ótimo, inclui a seqüência direcionadora, já o menor, que foi sintetizado a partir de um outro AUG permanece no citosol (Small et al. 1998).

A terminação da síntese protéica é desencadeada por um dos três códons de terminação (UAA, UAG, e UGA) e resulta na liberação do polipeptídio recém sintetizado do ribossomo. Contudo, existem casos nos quais o trinucleotídeo (códon de parada) não é suficiente para terminar a tradução eficientemente. (Tate et al. 1996). Os códons que precedem o códon de parada, a identidade do códon de parada e a base seguinte a ele têm influência na eficiência da terminação da tradução (Poole et al. 1995). Foi mostrado que códons de parada ineficientes conduzem a uma baixa produção protéica (Jin et al. 2002).

Devido ao nucleotídeo seguinte ao códon de parada ter se mostrado importante para a eficiência da terminação, os códons de terminação são freqüentemente considerados como compostos por quarto bases (Poole et al. 1995). Foi mostrado que o códon de terminação UAAU é o mais eficiente em *E. coli* (Poole et al. 1995). Lithwick and Margalit (2003) demonstraram que certamente existe uma preferência significante por UAAU em genes altamente expressos em *E. coli* e em *H. influenzae*.

São significativamente suprimidos padrões de dinucleotídeos CpG imediatamente após o códon de parada (Jacobs et al. 2000). Já a região anterior também parece apresentar alguns padrões, ainda que esta região, por ser codificante, deva ser influenciada por outros fatores tais como: *codon usage bias*, composição de aminoácidos ou estrutura protéica. Por exemplo,

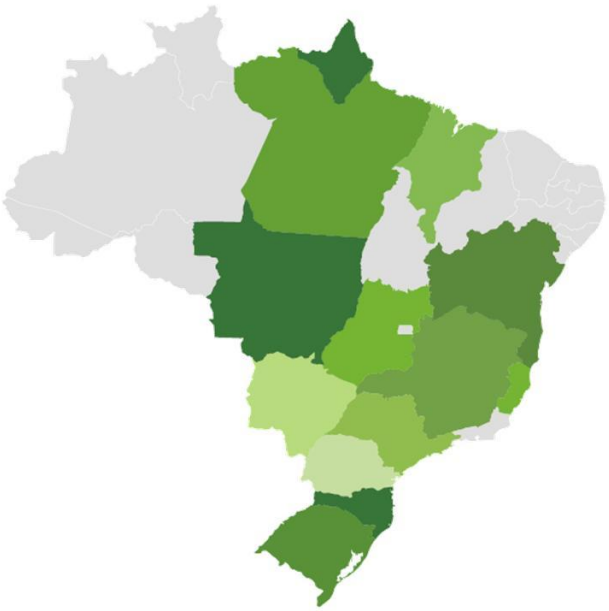
GpG dinucleotídeos são freqüentemente observados na região imediatamente anterior ao códon de terminação (Cavener e Ray, 1991), e tem sido reportado que apenas o último aminoácido tem influencia sobre a eficiência do processo de terminação (Mottagui-Tabar and Isaksson 1997).

O "Codon Adaptation Index" -CAI- relaciona a freqüência de códons em uma proteína com a freqüência destes códons em um grupo de referência de proteínas altamente expressas. Convencionalmente, proteínas ribossomais são usadas como grupo de referência de proteínas altamente expressas. Os valores de CAI variam entre 0 e 1, e quanto mais alto o valor o mais semelhante é o uso de códons no peptídeo particular em relação ao padrão de utilização que ocorre em genes altamente expressos (Sharp e Li 1987). Ozawa et al. (2002) observaram em eucariotos uma correlação positiva entre valores de CAI e o conteúdo de informação nas posições +1 e +2 depois do códon de parada. Como genes com valores de CAI elevado são geralmente mais expressos, os resultados de Ozawa et al. (2002) sugerem que genes expressos em níveis elevados tenham um consenso favorável ao redor do códon terminador.

3.2. A Cultura do Eucalipto no Brasil

O Eucalipto tem sua origem na Austrália e em outras ilhas da Oceania e chegou a Brasil na segunda metade do século XIX, quando foi introduzido com fonte de madeira que seria aplicada como dormentes nas linhas férreas que se instalavam no País (Ministério da Ciência e Tecnologia, 2003)

Atualmente o Brasil é o maior produtor mundial de celulose -6,3 milhões de toneladas por ano- e possui a maior área plantada de eucaliptos do mundo, mais de 3 milhões de hectares distribuídos do norte ao sul do país. (Figura 4).



Estado	Área Plantada (ha)
Minas Gerais	1.067.744
São Paulo	798.522
Bahia	527.386
Espírito Santo	204.035
Rio Grande de Sul	179,690
Paraná	114.996
Mato Grosso do Sul	113.432
Pará	106.033
Santa Catarina	61.166
Maranhão	60.745
Amapá	60.087
Goiás	47.542
Mato Grosso	42.417
Outros	27.409
TOTAL	3.407.205

Figura 4. Área e distribuição de florestas plantadas com eucalipto no Brasil entre 2005 e 2006 (Abraf, 2007)

Além da quantidade também se destaca a qualidade da cultura do eucalipto no Brasil, que apresenta o maior índice médio de produtividade do mundo, sendo de 40m^3 por hectare ao ano (Ministério da Ciência e Tecnologia, 2003).

Segundo relatório publicado pela Sociedade Brasileira de Silvicultura (ABRAF, 2007) a produção do eucalipto brasileiro se destina basicamente à produção de celulose e papel e ao carvão que abastece as siderúrgicas (Figura 5). As indústrias brasileiras que usam o eucalipto como matéria prima para a produção de papel, celulose e demais derivados representam 4% do nosso PIB, 8% das exportações e geram aproximadamente 150 mil empregos (Ministério da Ciência e Tecnologia, 2003).

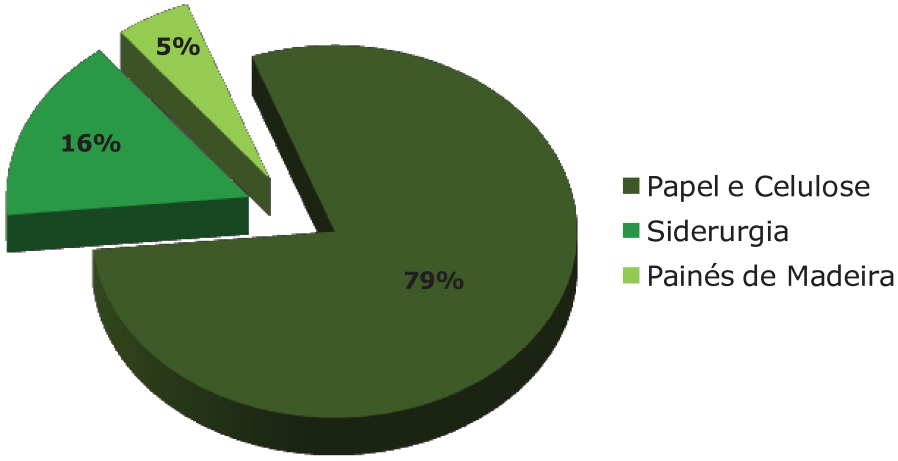


Figura 5. Área com florestas plantadas por segmento industrial (Abraf, 2007)

A produção brasileira de celulose vem crescendo a uma taxa de 6,4% ao ano desde 1997. No ranking mundial, o Brasil ocupa a 7ª posição, porém como fabricante de celulose de fibra curta, o país é o principal produtor mundial. Conforme o tipo de árvore se obtém a celulose de fibra curta ou de fibra longa. Esta característica torna o papel resultante mais absorvente ou mais resistente, respectivamente. Todos esses números demonstram a importância do eucalipto para a economia do País e se refletem na alta competitividade em um mercado altamente disputado (Abraf, 2007).

A produtividade média dos plantios de eucalipto que em 1990 era de aproximadamente 26m^3/ha.ano hoje é de aproximadamente 41m^3/ha.ano. Este crescimento permite afirmar que esta melhora na produção é resultado dos investimentos em pesquisa e desenvolvimento aplicados na silvicultura. Fato que faz com que a grande maioria das florestas de eucalipto plantadas hoje serem oriundas de plantios clonais de alta produtividade, adaptados e tolerantes a fatores adversos de clima, solo e água, entre outros (Abraf 2007). A maior parte destas variedades foi desenvolvida através de metodologias de melhoramento genético clássico, no entanto, cada vez mais é maior o interesse e o investimento em biotecnologia feitos pelo governo e pelas grandes empresas do setor. Duas iniciativas corroboram esta afirmação: o projeto Forests e o Projeto Genolyptus.

O projeto FORESTs, é uma iniciativa de 4 empresas e da FAPESP. O projeto conta com duas etapas: na primeira, foi feito o seqüenciamento do genoma expresso do eucalipto (ESTs), na segunda, foram realizadas correlações entre os genes seqüenciados às funções da planta. A Fapesp investiu 530 mil dólares na primeira fase, e as quatro empresas participantes (Votorantim, Suzano, Duratex e Ripasa), cerca de 250 mil reais. Os dados resultantes da pesquisa são compartilhados pelas empresas. O Genolyptus - do qual participam universidades, empresas e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) - tem abordagem diferente. São observadas e cruzadas árvores com diferentes características. Com o uso de mapeamento genético, seqüenciamento de DNA e mapeamento físico do genoma do eucalipto, serão identificadas as regiões do genoma que controlam funções ligadas principalmente a qualidade da madeira e resistência a doenças.

3.3. Objetivos do Projeto

O objetivo primordial deste trabalho foi fornecer informações que permitam aumentar os níveis de produção de proteínas em eucalipto geneticamente modificado através do aperfeiçoamento do processo de tradução.

Para isto identificou-se a presença de padrões na utilização de códons em seqüências ESTs de eucalipto e gerou informações que permitem aumentar os níveis de produção de proteínas de interesse em eucalipto geneticamente modificado. Em concreto, o presente projeto visou:

- Identificar códons mais utilizados em eucalipto
- Identificar variações na utilização do códon entre genes mais expressos e menos expressos e também entre genes expressos em diferentes tecidos
- Identificar padrões presentes nas regiões iniciadora e terminadora da tradução
- Produzir um software que indique as alterações necessárias na seqüência de DNA de qualquer gene de interesse

A fim de validar os dados levantados e o software de otimização foram construídos genes sintéticos e está em execução o processo para validar com ensaios *in vivo* a metodologia de otimização utilizada pelo software

4. MATERIAIS E MÉTODOS

4.1. Identificação das ORFs e Elaboração do Conjunto de Dados

A primeira etapa do projeto foi a identificação das regiões codificantes de proteínas (ORFs) a partir das seqüências depositadas no banco de dados do projeto FORESTs. Para isso foi elaborada com a linguagem de programação *Perl* uma ferramenta chamada ORFIS (Open Reading Frame Indentifier Script). Ela é capaz de comparar por similaridade os dados do banco FORESTs com dados de bancos públicos; obter o quadro correto de leitura e procurar pela metionina inicial e pelo códon de terminação, que não fossem identificados diretamente por similaridade. Além disto, esta mesma ferramenta apresenta funções de alinhamento e verificação da freqüência de cada nucleotídeo em diferentes posições da seqüência gênica. As verificações de homologia foram feitas com seqüências depositadas nos bancos NCBI (www.ncbi.nlm.nih.gov) e SWISSPROT (www.ebi.ac.uk/swissprot). Para obter um conjunto de dados mais confiáveis foram utilizadas apenas seqüências "Full Lenght" – aquelas que possuem similaridade tanto na região iniciadora quanto na terminadora.

4.2. Identificação de Variações na Utilização de Códons Sinônimos

Diversas análises foram realizadas para verificação da utilização de códons sinônimos. Para isso utilizamos os programas CodonW (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html), GCUA (General Códon Usage Analysis ftp://ftp.nhm.ac.uk/pub/gcua) e InCA (Interactive Condon Analysis http://www.bioinfo-hr.org/inca), desenvolvidos por outros grupos de pesquisas. Dentre estes três optamos por trabalhar com InCA. Este programa contabiliza a freqüência relativa de cada códon (codon usage), mostra o tamanho de cada ORF, o conteúdo de GC, além de desempenhar outras funções. Os dados foram submetidos a análises executadas com os programas R e Microsoft Excel. Estes estudos foram realizados com auxílio do grupo de bioinformática do Genômica Funcional.

4.3. Distribuição dos Genes em Classes

Após a elaboração do conjunto de dados, os genes selecionados foram agrupados de acordo com classe protéica (obtida por homologia), níveis e locais de expressão (dados FORESTs). Para isto extraímos do grupo de ORFs aquelas que representavam os 10 genes mais expressos, os 100 mais expressos e os 100 menos expressos de acordo com os dados de expressão *in silico* gerados pelo projeto FORESTs. Os locais de expressão também foram determinados por meio das informações de expressão *in silico*, e seguiram o agrupamento de tecido preferencial de expressão apresentado por Vicentini et al. (2005).

4.4. Estudo do Contexto de Iniciação e Terminação da Tradução

As análises do contexto de iniciação e terminação foram feitas através da ferramenta ORFIS que permitiu contabilizar a freqüência de cada base em cada posição e identificar possíveis consensos nestas regiões. Os consensos foram determinados através da regra 50/75 descrita por Cavener (1987). Uma base singular recebe o status de consenso e é escrita com letra maiúscula caso a freqüência relativa de um único nucleotídeo em uma determinada posição seja maior que 50% e duas vezes maior que a freqüência do segundo nucleotídeo mais freqüente. Quando apenas um nucleotídeo não preenche estes critérios, um par de bases pode ser determinado como consenso caso a soma das freqüências relativas destes dois nucleotídeos na mesma posição seja maior que 75%. Quando nenhum destes critérios é estabelecido em uma posição, é indicado o nucleotídeo mais freqüente em letras minúsculas.

4.5. Construção do Software

A partir dos padrões de utilização de códons identificados foi construído o Software LGF-EGO 05 (Eucalipto Gene Optimizer). Esse programa roda *on line* e pode ser acessado através do endereço: https://ipe.cbmeg.unicamp.br/cgi-

bin/pipeCodon/codon.cgi. O usuário pode optar entre três opções de otimização:

- Best Codons: Apenas as trincas mais utilizadas para cada aminoácido são utilizadas;
- Best Group of Codons: As trincas menos freqüentes para cada amionácidos são excluídas e é formado um grupo de códons mais freqüentes. O software seleciona deste grupo, de forma aleatória, a trinca que será utilizada;
- Best Couple of Codons: A relação entre códons vizinhos é utilizada como parâmetro na construção da seqüência gênica.

Como resultado final é exibido: a seqüência gênica otimizada, o número de códons modificados, e os conteúdos de GC e GC3 antes e depois do processo. O programa traduz (monta a seqüência de aminoácidos) a seqüência original e a otimizada e compara as duas cadeias peptídicas para ter certeza que não houve nenhuma troca de aminoácidos.

Um estudo individualizado de cada seqüência gerada é recomendável para se identificar possíveis introns, e mudanças pontuais na região problemática são a solução para este tipo de problema.

O programa foi construído com as linguagens de programação perl, R e html e seu registro foi feito junto ao INPI por intermédio da agência INOVA/UNICAMP (número de registro: 00071769).

4.6. Construção dos Genes Sintéticos

O genes sintéticos foram desenhados com o software LGF-EGO 05 tendo como molde a seqüência do gene mGFP5. Esta é uma versão do gene codificante de proteína verde fluorescente da água viva *Aequorea victoria* adaptada para expressão em plantas pois teve um íntron críptico retirado (Haseloff et al. 1997). Foram desenhadas duas seqüências otimizadas, uma contendo apenas os códons mais freqüentes para cada aminoácido (mGFP5-BC) e outra

utilizando relações entre códons vizinhos (mGFP5-BCC). Os genes sintéticos foram construídos pela empresa alemã GENEART.

4.7. Construção dos Vetores de Transformação

A partir dos plasmídios pCAMBIA 1302 e 2301 foi construído um vetor contendo os genes mGFP5 e nptII. O gene mGFP5 foi retirado do vetor 1302 com as enzimas de restrição HindIII e BstEII e clonado no vetor 2301 digerido pelas mesmas enzimas. Essa construção foi chamadada de pCAMBIA 2302 e será utilizada como controle em experimentos futuros. Ambos os genes sintéticos (mGFP5-BC e mGFP5-BCC) foram inseridos nos vetores 2301 da mesma forma. Primeiro foram retirados dos vetores originais após digestão com as enzimas NcoI e BstEII e inseridos no vetor 1302. Este vetor intermediário foi digerido com HinDIII e BsteII e o inserto foi clonado no vetor 2301 previamente digerido com as mesmas enzimas. Estes vetores foram chamados de 2302-BC e 2302-BCC. Além de PCR, foram feitos ensaios de digestão com as enzimas XhoI, BamHI PvuII e EcoRV para confirmar as construções.

4.8. Transformação de Epitélio de Cebola

Tiras de epiderme de cebola foram bombardeadas utilizando um canhão de hélio. Foi utilizado DNA plasmidial de cada construção (2302, 2302-BC e 2302-BCC) e do vetor pCAMBIA 1302, que foi utilizado como controle para expressão de GFP. Também como controle foi feita co-transformação com o plasmídio POL-RFP. Este plasmídio codifica a proteína RFP (red fluorescent protein) com um sinal que direciona esta proteína para mitocôndrias. Vinte quatro horas após o bombardeamento foi verificada a expressão de GFP e RFP por microscopia de fluorescência.

4.9. Transformação de Eucalipto

As tentativas de transformação de calos foram feitas de acordo com a metodologia apresentada por Sartoretto et al. 2002.

4.9.1. Obtenção de Plântulas

Sementes de eucalipto fornecidas pelo Centro de Energia Nuclear na Agricultura (CENA – USP) foram esterilizadas por imersão em etanol 70% (v\v) por 1 minuto seguido de 10 minutos em solução diluída de hipoclorito de sódio contendo chlorine 8% (w/v). Após esse tratamento foram lavadas quatro vezes em água estéril.

As sementes esterilizadas foram colocadas em meio básico MS [meia força MS macronutrientes, completo micronutrientes], suplementado com vitamina MS-B filtro esterilizada , sucrose 3% (w/v) e Agar 0,6% (w/v), pH=5,8 ajustado com 0,1 M KOH antes de autoclavagem por 20 minutos a 121°C.

As placas contendo as sementes foram mantidas sob temperatura de 26°C, sob fotoperíodo de 16h e intensidade luminosa de 40 mmoles fótons/m/s.

4.9.2. Indução dos Calos

Foram retirados cotilédones e hipocotilédones de plantas com pelo menos 10 dias de cultivo. Estes foram colocados em placas de petri previamente preparadas com meio MSM - 10µM 2,4D/ 2,5µM BAP. As placas foram mantidas no escuro. Foram feitos repiques a cada catorze dias.

Além dos calos obtidos de acordo com esta metodologia também foram feitas tentativas com calos fornecidos pelas empresas Suzano e VCP.

4.9.3. Transformação dos Calos

A metodologia de bombardeamento foi semelhante à utilizada para transformação de epitélio de cebola. Foi utilizado DNA plasmidial de cada construção (2302, 2302-BC e 2302-BCC) e do vetor pCAMBIA 1302, que foi utilizado como controle para expressão de GFP. Também como controle foi feita co-transformação com o plasmídio pCAMBIA 2301. Este plasmídio contem o gene repórter gusA e o gene de seleção nptII.

Para realização do teste histoquímico GUS, calos bombardeados foram incubados em solução X-GLUC durante 12h, no escuro, a 37 ^{o}C. A avaliação do teste foi realizada com auxílio do microscópio estereoscópico, observando-se a presença de coloração azulada no tecido analisado.

5. RESULTADOS E DISCUSSÃO

5.1. Identificação das Seqüências Codificantes e Elaboração de um Conjunto de Dados de Referência

Os genes contêm seqüências de nucleotídeos que são compostas por regiões reguladoras e uma região codificadora chamada de CDS. O CDS é a seqüência de nucleotídeos que corresponde aos aminoácidos na proteína predita. A busca pela CDS em uma seqüência desconhecida é feita por meio da verificação das possíveis ORFs (*Open Reading Frames*). O termo ORF indica uma porção de DNA com potencial para dar origem a um polipeptídio; Para isto uma ORF deve conter um códon de iniciação AUG e terminar com um dos três possíveis códons de terminação.

A primeira etapa deste projeto consistiu na identificação de ORFs e posterior extração de CDS das seqüências ESTs depositadas no banco FORESTS.

Existem diversas metodologias para identificação de ORFs, mas as duas principais são: algoritmos matemáticos capazes de predizer e localizar possíveis genes e ferramentas que "caçam" as ORFs através do rastreamento de todos os quadros de leitura do EST.

Dentre as ferramentas computacionais que tem como objetivo localizar possíveis genes, utilizando métodos preditivos matemáticos e estatísticos,as mais eficientes tem como base algoritmos de Hidden Markov Model (HMM), heurística ou uma combinação desses com outros métodos(mukherjee et al. 2005). Entre os programas mais utilizados estão GLIMMER (Delcher et al. 1999), GeneMark (Besemer et al. 2005), EasyGene (Larsen et al. 2003) e ESTScan (Iseli et al. 1999).

A ferramenta ORF Finder, disponível no site do NCBI, "lê" a seqüência EST em todos os possíveis quadros de leitura abertos, traduzindo a seqüência com a utilização do código genético padrão ou de um alternativo, específico para o organismo em estudo, e apresenta como resultado as regiões compreendidas entre um códon de início e um códon de terminação, fornecendo indícios sobre

o quadro de leitura correto de uma seqüência de DNA e sobre onde as regiões codificantes começam e terminam (Gibas e Jambeck, 2001).

A etapa seguinte à identificação das ORFs na seqüência de DNA é a comparação da seqüência de nucleotídeos e da seqüência de aminoácidos das proteínas codificadas por esses possíveis genes com seqüências depositadas em bancos de dados. Existem vários bancos de dados de acesso livre que agrupam informações que atendem as mais diversas questões biológicas. Os bancos de dados mais completos que armazenam seqüências de DNA e de proteínas são: o GenBank (Wheeler et al. 2006; Benson et al. 2006) (http://www.ncbi.nlm.nih.gov/Genbank/), o DDBJ (Tateno et al. 2002; Okubo et al. 2006), e o EMBL (Birney et al. 2006).

A busca por similaridade tem como principal objetivo identificar seqüências similares cuja função já tenha sido descrita experimentalmente e permitir identificar com maior precisão os CDSs.

Na grande maioria dos trabalhos científicos as seqüências de interesse são comparadas com aquelas depositadas nos bancos de dados, utilizando o programa BLAST (Basic Local Alignment Search Tool). O programa BLAST é uma ferramenta computacional criada em 1990 e muito utilizada devido à sua rapidez na obtenção dos resultados. O BLAST utiliza matrizes de valores (score), como a matriz BLOSUM62, para procurar pelo mais alto valor (grau de score) no alinhamento da seqüência de interesse (query) contra as seqüências de banco de dados (subject) (Altschul et al. 1990)

Uma comodidade da ferramenta Orf Finder é que as seqüências deduzidas podem ser alinhadas com as seqüências do GenBank utilizando a versão on-line do BLAST. Uma desvantagem é que esse processo é manual e deve ser feito para cada uma das seqüências.

Conforme apresentado, existem diversas metodologias disponíveis que permitem a extração de ORFs de dados obtidos através de seqüenciamento de cDNA. Submetemos nossos dados ao software ESTSCAN (Iseli et al., 1999) que utiliza algoritmos matemáticos para identificar seqüências codificantes, além de propor a capacidade de corrigir possíveis erros de seqüenciamento.

Apesar deste programa ser amplamente utilizado os dados que obtivemos com ele em testes preliminares não foram satisfatórios, pois o número de seqüências com erros de quadro de leitura foi muito elevado.

Já o método de identificação Orf Finder mostrou menos erros deste tipo, e demonstrou ser conveniente ao estudo. No entanto ele não permite a análise de grande volume de dados.

Pensando em um método que permitisse identificar as ORFs e realizar a busca de similaridade com seqüências já conhecidas (BLAST) de maneira otimizada e com grande volume de dados foi desenvolvida a ferramenta ORFIS. Que funciona com pressupostos semelhantes á ferramenta Orf Finder.

O programa ORFIS funciona em etapas, em um primeiro momento é feito Blast onde as seqüências ESTs são comparadas com seqüências depositadas em bancos públicos e de acordo com a seqüencia de maior similaridade é definido o quadro de leitura do EST. Na etapa seguinte o programa compara os códons de início e término da seqüência conhecida com a ORF identificada na seqüencia EST e determina o início e o fim do CDS, agora identificado.

Todos as seqüências consenso do banco FORESTs (cerca de 30.000) foram submetidas à ferramenta ORFIS que encontrou 2636 seqüências *full lenght* quando comparou as seqüências depositadas no banco de dados do projeto FORESTs com as seqüências depositadas no NCBI. A comparação com dados depositados no banco SWISSPROT identificou 1527 possíveis seqüências. O padrão de utilização de códons foi verificado para ambos os conjuntos e nenhuma diferença significativa foi observada entre eles. Objetivando um maior espaço amostral as análises posteriores foram feitas sobre o conjunto obtido com auxílio do banco NCBI.

Os dados obtidos com o programa ORFIS se mostraram de boa qualidade e validaram a metodologia que foi desenvolvida para a identificação de CDSs em grandes volumes de dados.

5.2. Identificação de Variações na Utilização de Códons Sinônimos

Conforme apresentado na introdução do presente trabalho o código genético é degenerado, isto é: um mesmo aminoácido pode ser codificado por trincas de nucleotídeos diferentes. E a freqüência como diferentes trincas codificantes de uma mesmo aminoácido – chamadas também de códons sinônimos - são utilizadas não é igual. Por exemplo, para o aminoácido histidina que pode ser codificado por dois códons distintitos – CAU e CAC – esperar-se-ia que cada uma das trincas fosse utilizada em metade das vezes em que o aminoácido ocorre, no entanto não é isso que acontece. Uma trinca é mais utilizada. Esse fenômeno recebe o nome de *Codon Usage Bias* ou utilização preferencial de códons.

Para identificar como os códons são utilizados nas seqüências de eucalipto submetemos os dados referentes aos CDSs extraídos do banco FORESTs ao programa INCA. O que este programa faz é contabilizar quantas vezes determinado aminoácido ocorre e também quantas vezes cada trinca de nucleotídeos ocorre.

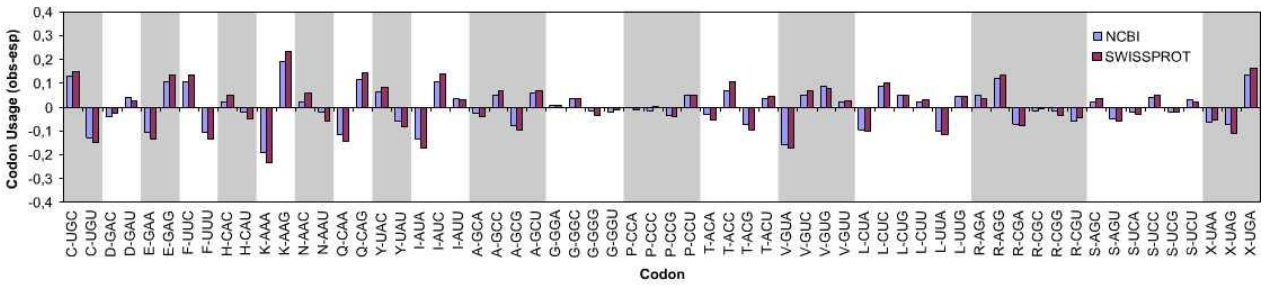


Figura 6. Padrão de *codon usage bias* apresentado pelas ORFs extraidas dos dados FORESTs. O eixo Y refere-se à diferença entre o *codon usage* observado em cada trinca e o valor referente a ocorrência aleatória da mesma; resultados positivos indicam uma utilização superior ao aleatório já valores negativos indicam uma utilização inferior. No eixo X estão representados os aminoácidos e as trincas correspondentes. Este artifício permite uma rápida análise gráfica comparativa do padrão de *codon usage bias* e foi utilizado nas figuras 1,2,3 e 4.

Como pode ser observado na figura 6 existe utilização preferencial de códons em eucalipto. Existem alguns aminoácidos que apresentam maior diferença na

utilização de códons sinônimos que outros. Foram realizados testes de qui-quadrado que mostraram que para todos aminoácidos a utilização das trincas não é aleatória e que a diferença na utilização é significativa.

É interessante notar o fato de que na maioria das vezes as trincas que ocorrem com maior freqüência são aquelas terminadas em C ou G. É notória a relação entre os padrões de *Codon Usage Bias* e o conteúdo de G+C genômico (Knight et al. 2001). No entanto não é possível afirmar quem é causa e quem é conseqüência. O conteúdo de GC identificado no eucalipto foi de 53%.

5.3. Codon Usage Bias de Acordo com Níveis de Expressão

Diversos trabalhos sugerem que genes expressos em níveis elevados possuem um padrão de utilização de códons que reflete as trincas mais freqüentes (Ikemura 1985). Identificar estas trincas é um dos objetivos principais deste trabalho. Para isto extraímos do grupo de ORFs aquelas que representavam os 10 genes mais expressos, os 100 mais expressos e os 100 menos expressos de acordo com os dados de expressão in silico gerados pelo projeto FORESTs. O padrão de utilização de códons desses conjuntos foi comparado com o apresentado por todo o conjunto de ORFs (figura 7). A grande maioria das trincas seguem a mesma orientação nos três grupos, mas é possível observar que existe um aumento na variação do *codon usage* observado menos o esperado no grupo de proteínas mais expressas (10+ e 100+), sugerindo que existe um maior grau de *codon usage bias* em genes altamente expressos.

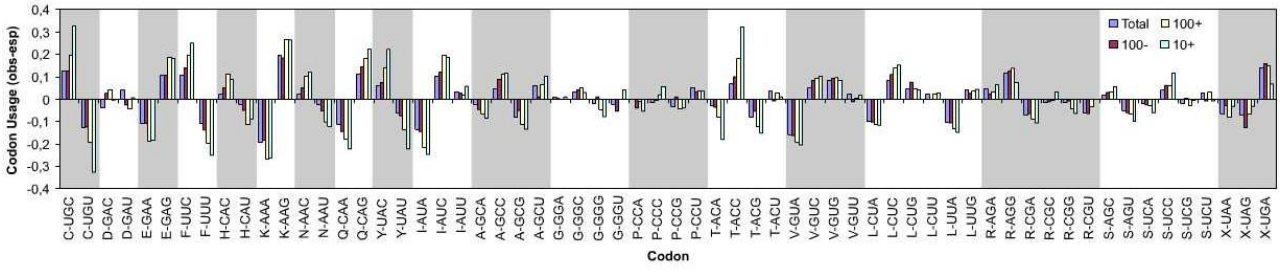


Figura 7. Padrão de códon usage bias apresentado pelas ORFs extraídas dos grupos de genes expressos em diferentes níveis

Este resultado é muito significativo, pois indica um vício mais intenso por códons ótimos em genes altamente expressos o que corrobora a idéia de que existe pressão seletiva agindo sobre o processo traducional, e mais especificamente agindo sobre a presença ou ausência de trincas de acordo com sua ocorrência no organismo.

5.4. Codon Usage Bias de Acordo com Classes Protéicas

As ORFs foram distribuídas em grupos distintos de acordo com a classe protéica que apresentavam segundo os dados de homologia. Isso permitiu que fosse verificada a possível existência de padrões diferenciais na utilização de códons dentro destas classes. Adicionalmente, possibilitou uma consolidação das análises sobre níveis de expressão visto que algumas proteínas, tais como proteínas ribossomais e histonas, são consideradas altamente expressas segundo a literatura (Kanaya et al. 2001).

Como mostra a figura 8 não foram observadas diferenças substanciais nos códons mais utilizados nas diferentes classes testadas, com exceção da trinca de terminação UAA, mais freqüentemente utilizada em histonas.

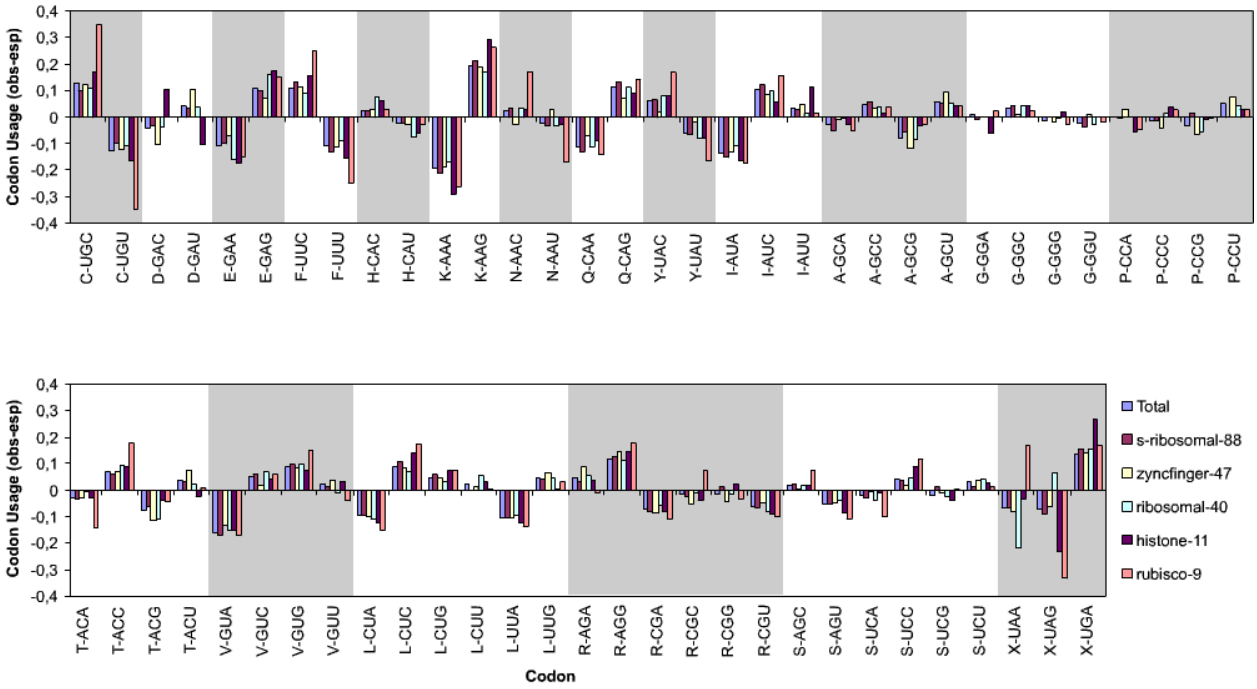


Figura 8. Perfil de códon usage bias em diferentes classes protéicas. O valor ao lado de cada classe indica o numero de seqüências que foram contabilizadas

.

5.5. Códon Usage Bias de Acordo com Local de Expressão

Plotkin et al. (2004) demonstraram que existem diferenças na utilização de códons sinônimos em genes expressos especificamente em seis tecidos adultos humanos diferentes e sugeriram que um controle traducional mediado por códons pode exercer uma importante função na regulação de genes tecido-específicos.

Os locais de expressão das ORFs foram determinados de acordo com os dados apresentados por Vicentini et al. (2005) e mostraram que o eucalipto utiliza, na grande maioria das vezes, de forma semelhante os mesmos códons preferenciais em todos os seus tecidos (figura 9). No caso da isoleucina é possível observar que as trincas mais e menos utilizadas são as mesmas nos diferentes tecidos, mas que existe um comportamento não uniforme para a trinca AUU. Para o ácido aspártico não foi possível definir apenas uma trinca

como ideal, visto que nas diferentes análises foi observado variação na trinca preferencial.

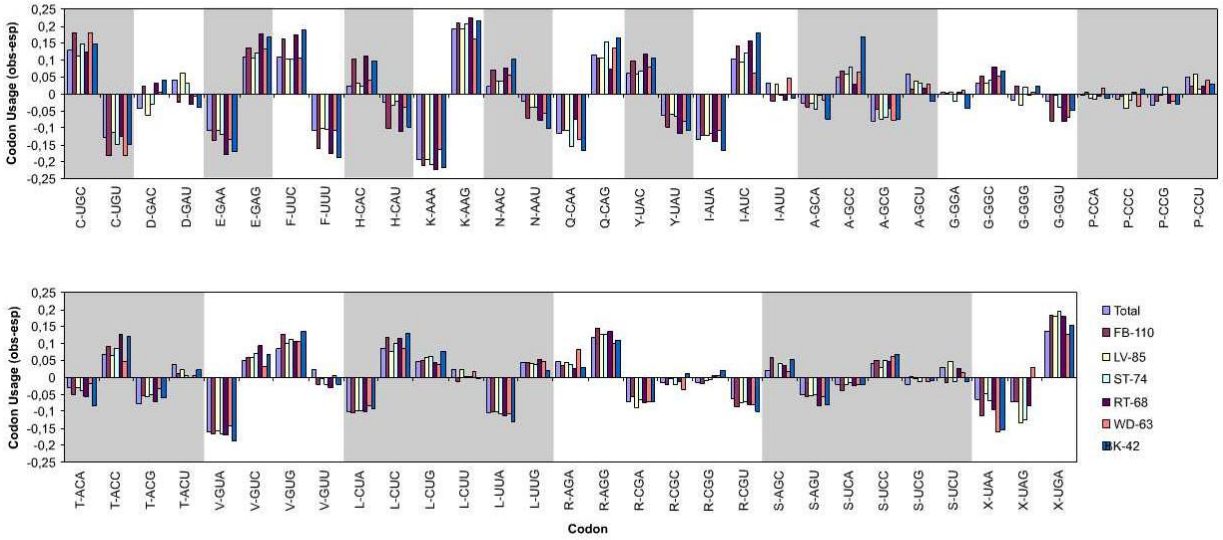


Figura 9. Perfil de códon usage bias de mRNAs expressos preferencialmente em diferentes tecidos. FB-flores e brotos; LV-folhas; ST- caule, plantas com seis meses; RT – raiz, plantas com seis meses, WD – corpo lenhoso, plantas entre 6 e 8 anos, BK – casca, plantas de 6 a 8 anos. Os números ao lado de cada tecido indicam o numero de seqüências que foram contabilizadas

É importante salientar que Plotkin et al. (2004) fizeram suas análises com genes específicos enquanto as análises de Eucalipto foram feitas sobre genes preferencialmente expressos, mas não exclusivos de cada tecido. Isto ocorre porque os dados de Plotkin foram retirados de bibliotecas tecido específicas, o que não ocorreu com nossos dados.

A análise conjunta dos dados de codon usage referentes ao nível de expressão, expressão em tecidos preferenciais e classe protéica permitiu apontar os códons mais freqüentes para cada aminoácido e para indicar o fim da tradução

5.6. Influência de Códons Vizinhos Sobre o Padrão de Utilização Preferencial de Códon

Durante as análises foram identificados alguns outros padrões que poderiam ser interessantes. O que mais chamou atenção foi uma possível relação entre códons adjacentes. Aparentemente existe uma relação entre a terceira base de um códon e a terceira base do códon vizinho.

Foi contabilizada a freqüência com que duplas de códon ocorriam a fim de estabelecer possíveis influências de um códon sobre a trinca seguinte. Como mostra a tabela 3 existe uma relação positiva entre a terceira base do códon anterior e a terceira base do códon seguinte, isto é na posição onde ocorrem a maior parte de mutações silenciosas, mutações onde a substituição de uma base não altera o aminoácido traduzido (Knight et al. 2001). Não foram observadas relações entres bases em outras posições das trincas.

Tabela 2. Freqüência com que ocorrem cada um dos quatro nucleotídeos na ultima posição de um trinca em função do nucleotídeo imediatamente anterior a ela.

		Terceira Base do Segundo Códon			
		A	C	G	T
Terceira Base do Primeiro Códon	A	0,223015	0,217838	0,25117	0,307978
	C	0,129092	0,383287	0,286629	0,200991
	G	0,166527	0,29331	0,321884	0,218279
	T	0,215205	0,236789	0,230391	0,317614

5.7. Contexto de Iniciação

A Tradução por ribossomos citosólicos geralmente ocorre no primeiro AUG do RNA mensageiro. No entanto, em mRNAS de eucariotos, o reconhecimento eficiente de um códon AUG como ponto de início da tradução, depende de vários fatores, em especial da seqüência de nucleotídeos que flanqueia o sítio de início (Kozak, 2005). Existem evidências de que o contexto que envolve o

códon de iniciação contribui para o controle do início da tradução (Kawaguchi et al. 2005). A seqüência ao redor do primeiro AUG, em especial a região não traduzida, também pode modular eficiência com a qual este AUG é reconhecido como início da tradução (Mignone et al. 2002). Se o primeiro AUG reside em um contexto favorável, a síntese protéica será iniciada. Quando o quadro é menos favorável, a síntese protéica poderá ter início no próximo AUG do mRNA.

Com alinhamento das seqüências codificantes baseado no ATG inicial foi verificada a freqüência com que cada base ocorre em cada posição (fig.10) e definido como contexto de iniciação a seguinte seqüência: g n g g g g g n g g c a:g A:G a:c g AUG G c g a:g n g:c g n g:c a:g a g g n g:c, que apresenta alta similaridade com os contextos apresentado por Joshi et al. (1997) para monocotiledôneas e eudicotiledôneas.

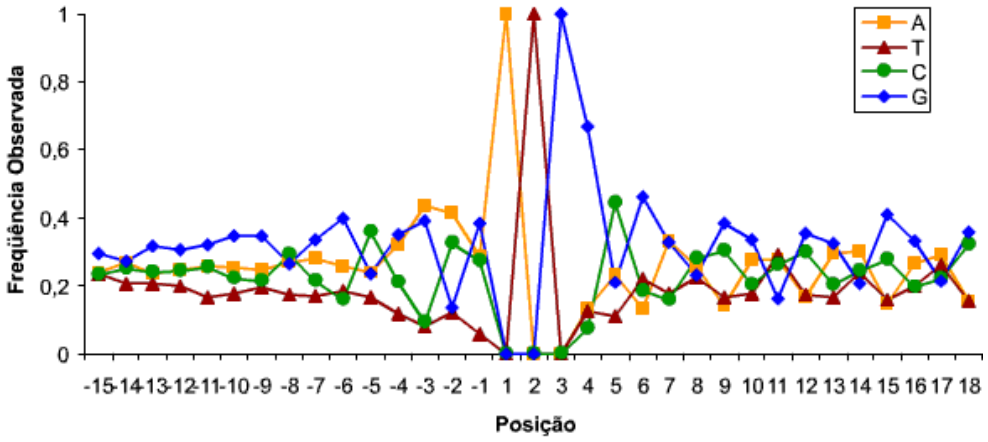


Figura 10. Freqüência relativa dos nucleotídeos ao redor do AUG inicial**.**

Esses dados corroboram a metodologia utilizada para identificação das ORFs que foi utilizada, pois mostram que as seqüências identificadas apresentam contexto de iniciação bem semelhante aos descritos na literatura.

Joshi et al. (1997) observaram que eudicotiledôneas apresentam uma região rica em A antes do AUG inicial, já monocotiledôneas apresentam a mesma região rica em G e C, e que existe uma periodicidade de G nas posições –3, -6

e –9 e nas posições +6 +9 +12 neste grupo de plantas. Segundo estes dados o contexto de iniciação do Eucalipto apresenta muito mais similaridade com contexto de monocotiledôneas do que com eudicotiledôneas, grupo ao qual pertence.

5.8. Contexto de Terminação

O término da tradução é determinado por um dos três códons de terminação (UAA, UGA ou UAG). O processo de término é bastante eficiente e os erros ocorrem à uma taxa estimada de 0,3% em leveduras. O principal erro que pode ocorrer é a incorporação de um aminoácido no lugar do códon de terminação. Um fator conhecido e determinante no término da tradução é a seqüência de nucleotídeos que circunda o códon de parada. Numerosos resultados indicam que o nucleotídeo imediatamente seguinte à trinca de término é um determinante crucial para o término eficiente. Estudos de larga escala demonstraram que a distribuição deste nucleotídeo não é aleatória (Namy et al. 2001)

As análises de *codon usage bias* identificaram o códon UGA como o mais utilizados para sinalizar a terminação da tradução nos genes de Eucalipto (figura.11). Esta trinca ocorre em 51% das ORFs enquanto UAA e UAG ocorrem em 26% e 23% respectivamente. Muitos autores sugerem que a terminação da tradução não seja desencadeada apenas por um dos três códons de parada, mas sim por um contexto de terminação que, além da trinca que indique a terminação, apresente bases específicas presentes em sua vizinhança. Dentre estas a que parece ser a mais importante é a que ocorre adjacente ao último nucleotídeo da trinca de parada. No entanto não identificamos a ocorrência preferencial de nenhum nucleotídeo após o códon de terminação.

A observação da figura 11 mostra que antes do códon de parada existe um padrão periódico na forma como os nucleotídeos ocorrem e que esta periodicidade deixa de ocorrer após o ponto de término da tradução. Este

mesmo fenômeno também pode ser observado na figura 10, onde começa a ocorrer a partir da base -6 antes do AUG inicial.

É amplamente conhecido que seqüências codificantes apresentam esse tipo de padrão e inclusive alguns autores sugerem a utilização deste fenômeno na construção de ferramentas que possibilitem a identificação de éxons (Eskesen et al. 2004).

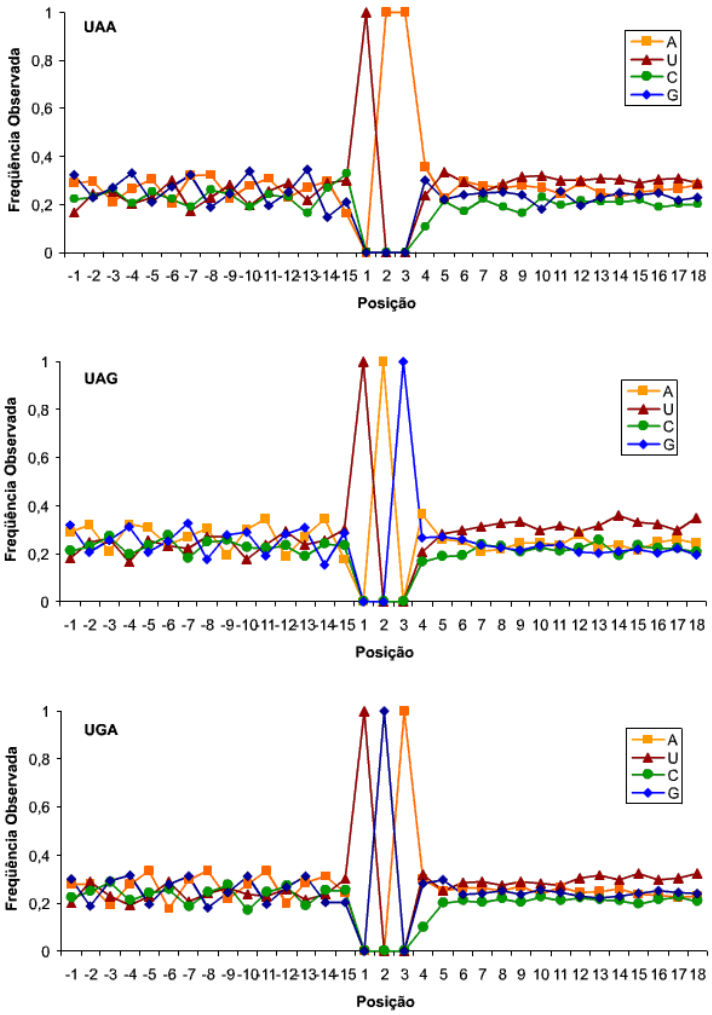


Figura 11. Freqüência relativa dos nucleotídeos ao redor dos três possíveis códons de terminação.

5.9. Construção dos Vetores

Trabalhos recentes reportam a transformação de eucalipto tanto por intermédio de *Agrobacterium* quanto por biolística. A maioria destes trabalhos utilizou o gene nptII, que confere resistência ao antibiótico kanamicina, como gene de resistência (Gonzalez et al. 2002; Satoretto et al. 2002). Desta forma optamos por vetores da série pCAMBIA que além de conter o gene nptII pudessem ser utilizados tanto em ensaios de biolística quanto em transformação via *Agrobacterium*.

Como nenhum vetor da série pCAMBIA contém simultaneamente os genes mGFP5 e nptII foi construído um vetor com essas características. Os genes sintéticos também foram inseridos em vetores semelhantes, de forma que a única diferença entre as três construções (2302, 2302-BC e 2302-BCC) fosse a região codificante de GFP. Todo o resto, promotor, terminador, gene de resistência, posição no vetor, deveria ser idêntico.

Ensaios de digestão com diversas enzimas de restrição mostraram que estes objetivos foram alcançados: os genes foram inseridos corretamente e os vetores são idênticos entre si (exceção feita a região codificante de GFP).

5.10. Transformação de Epitélio de Cebola

A transformação via biolística de células de epitélio de cebola permite a detecção da expressão de genes repórteres de uma forma relativamente simples. Apesar dos ensaios de restrição mostrarem que os genes de interesse haviam sido clonados corretamente, era importante saber se estes genes expressariam de forma correta a proteína verde fluorescente. A transformação das células de cebola (figura 12) mostrou que todos os plasmídios estão expressando GFP corretamente e podem ser utilizados na transformação de eucalipto. Como o promotor 35S utilizado em nossas construções, é mais eficiente em eucotiledôneas do que em monocotiledôneas acreditamos que níveis ainda melhores de expressão serão obtidos nos ensaios com eucalipto.

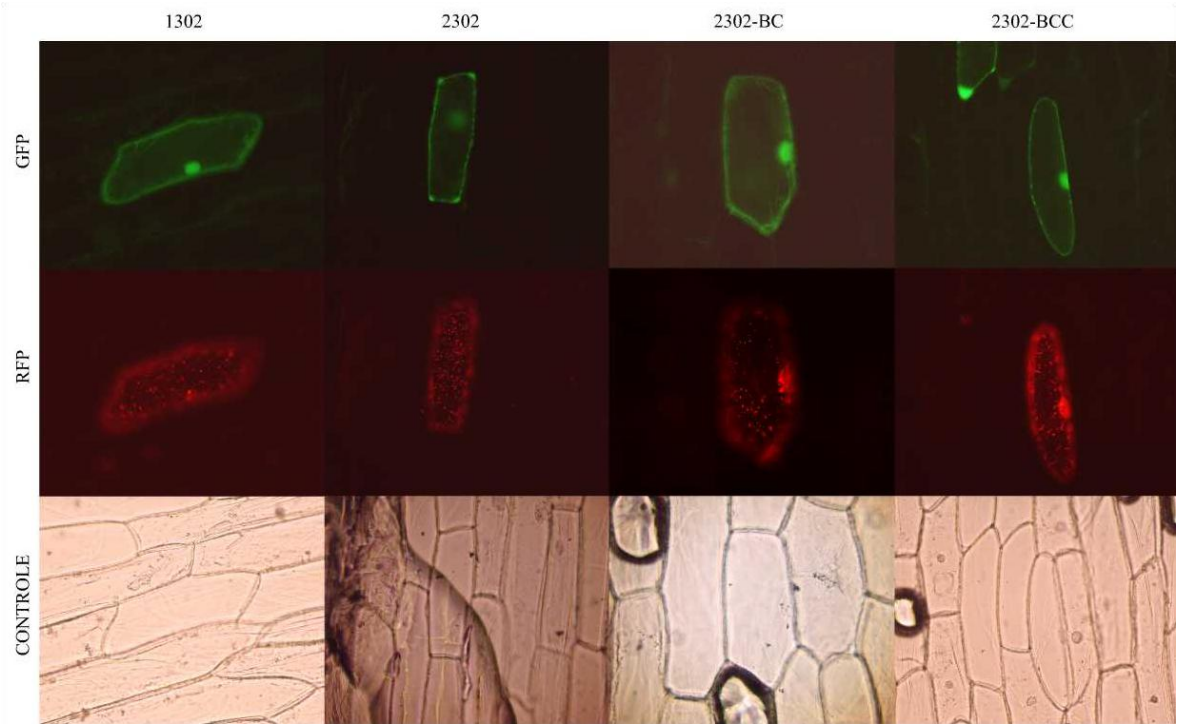


Figura 12. Células de epitélio de cebola transformadas por biolística expressando GFP e RFP. O plasmídio 1302 pertence a família pCambia e não foi alterado, sendo utilizado como controle do experimento. Todos os plasmídios 2302 são derivados do 1302, mas possuem como agente seletivos o gene nptII. Todos plasmídios 2302 codificam as mesmas proteínas. As diferenças entre eles são restritas a seqüência de nucleotídeos referentes à proteína GFP. A expressão da proteína RFP foi obtida através da co-transformação com o vetor RFP-POL.

5.11. Transformação de Eucalipto

A última etapa deste projeto consiste na validação *in vivo* da metodologia de otimização de seqüências gênicas à serem expressas em eucalipto. Para isso, os genes sintéticos desenhados com o software LFG-EGO foram obtidos e utilizados na transformação de tecidos de eucalipto, mas não foi possível obter sucesso nestes experimentos. Desta forma os ensaios *in vivo* foram adiados e estão sendo realizados em pareceria com a empresa Suzano Celulose e Papel.

A empresa Suzano Papel e Celulose possui uma metodologia própria e eficiente de transformação de eucaliptos e se propôs a realizar as transformações. No entanto a empresa não tem interesse em trabalhar com o gene GFP mas sim com o gene nptII, agente seletivo utilizado na seleção de eucaliptos transgênicos. Por este motivo todo o processo de otimização, síntese e clonagem realizado para o gene GFP foi reaplicado ao gene nptII.

Os vetores contendo as diferentes versões do gene nptII estão ao cuidados da Suzano que está realizando as tentativas de obtenção de eucalipto transgênico destes genes

6. CONCLUSÕES

O objetivo deste trabalho foi desenvolver e uma metodologia de otimização da síntese protéica em eucalipto em que através de mudanças nas bases de um gene fosse possível alterar seu nível de expressão sem promover qualquer alteração na seqüência da proteína.

Dentro deste objetivo a primeira etapa foi identificar os padrões de utilização de códons existentes nas seqüências de eucalipto. Esta etapa foi realizada com sucesso e consistiu em:

- Construir um programa chamado ORFIS, capaz de identificar as ORFs das seqüencias depositadas no banco FORESTs;
- Identificar os padrões de utilização de códons apresentado pelo eucalipto;
- Identificar relações entre nível de expressão, tecido e classes protéicas e variações na utilização de códons sinônimos
- Identificar o contexto de iniciação nas seqüência de eucalipto - bases vizinhas ao códon sinalizador do início da tradução;
- Identificar e determinar as relações existentes entre códons vizinhos.

A segunda etapa do projeto consistiu na análise dos dados obtidos e na construção de um programa de computador capaz de desenhar seqüência gênicas otimizadas de maneira automatizada. Este programa foi construído e recebeu o nome de LGF-EGO. E está etapa também foi concretizada com sucesso.

7. PERSPECTIVAS

A validação do processo de otimização está em andamento. A produção de calos transgênicos de eucalipto está sendo conduzida pela empresa Suzano Papel e Celulose. O método de transformação é feito através da utilização de *agrobacterium* e os calos produzidos serão utilizados para verificar a expressão dos genes otimizados.

Os vetores utilizados na transformação contém dois genes que serão expressos em eucalipto: o gene nptII e o gene gusA. Estão sendo construídas três variedades de calos transgênicas. O gene gusA é igual em todas elas. O gene nptI, no entanto, é variável. Uma linhagem de calos expressa uma versão sintética construída apenas com códons ótimos, outra linhagem expressa um gene construído de acordo com a metodologia de códons vizinhos e uma terceira linhagem expressa o gene sem alterações.

A análise da expressão protéica será realizada através de metodologias já conhecidas de quantificação da expressão dos gene nptII e gusA. Através destes dados será possível mensurar i*n vivo* qual o ganho promovido pelo processo de otimização.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Abraf - Anuário Estatístico Abraf – Associação Brasileira de Produtores de Florestas Plantadas, 2007 - http://www.abraflor.org.br/estatisticas.asp - acessado em março de 2008

Alberts B., A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, 2004

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

Andersson, G. E., and C. G. Kurland, 1991 An extreme codon preference strategy: codon reassignment. *Mol.Biol.Evol.* **8**: 530-544.

Baltimore, D., 1992 Viral RNA-dependent DNA polymerase. 1970. *Biotechnology* **24**: 3-5.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, 2006 GenBank. *Nucleic Acids Res.* **34**: D16-D20.

Bernardes, R.F., R. Otsuka, L.A. Colnago, 1998 Análise das Estruturas Secundárias das Proteínas do Glúten de Trigo em Estado Sólido por FTIR. Comunicado técnico **24**:1-4

Besemer, J., and M. Borodovsky, 2005 GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**: W451-W454.

Birney, E., D. Andrews, M. Caccamo, Y. Chen, L. Clarke *et al.* 2006 Ensembl 2006. *Nucleic Acids Res.* **34**: D556-D561.

Boyd, L., and C. S. Thummel, 1993 Selection of CUG and AUG initiator codons for Drosophila E74A translation depends on downstream sequences. *Proc.Natl.Acad.Sci.U.S.A* **90**: 9164-9167.

Cavener, D. R., 1987 Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res.* **15**: 1353-1361.

Cavener, D. R., and S. C. Ray, 1991 Eukaryotic start and stop translation sites. *Nucleic Acids Res.* **19**: 3185-3192.

Crick, F. 1958 On Protein Synthesis. Symp. Soc. Exp. Biol. XII, 139-163

Crick, F., 1970 Central dogma of molecular biology. *Nature* **227**: 561-563.

Deana, A., R. Ehrlich, and C. Reiss, 1998 Silent mutations in the Escherichia coli ompA leader peptide region strongly affect transcription and translation in vivo. *Nucleic Acids Res.* **26**: 4778-4782.

Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg, 1999 Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**: 4636-4641.

Eskesen S.T., F.N. Eskesen,B. Kinghorn, A. Ruvinsky, 2004 Periodicity of DNA in exons. *BMC Mol Biol*. **18:**5-12.

Gibas C, P. Jambeck, 2001. Developing Bioinformatics Computer Skills. Sebastopol: O'Reilly and Associates. 2001. ISBN:1-56592-664-1

González, E. R., A. Andrade; A. L. Bertolo, G. C. Lacerda, R. T. Carneiro, V. A. P. Defávari, M. T. V. Labate, C. A. Labate, 2002 *Production of transgenic Eucalyptus grandis x E.* urophylla *using sonication assisted Agrobacterium transformation (SAAT)*. Aust. J. Pl. Phys. **29:** 97-102

Gouy, M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**: 7055-7074.

Griswold, K. E., N. A. Mahmood, B. L. Iverson, and G. Georgiou, 2003 Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm. *Protein Expr.Purif.* **27**: 134-142.

Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**: 199-209.

Gustafsson, C., S. Govindarajan, and J. Minshull, 2004 Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**: 346-353.

Haruna, I., K. Nozu, Y. Ohtaka, and S. Spiegelman, 1963. An RNA "replicase" induced by and selective for a viral rna: isolation and properties. *Proc.Natl.Acad.Sci.U.S.A* **50**: 905-911.

Haseloff J., K.R. Siemering, D.C. Prasher, S. Hodge, 1997 Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic *Arabidopsis* plants brightly. *Proc.Natl.Acad.Sci.U.S.A* 94: 2122-7

Higgins S.J., B.D. Hames, 1999 Protein Expression: A Practical Approach, Oxford University Press

Hoekema, A., R. A. Kastelein, M. Vasser, and H. A. de Boer, 1987 Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. *Mol.Cell Biol.* **7**: 2914-2924.

Ikemura, T., 1981 Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J.Mol.Biol.* **151**: 389-409.

Ikemura, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol.Biol.Evol.* **2**: 13-34.

Jacobs, G. H., P. A. Stockwell, M. J. Schrieber, W. P. Tate, and C. M. Brown, 2000 Transterm: a database of messenger RNA components and signals. *Nucleic Acids Res.* **28**: 293-295.

Jacobs, G. H., P. A. Stockwell, M. J. Schrieber, W. P. Tate, and C. M. Brown, 2000 Transterm: a database of messenger RNA components and signals. *Nucleic Acids Res.* **28**: 293-295.

Jin, H., A. Bjornsson, and L. A. Isaksson, 2002 Cis control of gene expression in E.coli by ribosome queuing at an inefficient translational stop signal. *EMBO J.* **21**: 4357-4367.

Joshi, C. P., H. Zhou, X. Huang, and V. L. Chiang, 1997 Context sequences of translation initiation codon in plants. *Plant Mol.Biol.* **35**: 993-1001.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus *subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**: 143–155.

Kane, J. F., 1995 Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr.Opin.Biotechnol.* **6**: 494-500.

Kawaguchi, R., and J. Bailey-Serres, 2005 mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Res.* **33**: 955-965.

Kink, J. A., M. E. Maley, K. Y. Ling, J. A. Kanabrocki, and C. Kung, 1991 Efficient expression of the Paramecium calmodulin gene in Escherichia coli after four TAA-to-CAA changes through a series of polymerase chain reactions. *J.Protozool.* **38**: 441-447.

Knight, R. D., S. J. Freeland, and L. F. Landweber, 2001 A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**: RESEARCH0010.

Kozak, M., 1991 An analysis of vertebrate mRNA sequences: intimations of translational control. *J.Cell Biol.* **115**: 887-903.

Kozak, M., 2005 Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**: 13-37.

Larsen, T. S., and A. Krogh, 2003 EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC.Bioinformatics.* **4**: 21.

Lithwick, G., and H. Margalit, 2003 Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.* **13**: 2665-2673.

Loguercio, L. L., M. L. Barreto, T. L. Rocha, C. G. Santos, F. F. Teixeira *et al.* 2002 Combined analysis of supernatant-based feeding bioassays and PCR as a first-tier screening strategy for Vip -derived activities in Bacillus thuringiensis strains effective against tropical fall armyworm. *J.Appl.Microbiol.* **93**: 269-277.

Lukaszewicz, M., Feuermann1 M, B. Jerouville, A. Stas, and M. Boutry, 2000 In vivo evaluation of the context sequence of the translation initiation codon in plants. *Plant Sci.* **154**: 89-98.

Lutcke, H. A., K. C. Chow, F. S. Mickel, K. A. Moss, H. F. Kern *et al.* 1987 Selection of AUG initiation codons differs in plants and animals. *EMBO J.* **6**: 43-48.

Mignone, F., C. Gissi, S. Liuni, and G. Pesole, 2002 Untranslated regions of mRNAs. *Genome Biol.* **3**: REVIEWS0004.

Ministério da Ciência e Tecnologia, Projeto Genolyptus, 2003, http://ftp.mct.gov.br/especial/genolyptus4.htm

Molecular Biology of the Cell 4th ed.: New York and London: Garland Science

Mottagui-Tabar, S., and L. A. Isaksson, 1997 Only the last amino acids in the nascent peptide influence translation termination in Escherichia coli genes. *FEBS Lett.* **414**: 165-170.

Mukherjee, S., and S. Mitra, 2005 Hidden Markov Models, grammars, and biology: a tutorial. *J.Bioinform.Comput.Biol.* **3**: 491-526.

Mulinari, F., F. Staniscuaski, L. R. Bertholdo-Vargas, M. Postal, O. B. Oliveira-Neto *et al.* 2007 Jaburetox-2Ec: an insecticidal peptide derived from an isoform of urease from the plant Canavalia ensiformis. *Peptides* **28**: 2042-2050.

Nadershahi, A., S. C. Fahrenkrug, and L. B. Ellis, 2004 Comparison of computational methods for identifying translation initiation sites in EST data. *BMC.Bioinformatics.* **5**: 14.

Nambiar, K. P., J. Stackhouse, D. M. Stauffer, W. P. Kennedy, J. K. Eldredge *et al.* 1984 Total synthesis and cloning of a gene coding for the ribonuclease S protein. *Science* **223**: 1299-1301.

Namy, O., I. Hatin, and J. P. Rousset, 2001 Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.* **2**: 787-793.

Okubo, K., H. Sugawara, T. Gojobori, and Y. Tateno, 2006 DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.* **34**: D6-D9.

Oliveira, L. M., R. Paiva, J. R. F. Santana, R.C. Nogueira, F.P. Soares, L.C. Silva, 2007 Efeito de citocininas na senescência e abscisão foliar durante o cultivo in vitro de Annona glabra L. *Revista Brasileira de Fruticultura* **29**: 25-30.

Ozawa, Y., S.Hanaoka, R. Saito, T. Washio, S. Nakano, A. Shinagawa, M. Itoh, K. Shibata, P. Carninci, H. Konno, J. Kawai, Y. Hayashizaki, M. Tomita 2002 Comprehensive sequence analysis of translation termination sites in various eukaryotes. *Gene* **300**: 79–87

Plotkin, J. B., H. Robins, and A. J. Levine, 2004 Tissue-specific codon usage and the expression of human genes. *Proc.Natl.Acad.Sci.U.S.A* **101**: 12588-12591.

Poole, E. S., C. M. Brown, and W. P. Tate, 1995 The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *EMBO J.* **14**: 151-158.

Revista Ciencia Hoje - Versão on line http://cienciahoje.uol.com.br/controlPanel/materia/view/2667 - acessado em maio de 2008

Roy, P., S. B. Rondeau, C. Vezina, and G. Boileau, 1990 Effect of mRNA secondary structure on the efficiency of translational initiation by eukaryotic ribosomes. *Eur.J.Biochem.* **191**: 647-652.

Sartoretto, L. M., L. P. B. Cid, A. C. M. Brasileiro, 2002 Biolistic transformation of Eucalyptus grandis x E. urophylla callus. *Fuctional Plant Biology* **(29)**: 917-924.

Sharp, P.M., W.H. Li, 1987 The codon adaptation index – a measure of directional synonimous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281-1295.

Small, I., H. Wintz, K. Akashi, and H. Mireau, 1998 Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol.Biol.* **38**: 265-277.

Stenstrom, C. M., and L. A. Isaksson, 2002 Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. *Gene* **288** : 1-8.

Tate, W. P., E. S. Poole, M. E. Dalphin, L. L. Major, D. J. Crawford *et al.* 1996 The translational stop signal: codon with a context, or extended factor recognition element? *Biochimie* **78**: 945-952.

Tateno, Y., N. Saitou, K. Okubo, H. Sugawara, and T. Gojobori, 2005 DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res.* **33**: D25-D28.

Tateno, Y., T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou *et al.* 2002 DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* **30**: 27-30.

Temin, H. M., C. Y. Kang, and S. Mizutani, 1973 Endogenous RNA-directed DNA polymerase activity in normal cells. *Johns.Hopkins.Med.J.Suppl* **2**: 141-156.

Uchijima, M., A. Yoshida, T. Nagata, and Y. Koide, 1998 Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium. *J.Immunol.* **161**: 5594-5599.

Vicentini, R., and M. Menossi, 2007 TISs-ST: a web server to evaluate polymorphic translation initiation sites and their reflections on the secretory targets. *BMC.Bioinformatics.* **8**: 160.

Vicentini, R., F. T. Sassaki, M. A. GIMENES, I.G. MAIA, M. MENOSSI, 2005 In silico evaluation of the Eucalyptus transcriptome. Genetics and Molecular Biology, 28(3): 487-495.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese *et al.* 2008 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**: D13-D21.

Wu, X., H. Jornvall, K. D. Berndt, and U. Oppermann, 2004 Codon optimization reveals critical factors for high level expression of two rare codon genes in Escherichia coli: RNA stability and secondary structure but not tRNA abundance. *Biochem.Biophys.Res.Commun.* **313**: 89-96.

Zolotukhin, S., M. Potter, W. W. Hauswirth, J. Guy, and N. Muzyczka, 1996 A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J.Virol.* **70**: 4646-4654.

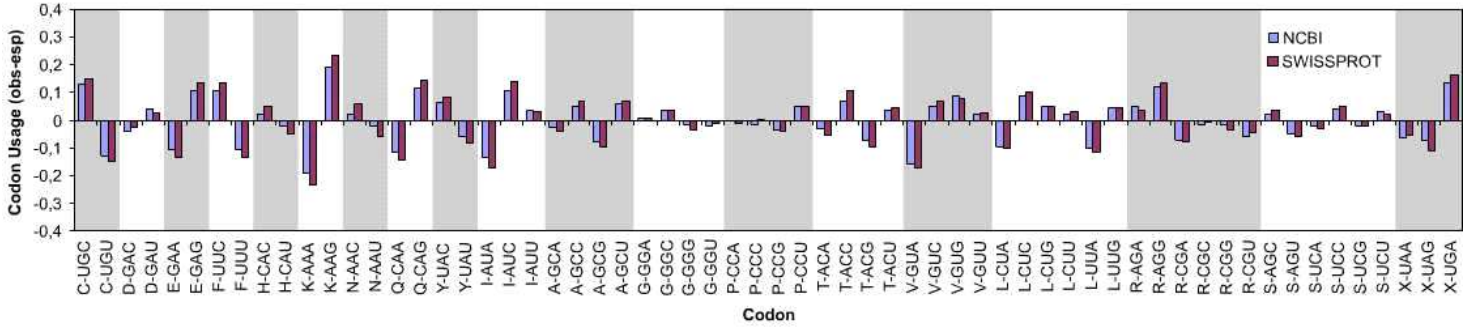


Figura 6 Padrão de *codon usage bias* apresentado pelas ORFs extraidas dos dados FORESTs. O eixo Y refere-se à diferença entre o *codon usage* observado em cada trinca e o valor referente a ocorrência aleatória da mesma; resultados positivos indicam uma utilização superior ao aleatório já valores negativos indicam uma utilização inferior. No eixo X estão representados os aminoácidos e as trincas correspondentes. Este artifício permite uma rápida análise gráfica comparativa do padrão de *codon usage bias* e foi utilizado nas figuras 1,2,3 e 4.

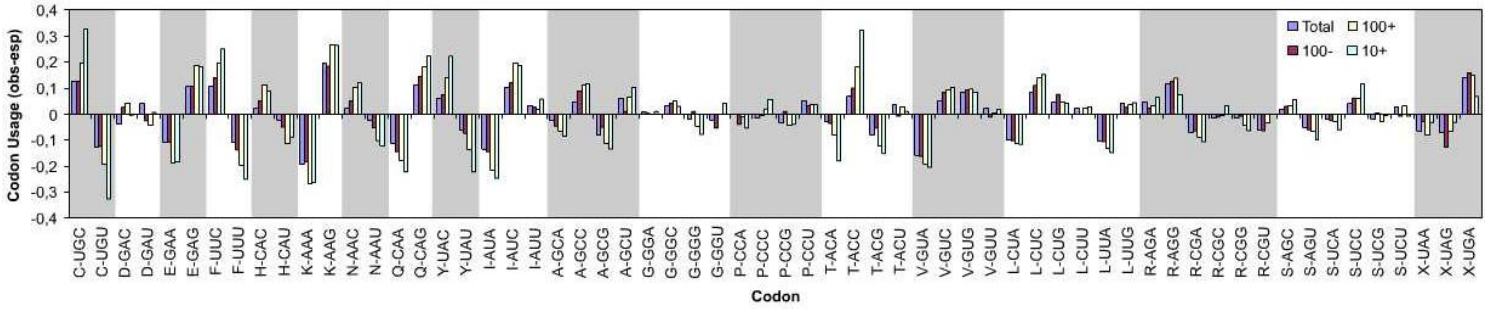


Figura 7 Padrão de códon usage bias apresentado pelas ORFs extraídas dos grupos de genes expressos em diferentes níveis

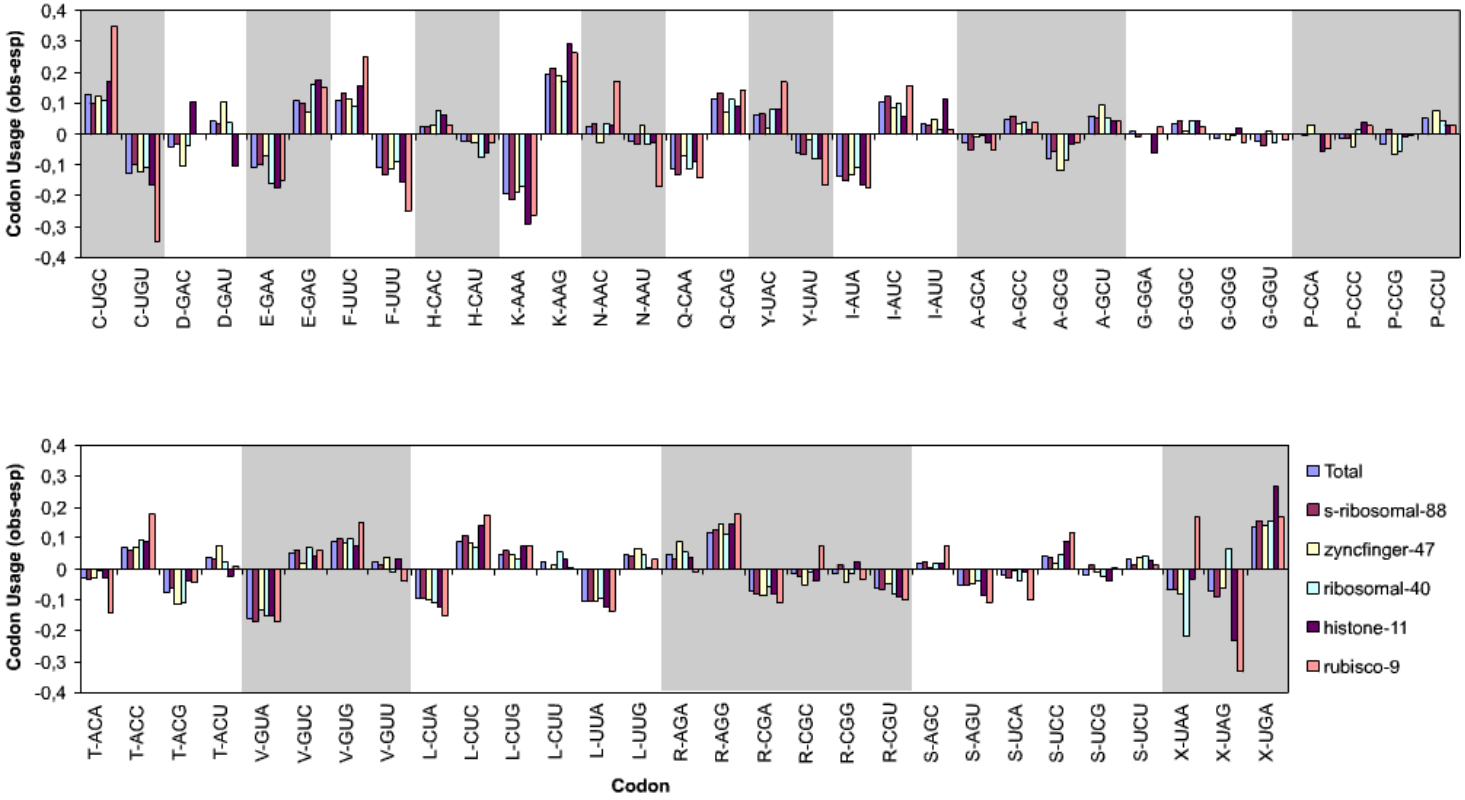


Figura 8 Perfil de códon usage bias em diferentes classes protéicas. O valor ao lado de cada classe indica o numero de seqüências que foram contabilizadas

.

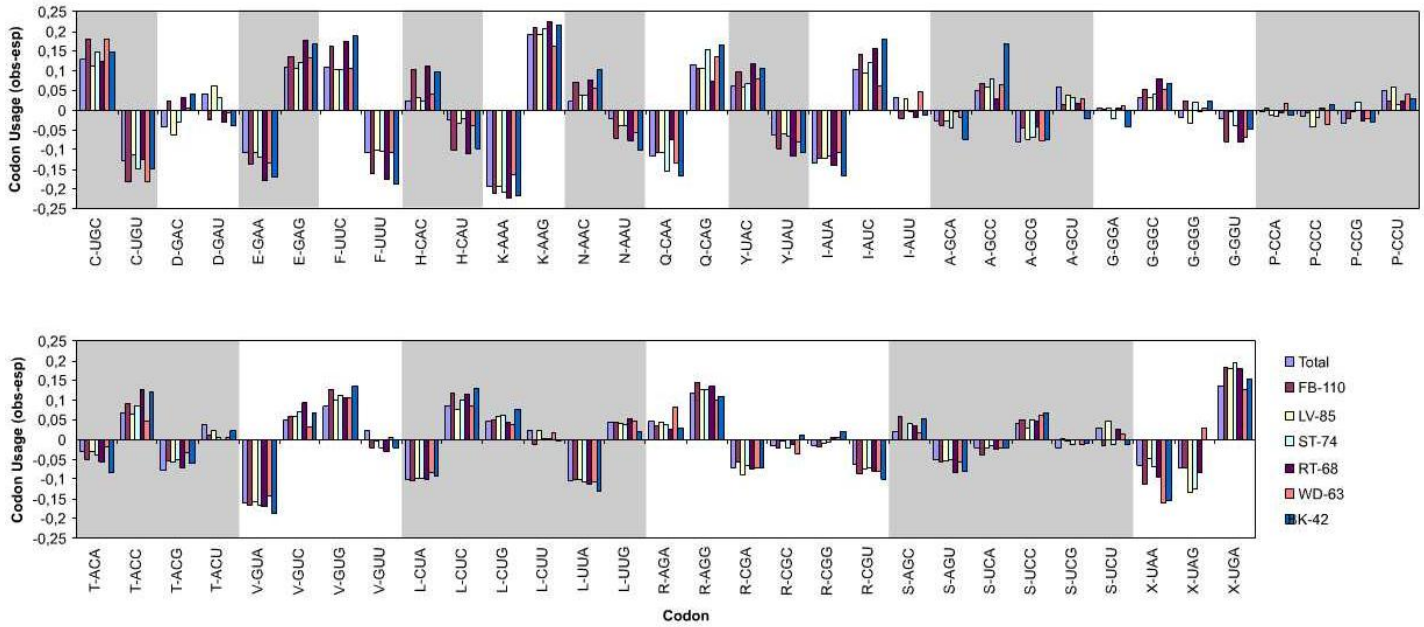


Figura 9 Perfil de códon usage bias de mRNAs expressos preferencialmente em diferentes tecidos. FB-flores e brotos; LV-folhas; ST- caule, plantas com seis meses; RT – raiz, plantas com seis meses, WD – corpo lenhoso, plantas entre 6 e 8 anos, BK – casca, plantas de 6 a 8 anos. Os números ao lado de cada tecido indicam o numero de seqüências que foram contabilizadas