Predição *in silico* de genes baseado no sistema MYOP

Aluno: Renato Cordeiro Ferreira Orientador: Alan Mitchell Durham
IME-USP, São Paulo (PIC 2013/2014 - Bolsista CNPq)

Introdução

Encontrar genes é uma tarefa essencial como guia na biologia molecular moderna. Eles são usados na síntese de proteínas - moléculas essenciais para o desenvolvimento de organismos complexos [2]. Uma ferramenta importante para identificar genes são os preditores de genes *ab initio*, que utilizam modelos estatísticos para automatizar o processo de busca e classificar as regiões do DNA com maior probabilidade de codificarem proteínas. Atualmente, várias destes programas estão disponíveis: Genscan, SNAP e Augustus dentre os mais usados. Neste estudo, comparamos a acurácia dos três contra o gerador de preditores MYOP [1], tendo como base *datasets* de validação de 6 diferentes organismos. Também analisamos o uso do Augustus e MYOP no genoma do sorgo (*Sorghum bicolor*), fornecendo uma *pipeline* automatizada para repetir este processo com outras espécies.

Materiais e Métodos

Dados

Para testar a acurácia dos preditores, foram usados 6 dos 10 *datasets* de validação desenvolvidos no grupo de pesquisas do Prof. Alan Durham, cada um contendo 2000 genes validados. As espécies e os preditores nas quais foram testadas são apresentados naTabela 1:

	MYOP	Augustus	SNAP	Gensan
A. thaliana	X	X	X	X
H. sapiens	X	X		X
C. elegans	X	X	X	
D. melanogaster	X	X	X	
Z. mays	X	X		X
O. sativa	X	X	X	

Tabela: *Datasets* de validação, cada um apresentando 2000 sequências contendo genes previamente classificados.

SNAP e Genscan não forneciam maneiras de fazer treinamento, e usamos apenas organismos pré-treinados por um deles. Assim, esses preditores poderiam se desempenhar melhor, pois incluiriam as sequências preditas nos conjuntos de treinamento.

O genoma do sorgo foi obtido do banco de sequências verificadas da NCBI (RefSeq), na forma de 10 cromossomos montados. Para gerar a predição de referência para comparações com MYOP e Augustus, foram usados 204.208 sequências expressas (*ESTs*) obtidas do *Plant Genome Database* (PlantGDB).

Comparação de preditores

Para validar programas que realizam classificação probabilística, podemos realizar uma **validação cruzada em *k-fold***. Neste trabalho, utilizamos **k=5** (400 genes por subconjunto):

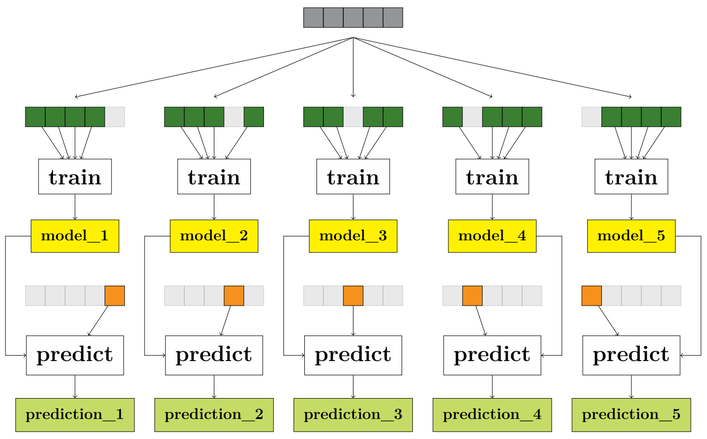

Figura: Validação-cruzada: dados são particionados em k conjuntos: k-1 para treinamento e 1 para predição. No final, obtém-se a média das estatísticas de acurácia alternando o conjunto de predição.

Comparando a rotulação conhecida com a realizada pelo preditor, podemos separar cada classificação em 4 grupos, e gerar estatísticas de **sensibilidade** (SN), **precisão** (PPV) e ***F-score***. Para tanto, utilizamos a ferramenta **SGEVal** para gerar os dados para os diagramas de Venn.

	Condição positiva	Condição negativa
Resultado positivo	Verdadeiro positivo	Falso positivo
Resultado negativo	Falso negativo	Verdadeiro negativo

Tabela: Verdadeiro positivo (TP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (TN).

$$\text{SN} = \frac{TP}{TP+FN} \qquad \text{PPV} = \frac{TP}{TP+FP} \qquad F\text{-}score = \frac{2 * SN * PPV}{SN+PPV}$$

Resultados

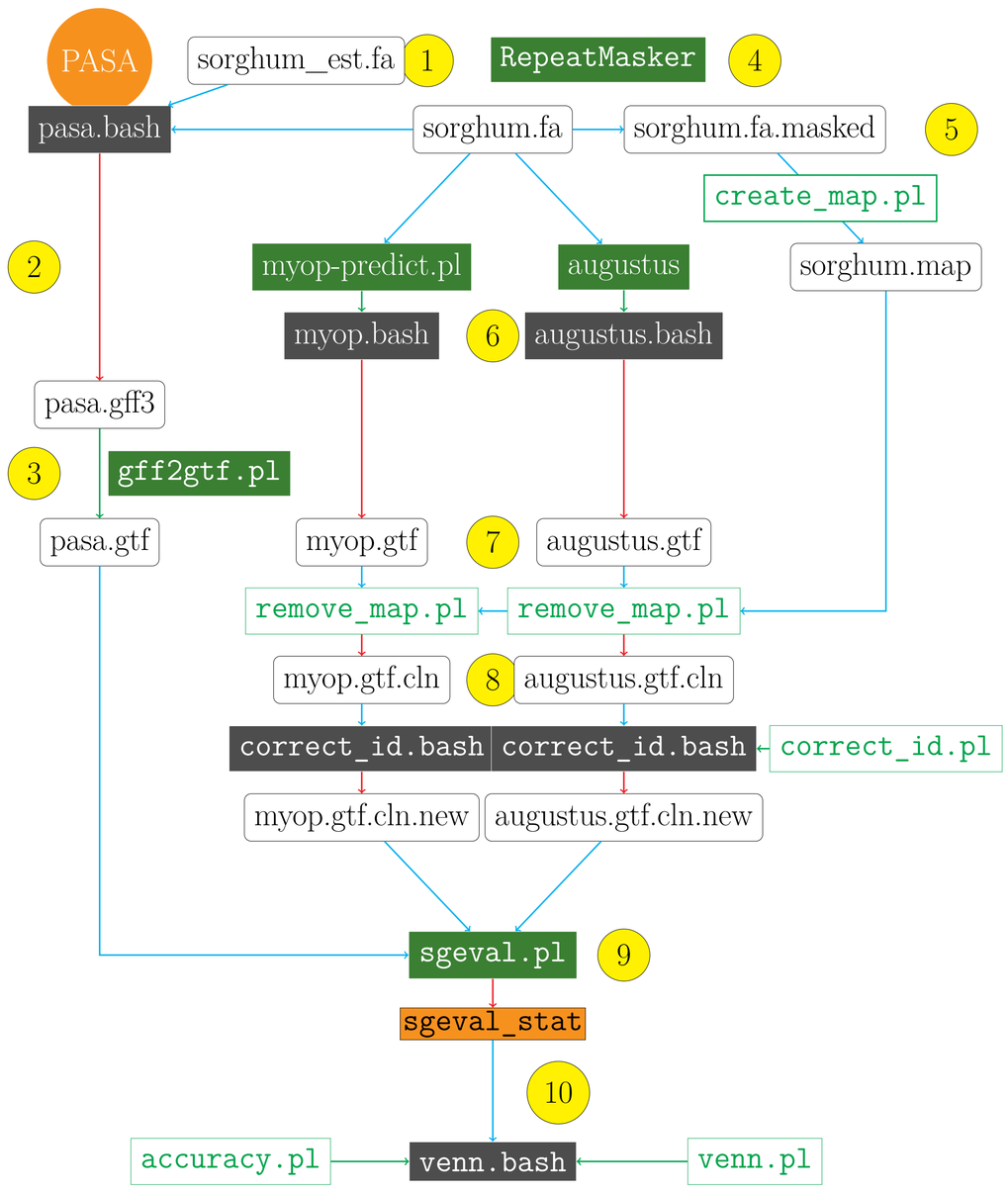

Figura: Pipeline de predição do sorgo.

1. Procurar os dados do genoma e de ESTs
2. Aplicar a *pipeline* do PASA para gerar o posicionamento dos ESTs sobre o genoma;
3. Criar predição de referência baseada nos ESTs a partir dos dados armazenados do PASA;
4. Mascarar as regiões com sequências de baixa complexidade e alto nível de repetições (RepeatMasker);
5. Gerar um mapa das regiões mascaradas, com relação às posições absolutas delas no genoma;
6. Gerar predições de genes *ab initio* utilizando os programas MYOP e Augustus;
7. Retirar as predições incorretas do Augustus e MYOP, feitas sobre regiões mascaradas, conforme registradas no mapa de mascaramentos;
8. Renumerar as identificações dos genes e transcritos (*gene id* e *transcript id*) das sequências preditas;
9. Comparar, via o programa SGEval [1] a eficiência MYOP e Augustus contra os a predição de referência gerada pelo banco de dados do PASA;
10. Criar um diagrama de Venn com as predições de éxons e íntrons do PASA, MYOP e Augustus;
11. Comparar as predições conflitantes contra um banco de proteínas, usando um sistema de *scores* que favoreça os alinhamentos mais longos usando o programa Blast.
12. Usar as sequências preditas em comum de MYOP e AUGUSTUS e as escolhidas conforme o critério acima para gerar uma reanotação do genoma do Sorgo. Usar a *pipeline* do PASA para realizar a junção dos ESTs, genoma e predições.

		Gene			Éxon			Nucleotídeo		
		PPV	SN	F-score	PPV	SN	F-score	PPV	SN	F-score
A. thaliana	augustus	53.063 ± 3.574	50.150 ± 3.527	51.563 ± 3.537	86.025 ± 0.869	81.402 ± 1.354	83.648 ± 1.102	98.241 ± 0.500	95.731 ± 0.784	96.969 ± 0.623
	genscan	23.622 ± 2.589	22.700 ± 2.363	23.151 ± 2.471	70.321 ± 1.023	59.377 ± 1.453	64.379 ± 1.107	96.001 ± 0.937	86.045 ± 1.101	90.750 ± 1.005
	myop	69.088 ± 1.797	61.600 ± 2.004	65.127 ± 1.901	89.826 ± 0.499	85.552 ± 1.330	87.634 ± 0.926	98.614 ± 0.373	95.971 ± 0.758	97.274 ± 0.564
	snap	41.062 ± 2.784	42.400 ± 2.262	41.711 ± 2.474	78.330 ± 1.320	73.893 ± 1.779	76.045 ± 1.550	97.526 ± 0.389	93.573 ± 0.723	95.508 ± 0.546
H. sapiens	augustus	19.384 ± 1.779	27.750 ± 2.080	22.819 ± 1.913	73.029 ± 1.168	71.770 ± 0.415	72.386 ± 0.455	75.423 ± 1.779	86.871 ± 0.757	80.732 ± 1.138
	genscan	9.523 ± 1.493	18.750 ± 1.891	12.622 ± 1.725	54.836 ± 2.378	74.529 ± 0.517	63.161 ± 1.736	61.540 ± 2.334	90.069 ± 0.587	73.099 ± 1.769
	myop	21.580 ± 0.840	24.750 ± 0.742	23.054 ± 0.778	71.076 ± 2.370	71.425 ± 0.452	71.226 ± 1.124	73.780 ± 2.628	86.799 ± 0.766	79.738 ± 1.640
C. elegans	augustus	26.742 ± 2.094	29.675 ± 2.401	28.129 ± 2.220	75.070 ± 1.944	74.374 ± 2.005	74.701 ± 1.556	88.001 ± 2.341	91.147 ± 1.190	89.529 ± 1.405
	myop	43.791 ± 1.990	44.151 ± 2.490	43.961 ± 2.170	78.925 ± 1.965	81.296 ± 1.398	80.070 ± 1.038	87.917 ± 2.396	93.631 ± 0.730	90.665 ± 1.302
	snap	23.755 ± 1.408	29.978 ± 1.237	26.500 ± 1.306	71.505 ± 2.500	75.381 ± 1.071	73.377 ± 1.663	87.410 ± 2.631	92.245 ± 1.015	89.750 ± 1.739
D. melanogaster	augustus	45.288 ± 4.295	54.311 ± 2.556	49.332 ± 3.423	73.510 ± 4.348	77.859 ± 2.541	75.577 ± 3.141	87.107 ± 3.694	93.761 ± 1.547	90.275 ± 2.349
	myop	54.881 ± 2.982	58.702 ± 1.825	56.701 ± 2.254	75.607 ± 4.359	80.634 ± 1.379	77.988 ± 2.728	87.305 ± 4.118	94.336 ± 1.182	90.637 ± 2.453
	snap	32.592 ± 2.715	42.143 ± 1.640	36.706 ± 2.112	63.563 ± 3.629	71.207 ± 2.621	67.120 ± 2.735	86.943 ± 3.412	91.311 ± 1.077	89.042 ± 2.023
Z. mays	augustus	34.304 ± 3.709	36.100 ± 2.267	35.146 ± 2.930	62.985 ± 2.873	58.382 ± 2.079	60.560 ± 1.968	74.901 ± 4.259	84.985 ± 0.657	79.574 ± 2.657
	genscan	7.094 ± 0.970	7.600 ± 1.200	7.337 ± 1.078	28.045 ± 1.459	21.989 ± 1.508	24.618 ± 1.273	68.071 ± 4.562	54.987 ± 1.129	60.759 ± 1.912
	myop	46.061 ± 2.655	38.200 ± 1.317	41.745 ± 1.724	66.799 ± 2.334	62.505 ± 2.167	64.549 ± 1.743	76.848 ± 4.124	85.387 ± 0.601	80.839 ± 2.362
O. sativa	augustus	26.117 ± 1.285	25.614 ± 1.761	25.860 ± 1.515	59.454 ± 1.260	50.195 ± 2.367	54.417 ± 1.839	91.333 ± 0.863	87.207 ± 1.259	89.221 ± 1.061
	myop	37.999 ± 0.878	30.777 ± 1.103	34.001 ± 0.905	64.024 ± 1.919	55.450 ± 2.509	59.413 ± 2.105	91.751 ± 0.641	89.256 ± 0.895	90.484 ± 0.641
	snap	18.869 ± 1.327	21.053 ± 1.909	19.898 ± 1.585	42.109 ± 2.138	29.335 ± 2.315	34.571 ± 2.318	85.410 ± 1.115	54.309 ± 1.619	66.391 ± 1.479

Figura: Resultados da predição de genes completos, éxons e nucleotídeos do MYOP, Augustus, SNA e Genscan sobre os genomas dos 6 organismos. Destacados, o maior valor absoluto para a coluna, com verde representando o MYOP e vermelho, outros preditores.

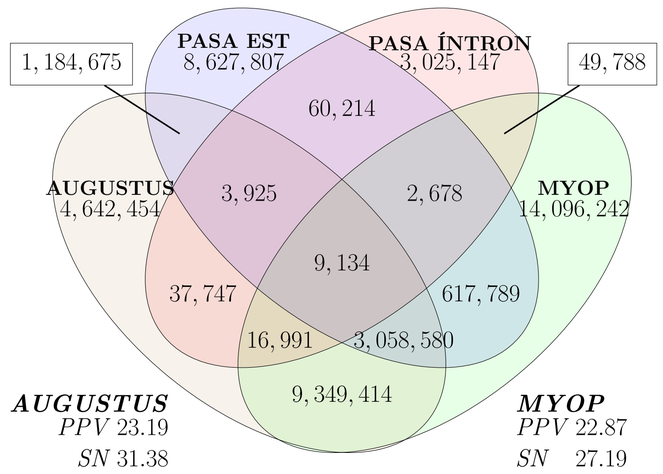

Figura: Diagrama de Venn com 4 conjuntos: ESTs alinhados com auxílio do PASA no genoma do sorgo, íntrons derivados desse alinhamento, predições do MYOP e do Augustus. A separação dos éxons e íntrons gerados pelo PASA a partir dos ESTs permite que o SGEval obtenha dados de SN, PPV e *F-score* quando há eventos de *splicing* alternativo.

IME-USP

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Conclusões e Trabalhos futuros

Validação

As predições realizadas sobre os 6 organismos, com comparações entre os 4 preditores, demonstraram uma clara vantagem do MYOP com relação aos outros preditores de gene. Em muitos casos, o preditor superou os outros, na média, mais que um desvio-padrão.

Anotação do Sorgo

Os resultados de sensibilidade e precisão do Augustus e do MYOP foram muito aquém do esperado ao calcular as estatísticas utilizando o SGEval. Essa diminuição ocorreu por conta da diferença na metodologia aplicada: parte das contagens realizadas pelos programas considerava como falso positivo predições que poderiam estar corretas (mas não eram validadas pelos ESTs). Novas contagens precisam ser geradas para considerar corretamente a acurácia de ambos os preditores.

Referências

[1] A. Y. Kashiwabara, **MYOP/ToPS/SGEval: Um ambiente computacional para estudo sistemático de predição de genes**.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Margin Raff and Peter Walter, "**Molecular Biology of the Cell**", *Garland Scence*, 4th, 2002.