Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Rua Walter Ianni, 255 - São Gabriel, Belo Horizonte - MG

Extração de regras conservativas para predição de início de sítio de tradução de proteínas utilizando análise formal de conceitos

Leandro M. Ferreira, Cristiano Lacerda Nunes Pinto, Sérgio M. Dias,
Cristiane N. Nobre, Luís E. Zárate
leandromferreira.eng@gmail.com

30 de outubro de 2019

Content

Introdução

- Características conservativas que definem os processos de tradução e transcrição utilizadas pelas células para interpretar informações genéticas
- *Coding Sequence*.
- Técnica de analise formais de conceitos (AFC)

Analise Formal de Conceito

- AFC é uma técnica da matemática introduzida no início dos anos 80 por Rudolf Wille [7].
- Depende da construção do contexto formal para representar problema especifico.
- Utiliza da teoria dos reticulados conceituais para organizar hierarquicamente objetos a partir de conjuntos de conceito formais composto por um par ($objetos$,$atributos$) e sua relação.
- Pode-se obter regras de implicação (comumente chamadas de implicações), que são regras indicativas de relação entre subconjuntos de atributos relacionados aos objetos.

Definição

Um contexto formal consiste de dois conjuntos e uma relação binária entre eles. Mais precisamente, um contexto formal é uma tripla (G, M, I) em que $I \subseteq G \times M$, sendo os elementos do conjunto G chamados de *objetos*, os elementos do conjunto M chamados de *atributos* e I chamada de *relação de incidência*. Em outras palavras $(g, m) \in I$ ou simplesmente gIm deve ser lida como "o objeto g contem o atributo m ".

Tabela: Exemplo de contexto formal

	-9				**-8**				**-7**				**-6**			
	a	**t**	**c**	**g**	a	t	c	g	a	t	c	g	a	t	c	g
1	X	.	.	.	X	.	.	.	X	.	.	.	X	.	.	.
2	.	.	.	X	.	.	X	.	.	X	.	.	.	X	.	.
3	.	.	.	X	.	.	X	.	.	.	X	.	.	.	X	.
4	.	.	.	X	.	.	X	.	.	.	.	X	.	.	.	X

Definição

Os *conceitos formais* obteníveis a partir de um contexto formal (G, M, I) são pares ordenados (A, B), em que $A \subseteq G$ e $B \subseteq M$, sendo que cada objeto em A possui todos os atributos em B e cada atributo em B é atributo de todos os objetos em A. Em outras palavras, (A, B) é um conceito formal se e somente se $A' = B$ e $B' = A$. Os conjuntos A e B são denominados *extensão* e *intenção* do conceito, respectivamente. Para se referenciar o conjunto de todos os conceitos formais existentes em um contexto formal (G, M, I), será usada a notação $\mathcal{B}(G, M, I)$.

Definição

Seja um contexto formal cujo conjunto de atributos é M. Uma implicação é uma expressão $P \rightarrow Q$, em que $P, Q \subseteq M$.

Em outras palavras: todo objeto que possui os atributos de P possui os atributos de Q.

Base de dados considerada

Materiais e métodos

- Extraída do repositório da RefSeq [4] da NCBI (*National Center for Biotechnology Information*) em 22 de Abril de 2014.
- Os dados extraídos são referente a 4 organismos *Rattus novergicus* (1383 moléculas), *Mus musculus* (1097 moléculas), *Homo sapiens* (21528 moléculas)e *Drosophila melanogaster* (27764 moléculas).
- Cada molécula é identificada de acordo com o nível de inspeção, e classificadas como: *Model, Inferred, Predicted, Provisional, Reviewed, Validated e WGSk*.

Base de dados considerada

Materiais e métodos

PUC Minas
Instituto de Ciências Exatas e Informática

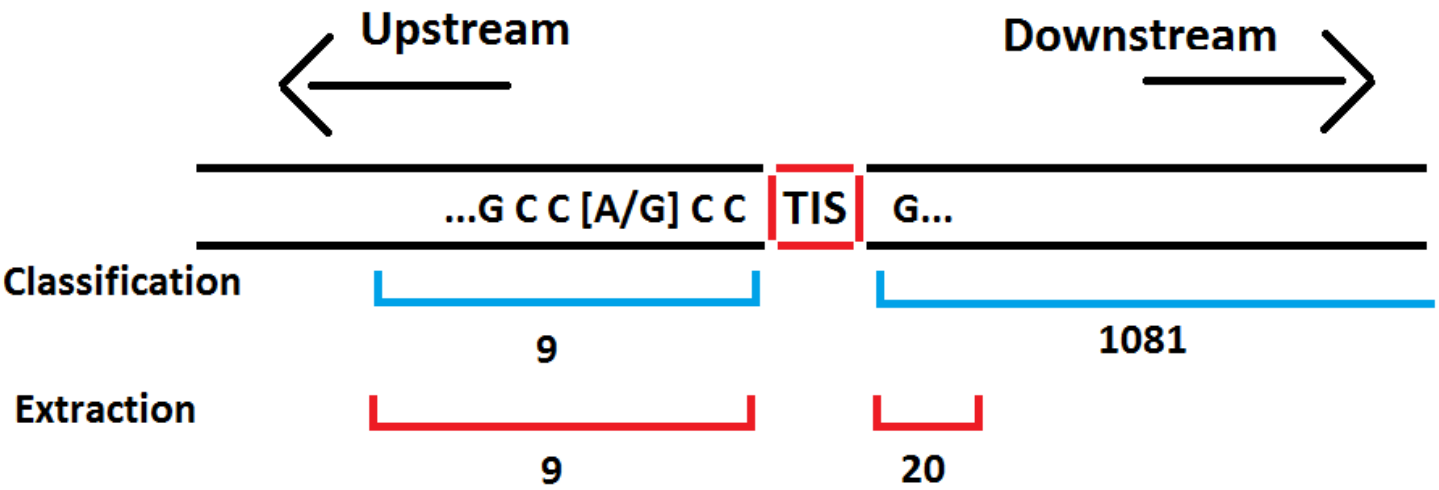
Tabela: Quantidade de sequencias de cada base

	Positivos	**Negativos**	**Total**
Rattus Novergicus	66	101	167
Mus musculos	398	632	1030
Homo sapiens	9716	16085	25801
Drosophila melanogaster	10122	25725	35847

Definição tamanho da janela

Materiais e métodos

Figura: Definição do Tamanho da janela



Definição tamanho da janela

Materiais e métodos

- Influência direta na qualidade do modelo de predição[5, 3].
- Janelas de tamanho assimétrico proporcionam uma maior acurácia [5].
- Modelo de escaneamento do ribossomo e o consenso de [1] que identifica um padrão conservativo nas posições:

-6	-5	-4	-3	-2	-1	$+1$	$+2$	$+3$	$+4$
G	C	C	$[A/G]$	C	C	A	T	G	$[G]$

- Foi utilizada por [6], onde identificou-se também a conservação da posição -7.
- Para a região *downstream*, foi verificado em [2] que quando maior esta região maior é a acurácia alcançada pelo classificador SVM, adotando assim o tamanho de 1081 nucleotídeos na região *downstream*.

Find Implications

Dado um contexto formal (X, Y), o algoritmo procura implicações de (X, Y) onde existe implicações $A \rightarrow B$, com $A \bigcap B = \emptyset$, $A \subset Y, B = Y - A$, de tal forma que esta implicação não possa ser obtida do conceito (W,Z).

Complexidade

$$O(|C|k^2|M|q)$$

C é o numero de conceito,

k é o maior numero de atributos na premissa,

M é o numero de atributos

q é o maior numero de relação por conceito.

Classificador SVM
Materiais e métodos

PUC Minas
Instituto de Ciências Exatas e Informática

- É uma técnica de aprendizado de máquina, capaz de resolver problemas de classificação lineares e não lineares,
- Faz separação de exemplos utilizando uma superfície de decisão linear e aumentando a distancia entre os pontos de treinamento.[5]

Parametros do SVM

Materiais e métodos

PUC Minas
Instituto de Ciências Exatas e Informática

Função *kernel* gaussiana RBF (*Radial Basis Function*) definida pela equação 1 e seu parâmetro é representado pelo símbolo *gamma* (γ).

$$K(x_i, x_j) = \exp(-\gamma \|x - x'\|^2) \tag{1}$$

$$Precisao = 100 \cdot \frac{TP}{TP + FP} \tag{2}$$

$$Sensibilidade = 100 \cdot \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 \cdot \frac{Precisao \cdot Sensibilidade}{Precisao + Sensibilidade} \tag{4}$$

Cross-validation

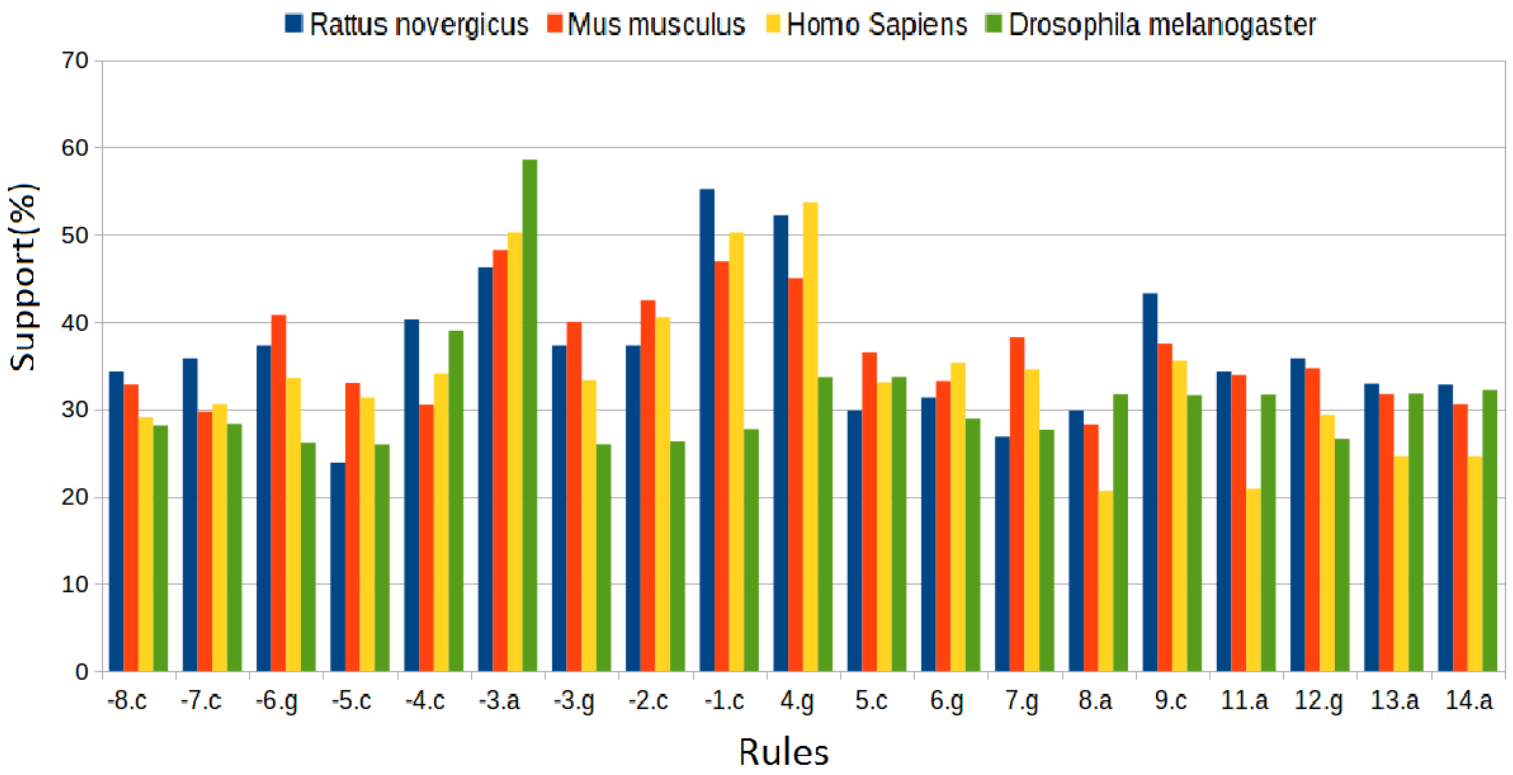
Consiste em subdividir o conjunto de dados disponível em 10 dobras de mesmo tamanho, sendo que 9 destas dobras são utilizadas para o treinamento e 1 dobra para a validação.

Resultados

Tabela: Regras de implicações extraídas e seu suporte

	Rattus novergicus		*Mus musculus*		*Homo Sapiens*		*Drosophila melanogaster*		Geral	
Premissa	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
-8.c	34,3%	21,6%	32,8%	22,3%	29,1%	26,7%	28,1%	21,7%	31,1%	23,0%
-7.c	35,8%	27,5%	29,7%	21,1%	30,6%	25,7%	28,3%	21,6%	31,1%	24,0%
-6.g	37,3%	17,6%	40,8%	31,3%	33,6%	28,7%	26,1%	21,2%	34,5%	24,7%
-5.c	23,9%	28,4%	33,0%	28,5%	31,3%	25,4%	25,9%	19,9%	28,5%	25,6%
-4.c	40,3%	18,6%	30,5%	25,4%	34,0%	24,8%	39,1%	18,5%	35,9%	21,8%
-3.a	46,3%	23,5%	48,2%	26,9%	50,2%	25,4%	58,6%	26,5%	50,8%	25,5%
-3.g	37,3%	27,5%	40,1%	25,7%	33,3%	28,3%	26,0%	18,5%	34,1%	25,0%
-2.c	37,3%	16,6%	42,5%	22,3%	40,5%	21,9%	26,3%	20,0%	36,7%	20,2%
-1.c	55,2%	25,5%	46,9%	23,5%	50,2%	27,7%	27,7%	17,3%	45,0%	23,3%
4.g	52,2%	33,3%	45,0%	24,6%	53,7%	28,5%	33,6%	17,4%	46,1%	25,9%

Figura: Valores de Suporte de cada regra



Resultados

PUC Minas
Instituto de Ciências Exatas e Informática

Figura: Vetores suporte adicionados

1083	1084	V(-8.C)	V(-7.C)	V(-6.G)	V(-5.C)	V(-4.C)	V(-3.Purin	V(-2.C)	V(-1.C)	V(4.G)	V(5.C)	V(6.G)	V(7.G)	V(8.A)	V(9.C)	V(11.A)	V(12.G)	V(13.A)	V(14.A)	Tis-Class
a	t	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	1	TIS
a	g	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	1	TIS
g	g	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1	TIS
t	g	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	1	TIS
c	g	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	1	TIS
t	g	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	0	1	TIS

Resultados

Tabela: Parâmetros obtidos utilizando o método *Grid Search*

	Sem adicionar as características		Adicionado as características	
	C	*Gamma*	*C*	*Gamma*
Rattus norvegicus	128	$3.051757812 \times 10^{-5}$	32	$3.0517578125 \times 10^{-5}$
Mus musculus	8	$1.220703125 \times 10^{-4}$	8	4.8828125×10^{-4}
Homo sapiens	2	4.8828125×10^{-4}	128	4.8828125×10^{-4}
Drosophila melanogaster	32	4.8828125×10^{-4}	8	4.8828125×10^{-4}

Resultados

Tabela: Resultados libSVM

	Sem adicionar as características			Com as características		
	Presição	Sensibilidade	F-measure	Presição	Sensibilidade	F-measure
Rattus norvegicus	89.4%	89.2%	89.1%	89.4%	89.2%	89.1%
Mus musculus	97.9%	97.9%	97.8%	98.8%	98.7%	98.8%
Homo sapiens	98.03%	97.74%	97.89%	98.2%	98.2%	98.2%
Drosophila melanogaster	96.86%	96.8%	96.8%	98.2%	98.2%	98.2%

Conclusão

- Utilizado um janelamento da *downstream* de 20 nucleotídeos e 9 na *upstream* para a extração, conseguimos encontramos um total de 19 regras de implicação incluindo as regras de Kosak.

- Reforçar o acontecimento das regras de implicações que formam as sequências SIT

Conclusão

Em trabalhos futuros deve-se analisar o uso de janelas maiores tanto na *upstream* quanto na *downstream* para a extração de regras, assim dano um maior suporte para as sequências positivas da base de treinamento com janelas bem maiores.

Referencias

[1] Marilyn Kozak.
Compilation and analysis of sequences upstream from the translational start site in eukaryotic mrnas.
Nucleic Acids Research, 12(2):857–872, 1984.

[2] C. P. Lacerda.
Aprendizado transdutivo aplicado à predição de sítio de início de tradução de proteínas.
Master's thesis, Pontifíca Universidade Católica de Minas Gerais, Belo Horizonte, MG, Brasil, 2015.

[3] HUIQING LIU and LIMSOON WONG.
Data mining tools for biological sequences.
Journal of Bioinformatics and Computational Biology, 01(01):139–167, 2003.

[4] Kim D. Pruitt and Donna R. Maglott.
Refseq and locuslink: Ncbi gene-centered resources.
Nucleic Acids Research, 29(1):137–140, 2001.

[5] Lívia Márcia Silva, Felipe Carvalho de Souza Teixeira, José Miguel Ortega, Luis Enrique Zárate, and Cristiane Neri Nobre.
Improvement in the prediction of the translation initiation site through balancing methods, inclusion of acquired knowledge and addition of features to sequences of mrna.
BMC Genomics, 12(4):1–20, 2011.

[6] George Tzanis, Christos Berberidis, and Ioannis Vlahavas.
Mantis: a data mining methodology for effective translation initiation site prediction.
In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 6343–6347. IEEE, 2007.

[7] R. Wille.
Restructuring lattice theory: an approach based on hierarchies of concepts.
I. Rival (Ed.): Ordered Sets, pages 445–470, 1982.

Obrigado