CLASSIFICAÇÃO DE CÉLULAS CANCERÍGENAS UTILIZANDO ALGORITMO GENÉTICO MULTI-POPULACIONAL

José Eurípedes Ferreira de Jesus Filho

Universidade de São Paulo – IME – jeferreirajf@gmail.com

AGENDA

- O problema
- Proposta do trabalho
- Algoritmo Genético
  - Codificação
  - Aptidão
  - Recombinação
  - Mutação
  - Hierarquia
  - Multi-populações
- Resultados computacionais
- Conclusões e trabalhos futuros
- Referências

O PROBLEMA

- Amaral (2007) utilizou a base NC160 de expressões gênicas para construir regras do tipo IF-THEN para classificação de células cancerígenas.
- A base NC160 foi criada a partir da colaboração de dois laboratórios, *Brown/Bolstein* e *Laboratory of Developmental Therapeutics*, nos EUA.
- A base possui a expressão gênica de mais de 8000 genes, sendo 61 amostras de 9 tipos de câncer diferentes: 7 de mama (C_1), 6 do sistema nervoso central (C_2), 7 de cólon (C_3), 6 de leucemia (C_4), 8 de melanoma (C_5), 9 de pulmão (C_6), 6 de ovário (C_7), 8 do sistema renal (C_8) e 4 de células reprodutivas (C_9).

O PROBLEMA

- Através de vários trabalhos da literatura, os mais de 8000 genes da base NC160 foram reduzidos.
- B_1: Reduzida para 13 genes através da aplicação de um Algoritmo Genético (AG) em conjunto com um classificador de máxima verossimilhança.
- B_2: Reduzida para 20 genes através da aplicação de um método *between-group/within-group*.
- B_3: Reduzida para 17 genes através da aplicação de um método *signal-to-noise/one-vs.-all*.
- B_4: Reduzida para 13 genes através da aplicação de uma variação do método utilizado na redução da base B_1.
- Assim, B_1 e B_4 foram gerados através de métodos AG/MV e B_2 e B_3 foram gerados a partir de métodos de *ranking*.

PROPOSTA DO TRABALHO

- Modificar o AG proposto por Amaral (2007), acrescentando o uso de multi-população e hierarquia estruturada para a construção das regras de alto nível.
- Utilizar o mesmo modelo da construção das regras IF-THEN.
- Perspectiva de melhorar os resultados por causa de trabalhos anteriores que trabalhavam com codificações tão complexas quanto as de Amaral.

ALGORITMO GENÉTICO - CODIFICAÇÃO

- Cada indivíduo é composto por n alelos e representa uma regra.
- Cada alelo $i, i = 1, \ldots, n$, possui o formato (I_i, P_i, O_i, V_i).
- I_i: Índice do alelo i em relação a posição na base de dados original.
- P_i: Possui valor entre 0 e 10. Caso seja maior que 7, o termo referente ao alelo é incluído na regra final.
- O_i: Indica o operador do termo, que pode ser $\geq$ ou $<$.
- V_i: Valor entre o menor e o maior valor de expressão gênica do gene I_i.

6 | 2 | < | 0,315	14 | 8 | ≥ | -0,812	42 | 9 | ≥ | 1,150	88 | 0 | < | -1,653	111 | 5 | ≥ | -0,451

if 14 ≥ -0, 812 *AND* 42 ≥ -1, 150 *then*

ALGORITMO GENÉTICO - APTIDÃO

- Aptidão baseada em dois indicadores comumente utilizados em domínios médicos: *sensibilidade* e *especificidade*.
- Sensibilidade (Se):

$$Se = \frac{tp}{tp + fn}.$$

- Especificidade (Es):

$$Es = \frac{tn}{tn + fp}.$$

- Aptidão (Apt):

$$Apt = Se \times Es.$$

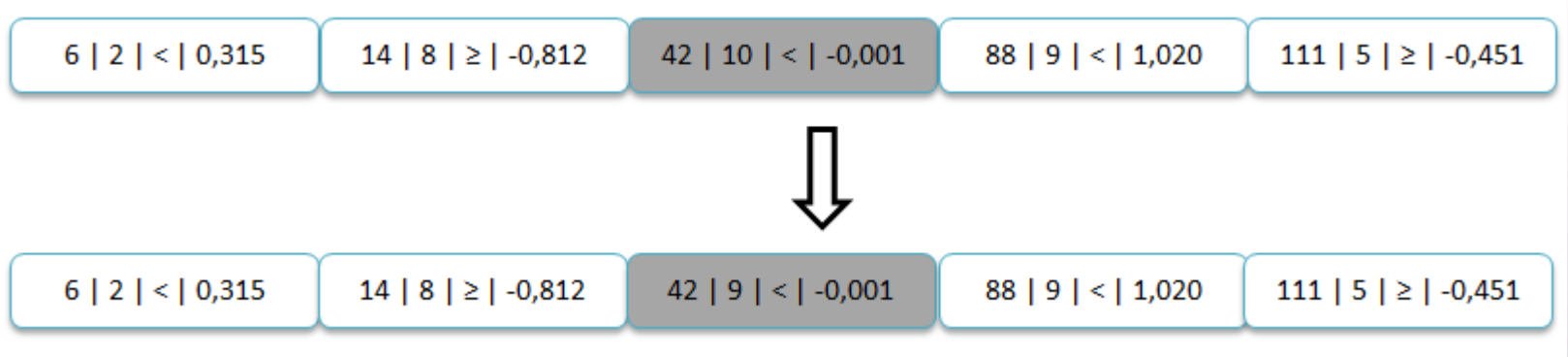
ALGORITMO GENÉTICO - RECOMBINAÇÃO

- *Crossover* de 2 pontos.

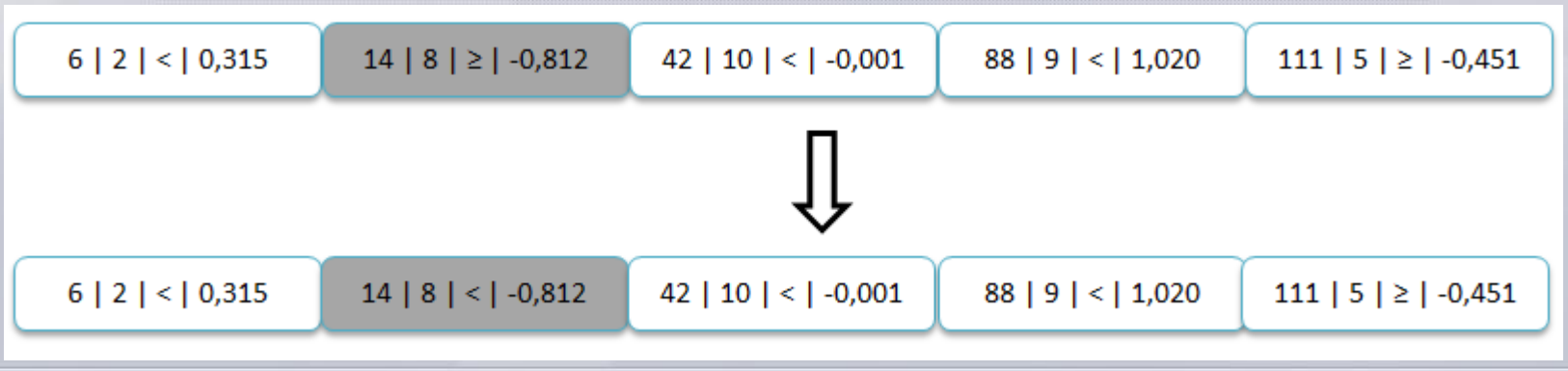
6 | 2 | < | 0,315	14 | 8 | ≥ | -0,812	42 | 9 | ≥ | 1,150	88 | 0 | < | -1,653	111 | 5 | ≥ | -0,451
6 | 0 | ≥ | 0,420	14 | 2 | < | 0,111	42 | 10 | < | -0,001	88 | 9 | < | 1,020	111 | 10 | < | 3,975
6 | 2 | < | 0,315	14 | 8 | ≥ | -0,812	42 | 10 | < | -0,001	88 | 9 | < | 1,020	111 | 5 | ≥ | -0,451

ALGORITMO GENÉTICO - MUTAÇÃO

- P_i: Acréscimo ou decréscimo em uma unidade.

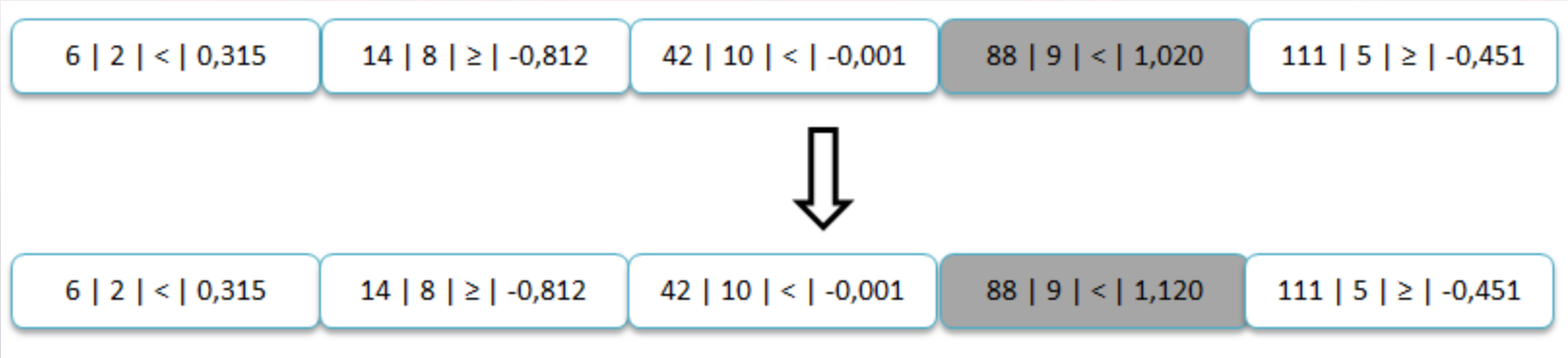


- O_i: Inversão no operador utilizado.

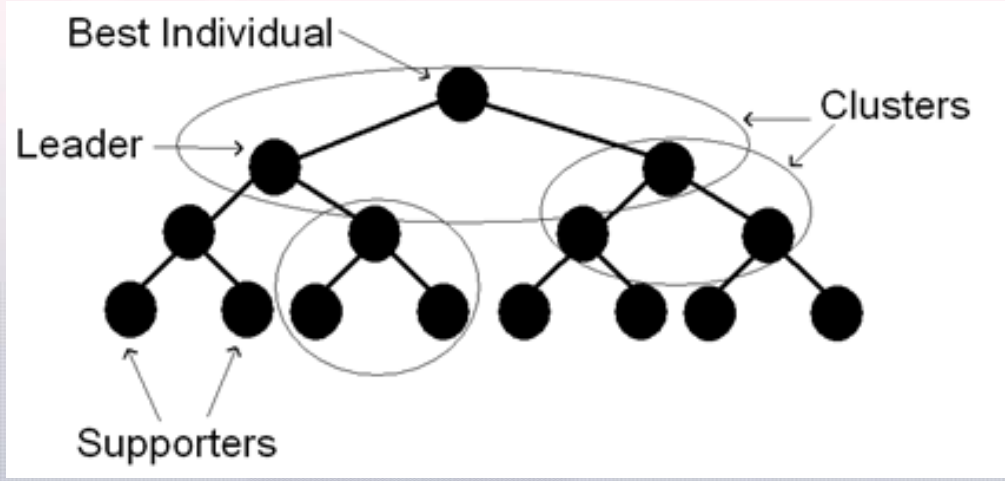


ALGORITMO GENÉTICO - MUTAÇÃO

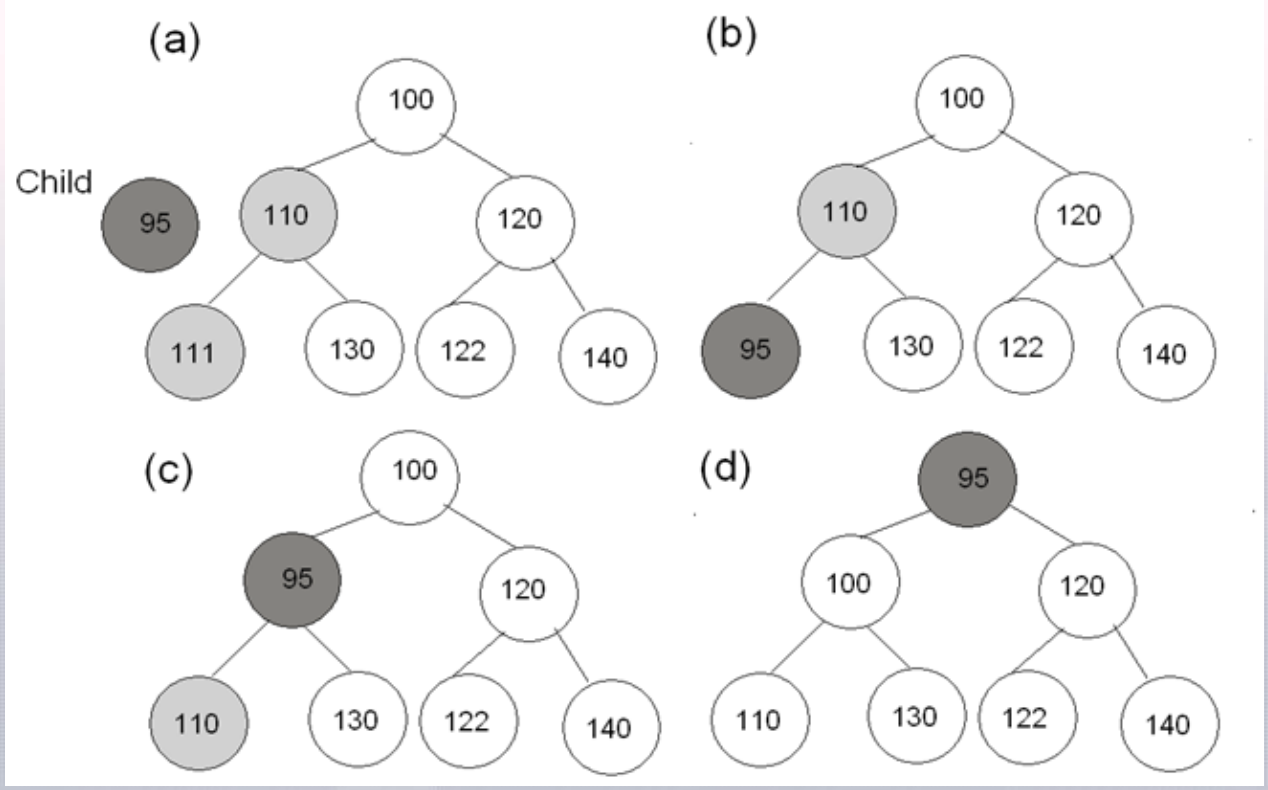
- V_i: Acréscimo ou decréscimo em 0,1.



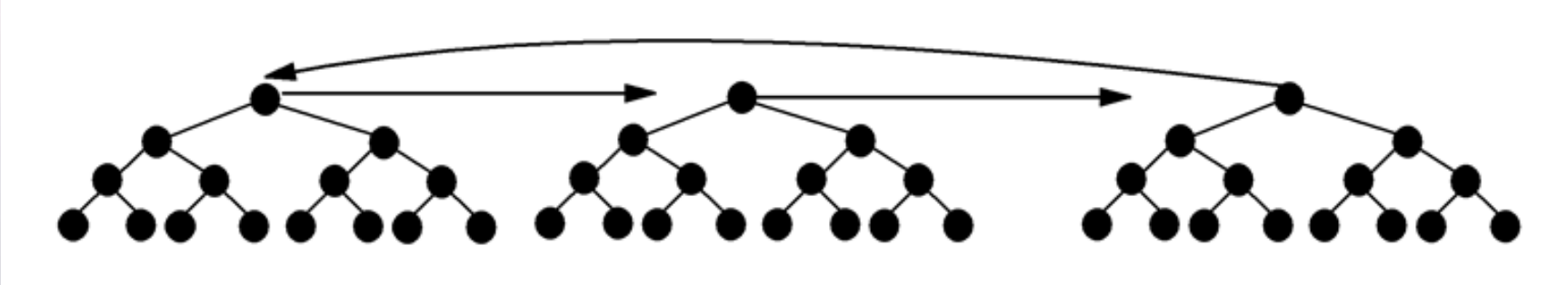
ALGORITMO GENÉTICO - HIERARQUIA



ALGORITMO GENÉTICO - HIERARQUIA



ALGORITMO GENÉTICO – MULTI-POPULAÇÕES



RESULTADOS COMPUTACIONAIS

- Os testes foram realizados em um computador Intel Core Quad 9400, 2.66 GHz com 4GB de RAM DDR2 800 MHz.
- O AG proposto foi implementado em C++ e os parâmetros do AG foram: taxa de recombinação de 20; taxa de mutação de 0.9; 3 populações com 15 indivíduos cada.
- Para cada classe, o AG foi executado inicialmente 20 vezes e posteriormente 50 vezes. Abordagem defendida por conta da baixa quantidade de amostras nas bases.
- Cada base reduzida, $B_1 - B_4$, foi dividida em 3 partes, mantendo a proporção do número de amostras por classe. As duas primeiras partes foram utilizadas na fase de treinamento com o AG e a terceira parte na fase de teste (validação) da regra obtida pelo AG.

RESULTADOS COMPUTACIONAIS - B_1

	Amaral		AG 20 Exec			AG 50 Exec		
Classe	Apt_{trein}	Apt_{teste}	Apt_{trein}	Apt_{teste}	CPU(s)	Apt_{trein}	Apt_{teste}	CPU(s)
C_1	0,80	0,00	1,00	0,00	0,05	1,00	0,00	0,07
C_2	1,00	1,00	1,00	1,00	0,04	1,00	1,00	0,03
C_3	1,00	0,50	**1,00**	**1,00**	**0,08**	**1,00**	**1,00**	**0,02**
C_4	1,00	1,00	1,00	0,88	0,05	1,00	1,00	0,05
C_5	1,00	1,00	1,00	1,00	0,06	1,00	1,00	0,05
C_6	1,00	0,00	**1,00**	**0,58**	**0,10**	1,00	0,53	0,05
C_7	1,00	0,00	**1,00**	**0,94**	**0,04**	1,00	0,50	0,05
C_8	1,00	0,67	**1,00**	**1,00**	**0,07**	**1,00**	**1,00**	**0,03**
C_9	1,00	0,90	**1,00**	**0,94**	**0,02**	**1,00**	**0,94**	**0,02**
Média	0,98	0,56	**1,00**	**0,81**		1,00	0,77	

RESULTADOS COMPUTACIONAIS - B_2

	Amaral		AG 20 Exec			AG 50 Exec		
Classe	Apt_{trein}	Apt_{teste}	Apt_{trein}	Apt_{teste}	**CPU(s)**	Apt_{trein}	Apt_{teste}	**CPU(s)**
C_1	0,94	0,47	**1,00**	**0,50**	**0,38**	**1,00**	**0,50**	**13,83**
C_2	1,00	0,00	**1,00**	**1,00**	**0,08**	1,00	0,94	0,02
C_3	1,00	0,50	1,00	0,94	0,06	**1,00**	**1,00**	**0,06**
C_4	1,00	0,00	**1,00**	**1,00**	**0,02**	**1,00**	**1,00**	**0,06**
C_5	1,00	0,94	**1,00**	**1,00**	**0,53**	**1,00**	**1,00**	**0,07**
C_6	1,00	0,43	**1,00**	**1,00**	**2,49**	1,00	0,93	1,72
C_7	**1,00**	**0,50**	0,97	0,50	0,07	0,97	0,50	0,04
C_8	1,00	0,67	1,00	0,94	1,15	**1,00**	**1,00**	**0,14**
C_9	1,00	0,95	**1,00**	**1,00**	**0,08**	**1,00**	**1,00**	**0,02**
Média	0,99	0,50	**1,00**	**0,88**		1,00	0,87	

RESULTADOS COMPUTACIONAIS - B_3

	Amaral		AG 20 Exec			AG 50 Exec		
Classe	Apt_{trein}	Apt_{teste}	Apt_{trein}	Apt_{teste}	**CPU(s)**	Apt_{trein}	Apt_{teste}	**CPU(s)**
C_1	1,00	0,00	**1,00**	**0,38**	**0,42**	1,00	0,00	0,02
C_2	1,00	1,00	1,00	1,00	0,05	1,00	1,00	0,02
C_3	1,00	0,47	1,00	0,81	0,09	**1,00**	**0,94**	**0,07**
C_4	1,00	0,00	**1,00**	**1,00**	**0,03**	**1,00**	**1,00**	**0,00**
C_5	1,00	1,00	1,00	1,00	0,05	1,00	1,00	0,02
C_6	1,00	0,67	**1,00**	**0,87**	**0,17**	1,00	0,62	0,11
C_7	1,00	0,47	**1,00**	**1,00**	**0,05**	**1,00**	**1,00**	**0,06**
C_8	1,00	0,67	0,97	1,00	0,22	0,97	1,00	0,11
C_9	1,00	0,00	**1,00**	**0,94**	**0,05**	1,00	0,88	0,01
Média	1,00	0,48	**1,00**	**0,89**		1,00	0,83	

RESULTADOS COMPUTACIONAIS - B_4

	Amaral		AG 20 Exec			AG 50 Exec		
Classe	Apt_{trein}	Apt_{teste}	Apt_{trein}	Apt_{teste}	**CPU(s)**	Apt_{trein}	Apt_{teste}	**CPU(s)**
C_1	0,92	0,44	**1,00**	**0,50**	**0,10**	**1,00**	**0,50**	**0,05**
C_2	1,00	1,00	1,00	1,00	0,02	1,00	1,00	0,02
C_3	1,00	0,78	1,00	0,88	0,06	**1,00**	**1,00**	**0,09**
C_4	1,00	0,44	1,00	0,94	0,02	**1,00**	**1,00**	**0,05**
C_5	1,00	1,00	1,00	1,00	0,07	1,00	1,00	0,02
C_6	1,00	0,31	1,00	0,53	0,19	**1,00**	**0,62**	**0,38**
C_7	**1,00**	**0,50**	1,00	0,00	0,03	1,00	0,41	0,05
C_8	1,00	0,67	**1,00**	**1,00**	**0,04**	**1,00**	**1,00**	**0,04**
C_9	0,97	1,00	1,00	0,88	0,00	1,00	0,94	0,02
Média	0,99	0,68	1,00	0,75		**1,00**	**0,83**	

CONCLUSÕES E TRABALHOS FUTUROS

- Considerando as quatro bases reduzidas de expressões gênicas, o AG multi-populacional proposto superou os resultados de Amaral (2007) em 24 das 36 regras. Igualou os resultados em 8 regras e perdeu em apenas 2 regras.
- As 2 regras restantes são inconclusivas uma vez que ou o resultado da fase de treinamento ou o resultado da fase de teste é levemente inferior.
- Como trabalho futuro, sugere-se a adição de algum outro critério na avaliação da aptidão.
- Além disso, seria interessante executar mais testes do método, como outras composições das frações da base para treinamento e avaliação.

REFERÊNCIAS

- L. R. Amaral, **Mineração de Regras para Classicação de Oncogenes Medidos por Microarray Utilizando Algoritmos Genéticos**, *Dissertação de Mestrado*, UFU, Brasil, 2007.

- D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jerey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein and P. O. Brown, **Systematic variation in gene expression patterns in human cancer cell lines**, *Nature Genetics*, 2000.