CLASSIFICAÇÃO DE PROTEÍNAS COM REDES NEURAIS ARTIFICIAIS

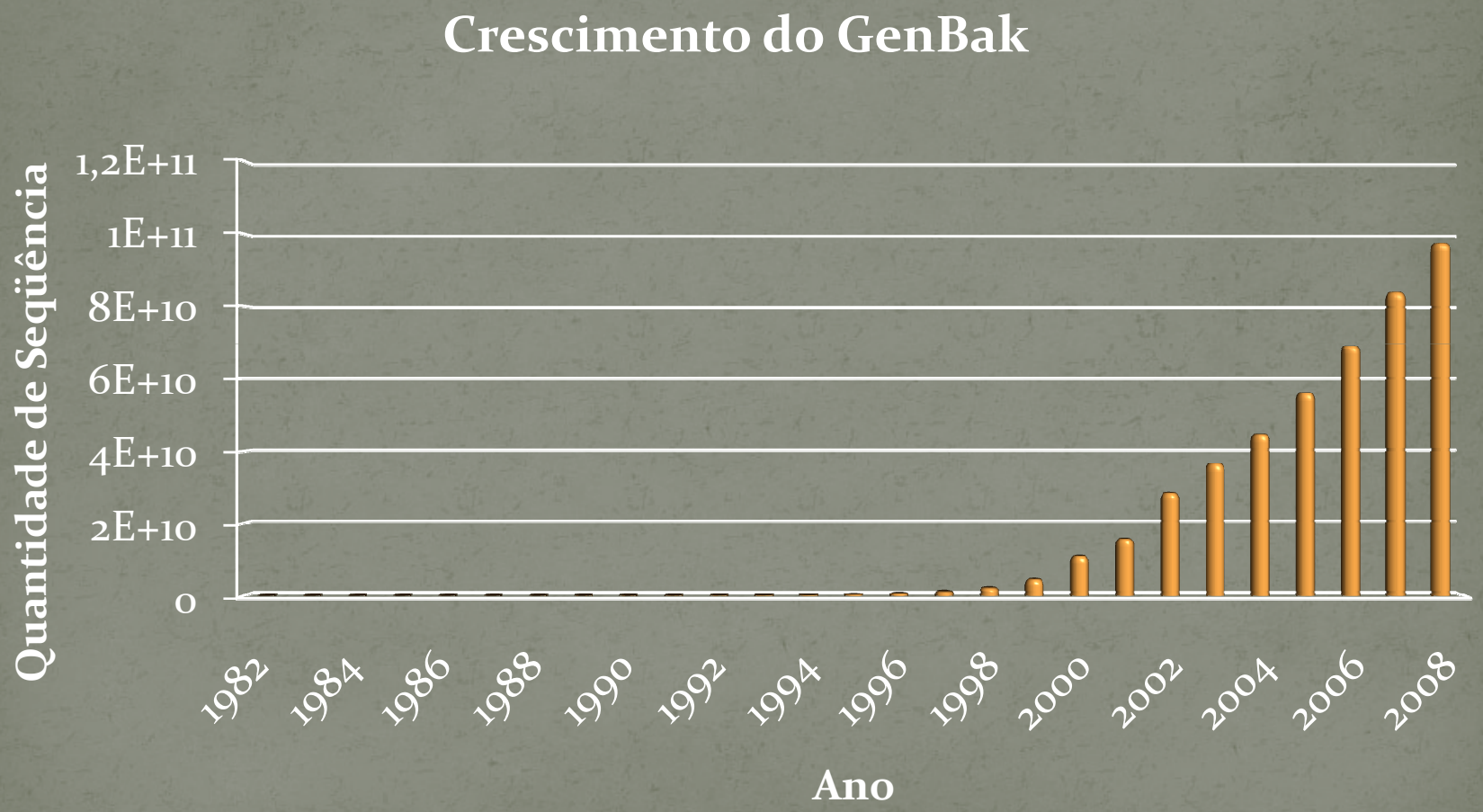
Tiago Amador Coelho

Orientador : Dr. Thiago de Souza Rodrigues

Roteiro

- Motivação
- Proteínas
- Banco de Dados COG
- Redes Neurais Artificiais
- Sequence Coding By Sliding Window (SCSW)
- Metodologia
- Resultados
- Conclusão
- Trabalhos Futuros

Motivação


Crescimento do GenBak
Quantidade de Seqüência
1,2E+11
1E+11
8E+10
6E+10
4E+10
2E+10
0
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
Ano

Motivação

- A partir do seqüenciamento do genoma, a geração de dados tem como objetivo, a predição do conjunto de proteínas existentes no organismo e a funcionalidade que cada proteína desempenha, para melhor entender o funcionamento do organismo

- Existem dois métodos que podem ser seguidos:
  - Laboratorial
  - Computacional

Objetivo

- Construção de um Classificador de Proteínas utilizando RNA
  - Classificar as seqüências ainda não classificadas
  - Verificar se seqüências já anotadas, podem ser reclassificadas pelo surgimento de um novo domínio

Proteínas

- O que são proteínas?
- Qual a composição das proteínas?
- Organização Estrutural das Proteínas
  - Estrutura Primária
  - Estrutura Secundária
  - Estrutura Terciária
  - Estrutura Quaternária

Proteínas – Estrutura Primária

primary structure
(amino acid sequence)

Banco de Dados COG

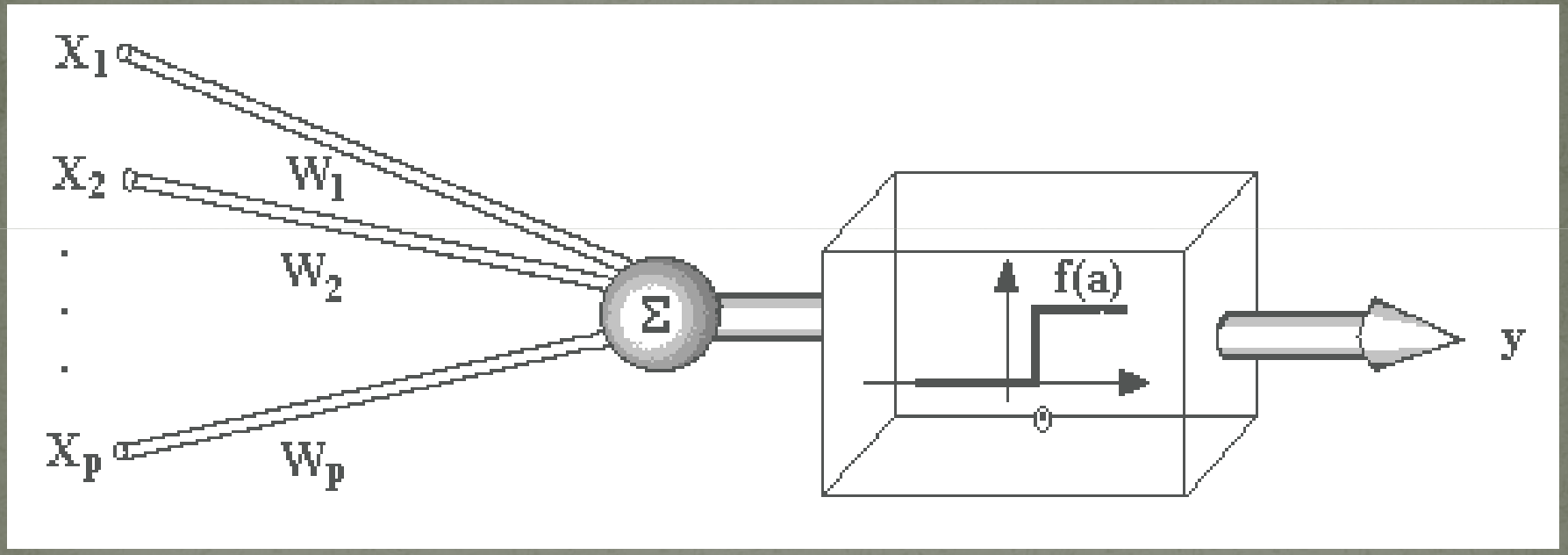
- NCBI - *National Center for Biotechnology Information*
  - Responsável por alguns bancos de dados públicos
  - Fonte de informação para a biologia molecular
  - Conduz pesquisas em biologia computacional
  - Desenvolve ferramentas
- COG - *Cluster of Orthologous Groups*
  - Tentativa de classificação filogenética das proteínas
  - As proteínas contidas são classificadas em 28 categorias funcionais

Redes Neurais Artificiais - RNA

- O que é RNA?
- Como funciona a RNA?

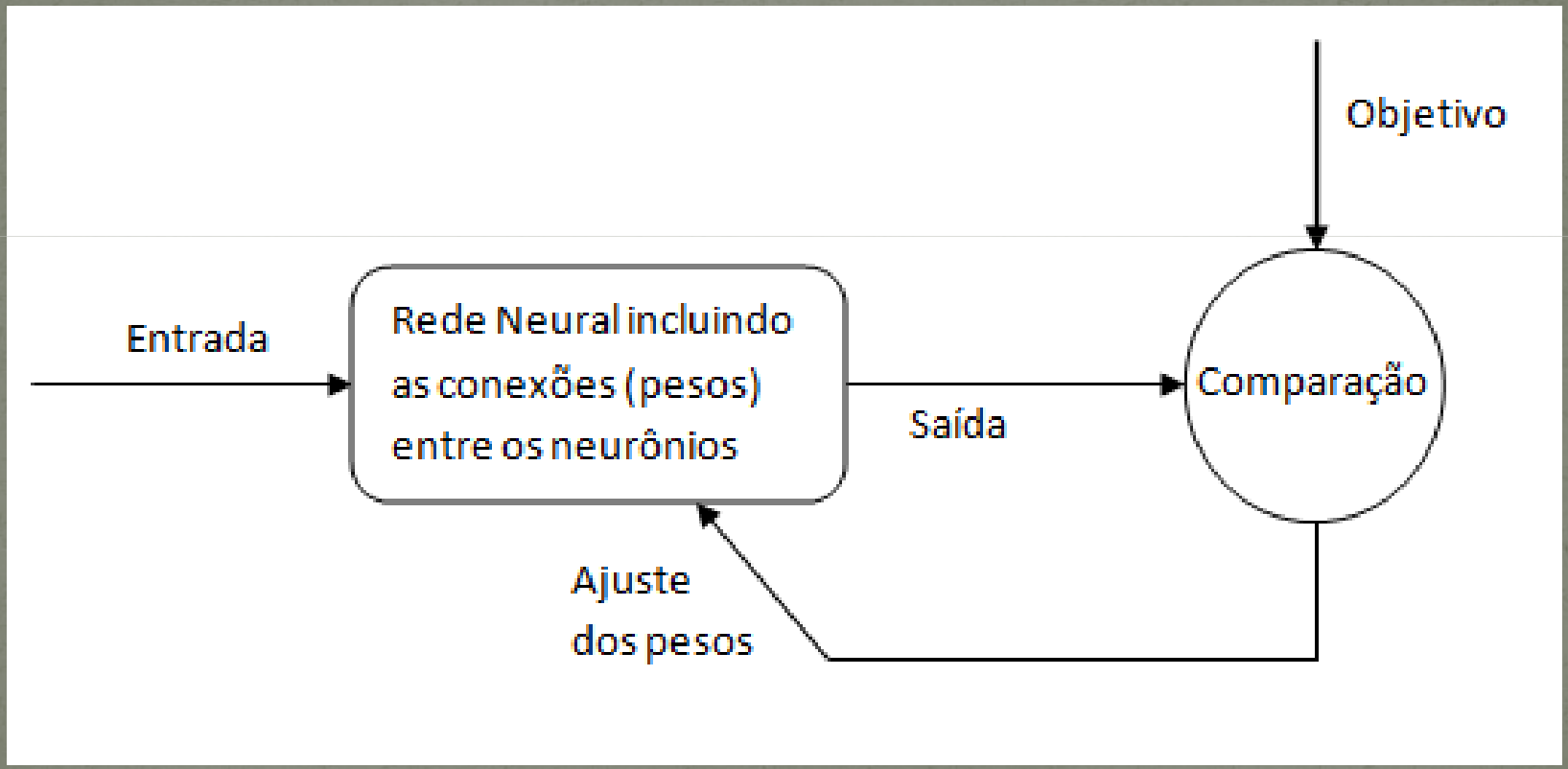
Redes Neurais Artificiais - RNA

Modelo de McCulloch e Pitts



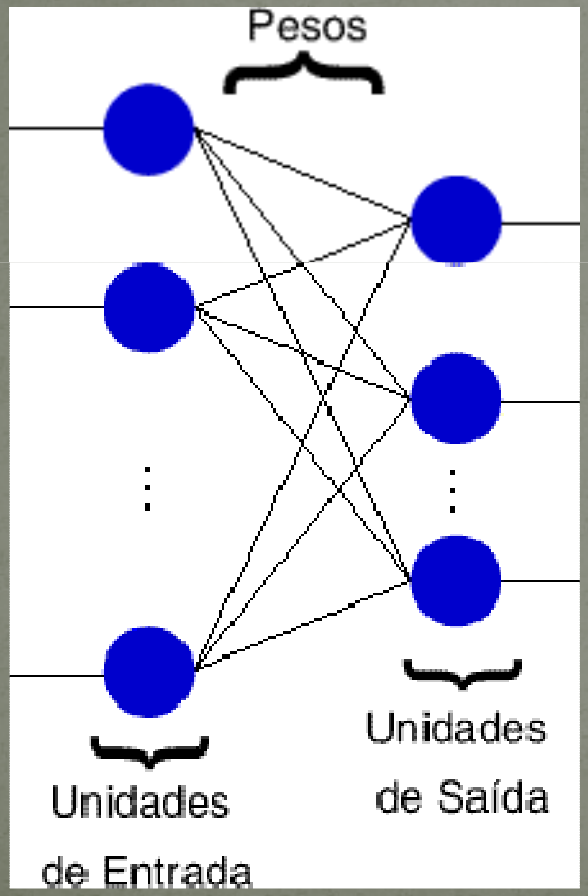
Redes Neurais Artificiais - RNA

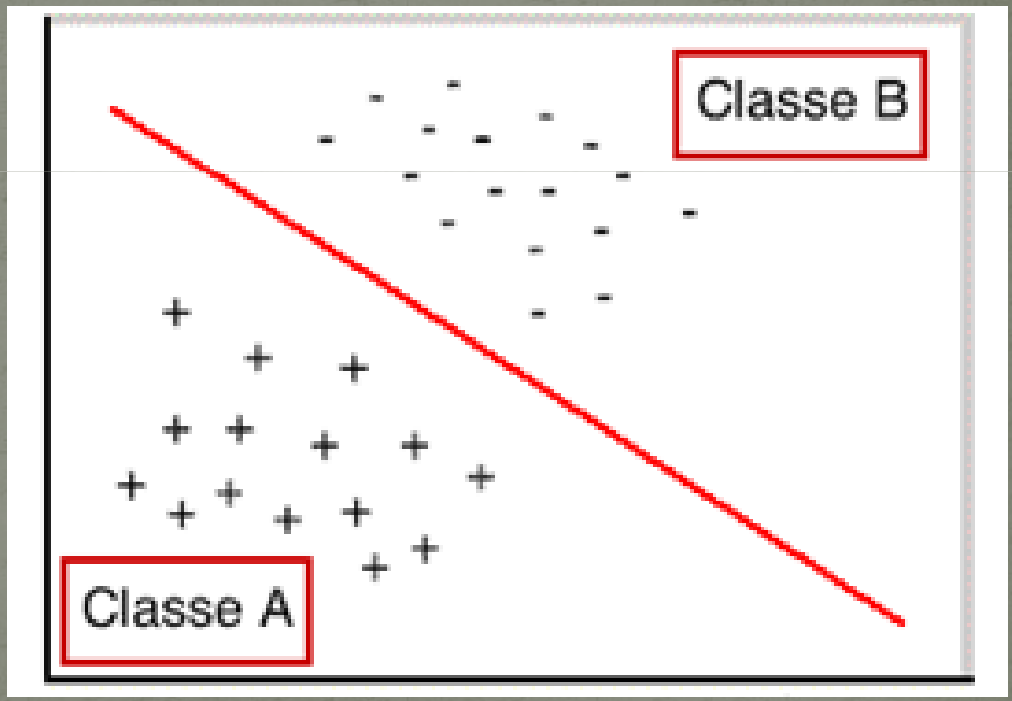
Treinamento



Redes Neurais Artificiais - RNA

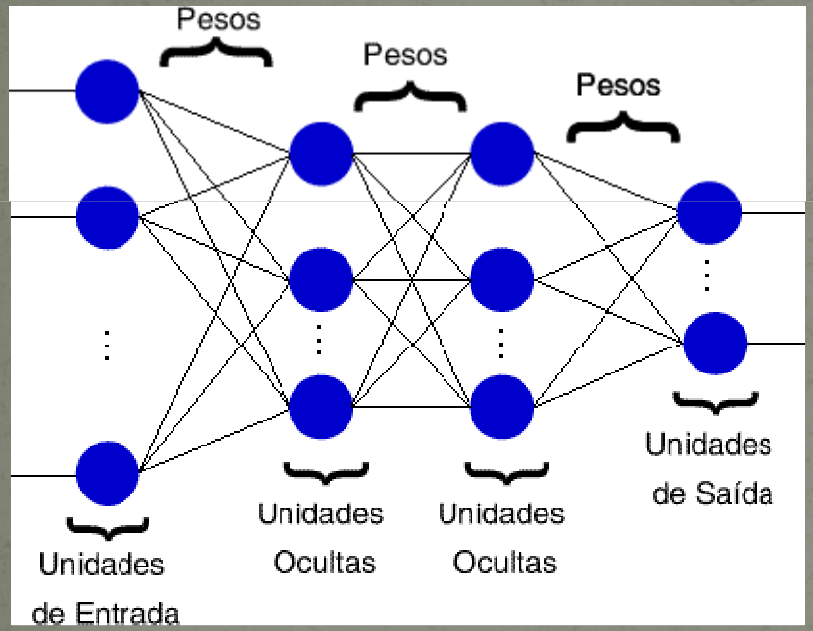
Perceptron (Rosemblatt 1958)

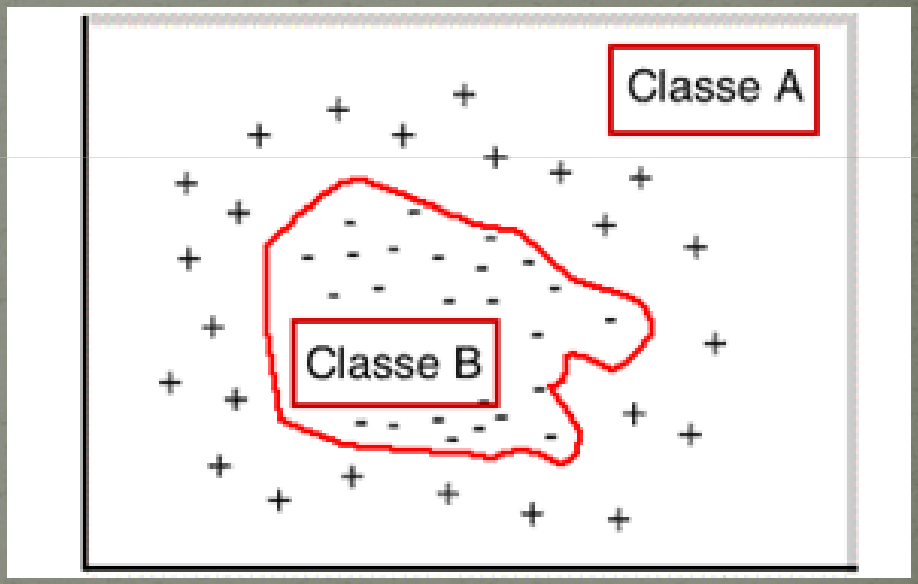




Redes Neurais Artificiais - RNA

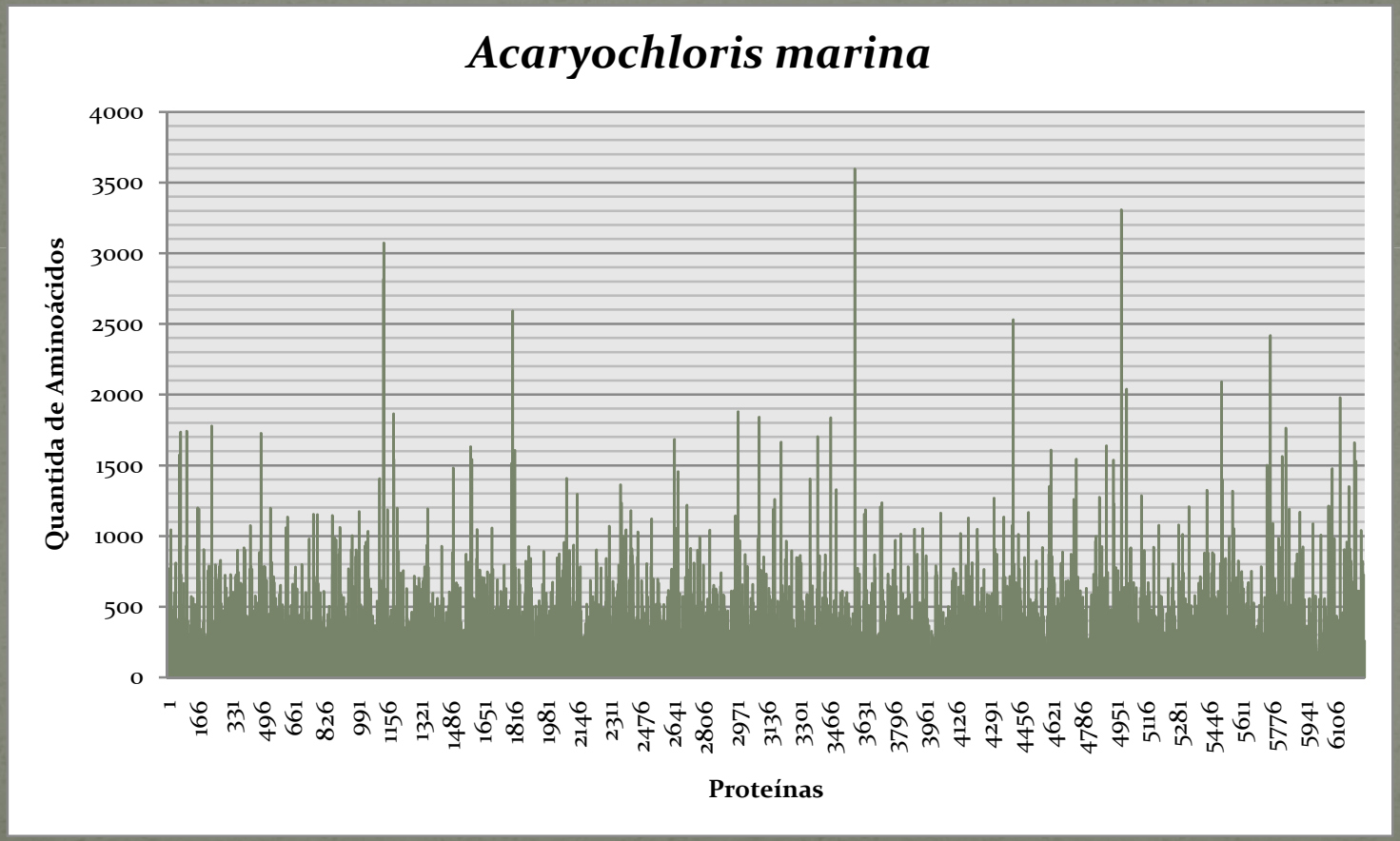
Multilayer Perceptron





Redes Neurais Artificiais - RNA

- Problema na Aplicação das RNA à Bioinformática



Sequence Coding By Sliding Window (SCSW)

- Dado uma seqüência S de tamanho N definida sobre um alfabeto α;
- Uma janela deslizante W_n de tamanho $1 \leq n \leq N$ é posicionada na posição 1 da seqüência S e vai sendo deslocada até a posição $N - n + 1$;
- Um vetor V_n de dimensão α^n é definido, onde cada posição corresponde a uma possível n – tupla dos elementos de α;

Sequence Coding By Sliding Window (SCSW)

- A cada deslocamento de W_n em S a posição de V_n correspondente à n – tupla encontrada é incrementada de 1;
- Após W_n atingir a posição N – n + 1 em S, o vetor V_n conterá a quantidade de cada n – tupla da seqüência percorrida e, independentemente do tamanho da seqüência, o vetor V_n terá dimensão α^n.

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
o	o	o	o	o	o	o	o	o

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
0	1	0	0	0	0	0	0	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
0	1	0	0	0	1	0	0	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
0	1	0	0	0	1	1	0	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
1	1	0	0	0	1	1	0	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
1	1	1	0	0	1	1	0	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
1	1	1	0	0	1	1	1	0

ABCAACBC

Sequence Coding By Sliding Window (SCSW)

AA	AB	AC	BA	BB	BC	CA	CB	CC
1	1	1	0	0	2	1	1	0

ABCAACBC

Metodologia

- Obtenção dos dados
  - NCBI
  - *Mycoplasma hyopneumoniae*
  - *Acaryochloris marina*
- Arquivo Fasta
- Obtenção dos códigos das proteínas pertencente a cada classe

Metodologia

>gi|72080389|ref|YP_287447.1| F0F1 ATP synthase subunit A [Mycoplasma hyopneumoniae 7448]

MMDFFRDWNQPQLFTLFILVFLVIILSIIFFHIKKAKIDESPSAVVLFAESYLIFIDDL
VETAGEGYINKVKPYIFSLFTFFLLGNLLSLVGLEPISTSISVTLSLAFVSWFGIFVVG
AIYSRWKYLSEFAKNPLKIIGIPAPLISLSFRMYGNLISGSVLLLIIYSGVQWIYQKIP
LGTFGNFNLPIVLIFPPFLIYFDIVGSLIQSFIFVILTTSYWGMEVNQDEARLKINKKQ
LNLQKI

>gi|72080390|ref|YP_287448.1| F0F1 ATP synthase subunit C [Mycoplasma hyopneumoniae 7448]

MNSIVNFSQQLIQNFQEVSQKTVADSSNLKAFAYLGAGLAMIGVIGVGAGQGYAA
GKACDAIARNPEAQK QVFRVLVIGTAISETSSIYALLVALILIFVG

- Separação das proteínas pelas classes funcionais

Metodologia

- Redução do alfabeto

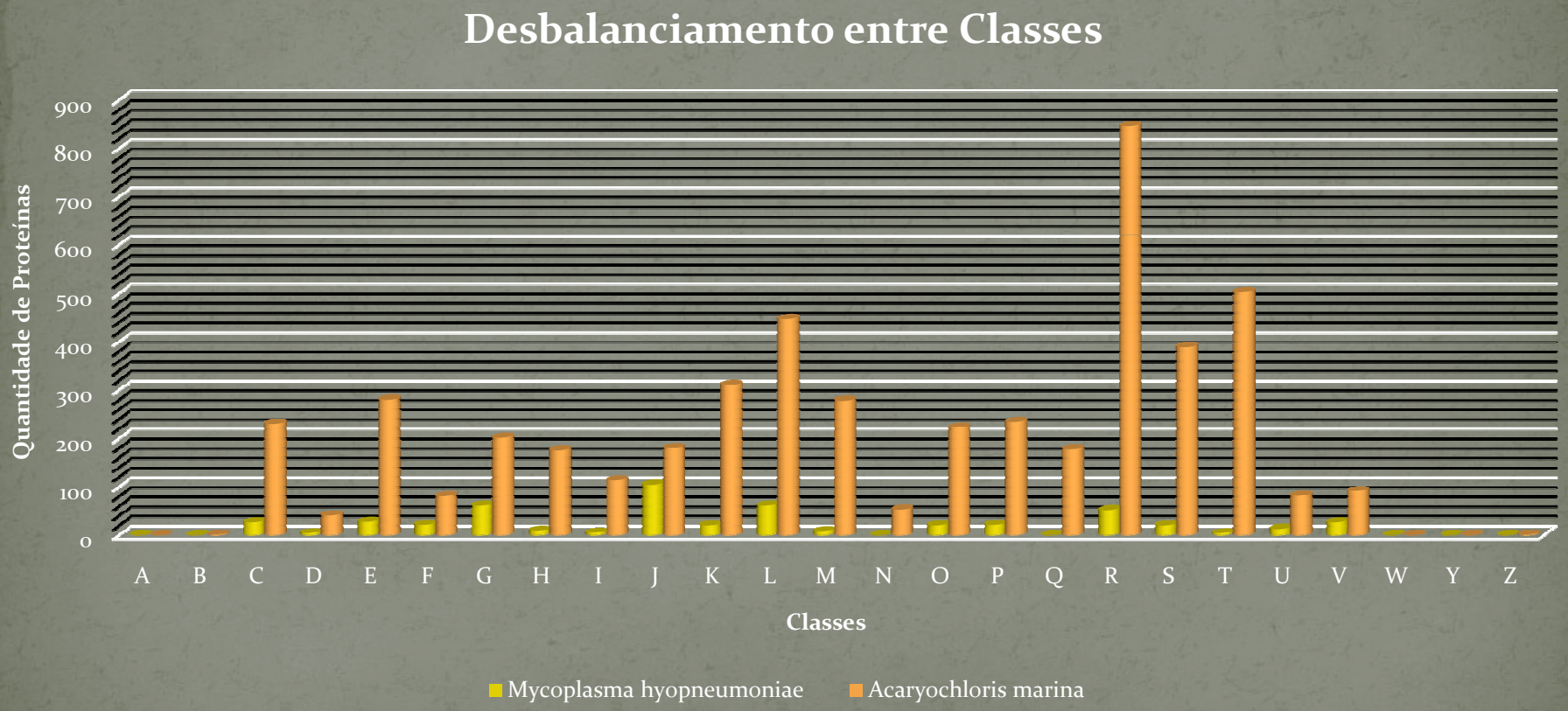
Exchange Group
H R K
C
F Y W
D N Q E
S T P A G
M I L V

Metodologia

- Codificação SCSW
  - Alfabeto de tamanho 6
  - Janela de tamanho 2
  - Vetor com dimensão 36

Metodologia

- Metodologia *One-Against-All*



Metodologia

- Seleção dos pontos da margem de Separação
  - *Condensed Nearest Neighbor* (CNN)
    - Distância Euclidiana entre todos os pontos
- Construção, Treinamento e Validação das Redes Neurais Artificiais
  - Uma RNA para cada classe de proteínas
  - Treinamento com os dados da *Mycoplasma hyopneumoniae*
  - Validação com os dados da *Acaryochloris marina*

Resultados

Dados de Entrada	Classificadores								
	C	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
Pertencente a classe (acerto)	85,65%	57,50%	0,00%	13,75%	0,00%	45,14%	71,68%	28,33%	10,61%
Pertencente a classe (erro)	5,65%	20,00%	69,29%	72,50%	100,00%	46,29%	21,24%	28,33%	74,92%
Não pertencente a classe (acerto)	12,15%	9,00%	66,92%	68,46%	95,86%	53,25%	24,40%	58,73%	81,75%
	L	**M**	**O**	**P**	**R**	**S**	**T**	**U**	**V**
Pertencente a classe (acerto)	0,00%	2,16%	22,52%	35,04%	7,57%	11,86%	6,97%	18,52%	1,10%
Pertencente a classe (erro)	38,03%	76,62%	50,45%	58,97%	36,76%	40,72%	60,96%	59,26%	96,70%
Não pertencente a classe (acerto)	72,72%	70,33%	55,20%	81,40%	33,99%	44,68%	48,83%	63,03%	94,00%

Conclusão

- Não obteve boa especificidade
- Boa generalização
- Tamanho da janela da metodologia SCSW

Trabalhos Futuros

- Utilização do modelo SCSW com outros tamanhos de janela;
- Utilização do modelo estendido da metodologia SCSW;
- Utilização do alfabeto de tamanho 20;
- Utilização de outros algoritmos de treinamento de RNAs.

Obrigado!