Assinado Digitalmente

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA ECONOMIA
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

CARTA PATENTE Nº PI 1106267-3

O INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL concede a presente PATENTE DE INVENÇÃO, que outorga ao seu titular a propriedade da invenção caracterizada neste título, em todo o território nacional, garantindo os direitos dela decorrentes, previstos na legislação em vigor.

(21) Número do Depósito: PI 1106267-3

(22) Data do Depósito: 12/08/2011

(43) Data da Publicação Nacional: 16/04/2013

(51) Classificação Internacional: G06F 19/16; G06F 19/24; G06N 3/02.

(30) Prioridade Unionista: US 12/855,366 de 12/08/2010.

(54) Título: MÉTODO, SISTEMA E APARELHO PARA PREDIZER E/OU RECONHECER E/OU CLASSIFICAR SEQUÊNCIAS BIOLÓGICAS

(73) Titular: FUNDAÇÃO UNIVERSIDADE DE CAXIAS DO SUL - UCS. CGC/CPF: 88648761000103. Endereço: Rua Francisco Getúlio Vargas, 1130 - Cidade Universitária, Caxias do Sul, RS, BRASIL(BR), 95.020-972

(72) Inventor: SCHEILA DE AVILA E SILVA; SERGIO ECHEVERRIGARAY LAGUNA; GÜNTHER JOHANNES LEWCZUK GERHARDT.

Código de Controle: 9204C444C084C881 75D4415BA8BA94E8

Prazo de Validade: 20 (vinte) anos contados a partir de 12/08/2011, observadas as condições legais

Expedida em: 18/08/2020

Assinado digitalmente por:
Liane Elizabeth Caldeira Lage
Diretora de Patentes, Programas de Computador e Topografias de Circuitos Integrados

Relatório Descritivo de Patente de Invenção

MÉTODO, SISTEMA E APARELHO PARA PREDIZER E/OU RECONHECER E/OU CLASSIFICAR SEQUÊNCIAS BIOLÓGICAS

Campo da Invenção

A presente invenção está no campo da bioinformática e biologia molecular. Especificamente, a presente invenção fornece um método, um sistema e um aparelho para predizer e/ou reconhecer e/ou classificar sequências biológicas, especialmente motivos de reconhecimento de sítio de ligação pouco conservados, compreendendo o uso de regras extraídas de processo de aprendizagem de redes neurais. A presente invenção é particularmente útil para a predição, reconhecimento e/ou classificação de promotores.

Antecedentes da Invenção

A predição e reconhecimento de promotores *in silico* é uma questão crucial em biologia molecular e um desafio em bioinformática. Os promotores são elementos cis-regulatórios (*cis-acting*) localizados antes do sítio de início da transcrição (TSS) da fase de leitura aberta (ORF). A expressão gênica começa com o reconhecimento do promotor pela enzima RNA polimerase (RNAP). Em bactérias, a holoenzima RNAP é composta por cinco subunidades (2 α, β, β', ω) e um fator de subunidade sigma (σ) adicional (Borukov e Nudler, 2003; Thiyagarajan e cols., 2005). A subunidade σ da RNAP é um regulador chave da expressão gênica de bactérias, porque é responsável pela interação específica de RNAP na região promotora. Os fatores σ controlam a iniciação da transcrição direcionando a ligação da RNAP para sequências promotoras específicas e dissociando ("melting") o DNA dupla-fita, assim, a transcrição de um determinado gene é dependente da σ associada à RNAP (Doucleff, 2007; Borukhov e Nudler, 2003; Hook-Barnard e cols., 2006).

Células bacterianas usam fatores σ alternativos, específicos para diversos subgrupos de promotores, a fim de se adaptar às mudanças

ambientais (Borukhov e Nudler, 2003). A *E. coli* possui vários fatores σ, os mais prevalentes são: σ24, σ28, σ32, σ38, σ54 e σ70 (o número indica o seu peso molecular). Cada família σ tem um papel na resposta bacteriana às condições ambientais e reconhece diferente sequência promotora consensual. Por exemplo, σ32 tem um papel na resposta ao choque térmico, σ28 está associado à expressão de genes flagelares durante o crescimento normal e σ70 é o principal fator responsável pela maior parte da atividade de transcrição na célula (Lewin 2008; Borukov e Nudler, 2003). Apesar da família, todos os promotores possuem dois importantes sítios de ligação para RNAP, a região -35 e -10 em relação aos nucleotídeos do TSS. Estes motivos são pouco conservados, particularmente entre famílias σ. O consenso canônico para as regiões -35 e -10 e o número de nucleotídeos entre elas são (Lewin, 2008):

σ32 - CCCTTGAA 13-15pb CCCGATNT

σ28 - CTAAA 15pb GCCGATAA

σ70- TTGACA 16-18 bp TATAAT

σ54- CTGGNA 6-bp TTGCA

Os motivos consensuais reconhecidos por σ24 e σ38 não foram descritos, devido a sua pouca conservação ou reduzido número de promotores confirmados.

A variação entre as sequências consensuais reconhecidas por cada fator σ, particularmente a posição relativa dos motivos conservados, limita a eficiência de uma abordagem de análise global. A predição de promotores deve ser feita para cada família σ separadamente, pois a análise de um dado promotor através da comparação com o motivo promotor consensual σ70 pode levar a resultado incorreto.

Compilações de promotores e análises permitiram o desenvolvimento de programas de computador que predizem a localização de sequências promotoras com base em sua homologia com sequências consensuais ou uma lista de referência de promotores (Polat e Günes, 2007). A abordagem clássica para a predição de promotores envolve o desenvolvimento de algoritmos que usaram matrizes de peso score-posição (PWMs). Esta metodologia apresenta

resultados, alinhando exemplos de sequências e estimando a preferência de base em cada posição de uma matriz (Gordon et al, 2006; Stormo, 2000; Hannenhalli e Wang, 2005).

Nos últimos anos, abordagens de Aprendizados de Máquina têm sido aplicadas para o reconhecimento e predição de promotores. Entre estes, Support Vector Machines (SVM), e Neural Network (NN) deram resultados promissores. Os métodos SVM usam um algoritmo de treinamento e podem representar funções complexas não-lineares. Este algoritmo tem como objetivo separar o conjunto de dados em duas classes por um hiperplano (Kapetanovich et al, 2004). O SVM pode ser aplicado para identificar importantes elementos biológicos: fatores de transcrição (Holloway, 2007), promotores (Polat e Gunes, 2007; Liang e Li, 2006), sítios de início da transcrição (Gordon et al, 2006; Gao, T. et al, 2009), entre outros.

Os NNs são ferramentas computacionais com funções complexas não-lineares. Eles têm sido usados para muitas aplicações biológicas, como predição de promotores (Demeler e Zhou, 1991; Burden et al, 2005; Rani et al, 2007), expressão gênica (Tan e Pan, 2005; Janga e Collado-Vides, 2007 proteína) e análise proteica (HellesFonseca, 2009; Chae et al, 2009). Os NNs são adequados para a predição e reconhecimento de promotores, devido à sua capacidade de identificar padrões degenerados, imprecisos e incompletos nessas sequências, e atingiram um elevado desempenho no processamento de grandes sequências do genoma (Cotik et al, 2005; Kalate et al, 2003). Além disso, a metodologia NN permite a extração de regras a partir de redes treinadas, o que pode ajudar na descoberta de regras biológicas aprendidas com os dados de entrada (Andrews et al, 1995).

Na literatura, existem alguns artigos descrevendo preditores de promotores, como BDGP [Reese MG (2001)], porém nenhum deles usa as regras extraídas de treinamentos de redes neurais mencionadas aqui, conforme descrito a seguir.

É um objeto da presente invenção um método, sistema e aparelho para predizer e/ou reconhecer e/ou classificar sequências biológicas, especialmente

as sequências de famílias com motivos de reconhecimento de sítios de ligação mal conservados, compreendendo o uso de regras de redes neurais. Em uma modalidade preferencial, a invenção fornece uma ferramenta de predição de promotores bacterianos, denotada até então como BacPP. Diferentemente de outras ferramentas de previsão, que empregam apenas sequências σ70 para a previsão de todos os promotores, BacPP é baseada em regras extraídas de processos de aprendizagem de NN para famílias de promotores σ24, σ28, σ32, σ38, σ54 e σ70. As informações obtidas das regras foi ponderada para maximizar a previsão de promotores e a classificação de acordo com seu fator σ.

Alguns documentos de patentes relacionados descrevem ferramentas de predição usando informações biológicas, como descrito a seguir.

O documento US 2010/0057419 descreve uma classificação *fold-wise* de proteínas compreendendo a predição de um padrão de dobra de uma proteína de interesse tendo um padrão de dobra desconhecido por formação de um sistema para correlacionar características estruturais ou de sequência para o padrão de dobra conhecido da proteína para prever os padrões de dobra da proteína, de preferência usando SVMs. A presente invenção descreve o uso de regras de redes neurais (NN) para classificar, predizer e/ou reconhecer sequências biológicas bacterianas pouco conservadas, o que não é citado no documento acima, e não inclui a predição específica de padrões de dobra de proteínas.

O documento US 2009/0111099 descreve um método de detecção e análise de promotores compreendendo a inserção de um candidato a sequência em um vetor compreendendo uma sequência TAG. A presente invenção descreve o uso de regras de redes neurais (NN) para classificar, predizer e/ou reconhecer sequências biológicas bacterianas pouco conservadas, o que não é citado no documento acima.

O documento US 2008/0147369 descreve métodos, sistemas e software para a identificação de biomoléculas funcionais compreendendo a geração de um modelo através da identificação de produtos vetoriais (*cross-products*)

usando algoritmos genéticos. A presente invenção descreve o uso de regras de redes neurais (NN) para classificar, predizer e/ou reconhecer sequências biológicas bacterianas pouco conservadas, o que não é citado no documento acima.

O documento WO 2007/059119 descreve sistemas e métodos para a identificação de indicadores de diagnóstico usando as regras das redes neurais, determinando a capacidade de resposta a uma terapia. A presente invenção está relacionada com o uso de regras de redes neurais (NN) para classificar, predizer e/ou reconhecer sequências biológicas bacterianas pouco conservadas, o que não é descrito em qualquer documento anterior, e não é aplicada a identificação de indicadores de diagnóstico.

Em vista do estado da técnica citado acima, pode-se ver que não foi encontrado nenhum antecedente relevante que revele uma abordagem matemática para validar mutações de proteínas, como o que foi aqui divulgado.

Os objetos e as vantagens da invenção aqui estabelecidos serão prontamente apreciados adiante, ou podem ser aprendidos através da prática com a invenção. Esses objetos e as vantagens são realizados e obtidos por meio de instrumentos e combinações apontados na descrição da invenção e nas reivindicações.

Sumário da Invenção

Em um primeiro aspecto, a presente invenção fornece um método, sistema e aparelho para predição biológica, usando interações não lineares entre módulos funcionais dentro de novas e inventivas ferramentas de bioinformática, aqui utilizadas para predizer características biológicas, especialmente sequências biológicas bacterianas pouco conservadas, como promotores.

Portanto, é um objeto da presente invenção um método, sistema e/ou aparelho para predizer e/ou reconhecer e/ou classificar motivos de reconhecimento de sítios de ligação pouco conservados, compreendendo o uso

de regras de redes neurais para sequências biológicas múltiplas pouco conservadas, sendo a sequência biológica preferencialmente um promotor.

É um objeto da presente invenção um método para predizer características biológicas compreendendo:

a) treinar/aprendizagem da NN para sequências "X" de motivos de reconhecimento de sítios de ligação pouco conservados;

b) extrair regras para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

c) substituir os valores protótipo da extração das regras da NN por um número inteiro para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

d) analisar as sequências para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

e) pontuar a sequência (> valor de corte); e

f)verificar se é um promotor da família fator sigma "X",

onde X significa a família de uma determinada sequência.

Preferencialmente, os motivos de reconhecimento de sítios de ligação pouco conservados são promotores bacterianos, pertencentes a diferentes fatores σ. Mais preferencialmente, o fator σ compreende σ24, σ28, σ32, σ38, σ54 e σ70.

É também um objeto da presente invenção um sistema que compreende o método para prever as características biológicas bacterianas descritas acima.

É também um objeto da presente invenção um aparelho compreendendo o método para prever as características biológicas bacterianas descritas acima.

Estes e outros objetos da invenção se tornarão mais evidentes para os técnicos na arte através da leitura da descrição detalhada abaixo.

Breve Descrição das Figuras

A Figura 1 mostra o gráfico de frequência WebLogo de protótipos NN para: a) promotores σ24; b) promotores σ28; c) promotores σ32; d) promotores σ38; e) promotores σ54, e; f) promotores σ70.

A Figura 2 mostra as respectivas estabilidades quando diferentes regras são aplicadas (regras 1-7).

A Figura 3 mostra a RMS em treino e simulações teste 1 e 2.

Descrição Detalhada da Invenção

Todos os exemplos mostrados no presente pedido devem ser entendidos apenas como exemplos ilustrativos, e não limitam o escopo da invenção. Todos os exemplos aqui citados, bem como formas semelhantes ou equivalentes para alcançar os objetivos da invenção estão abrangidas pela presente invenção.

Sistema, Método, Aparelho para Predizer/Reconhecer/Classificar Sequências Biológicas

É um objeto da presente invenção um método, sistema e/ou aparelho para predizer e/ou reconhecer e/ou classificar sequências biológicas especialmente motivos de reconhecimento de sítios de ligação pouco conservados, compreendendo o uso de regras de redes neurais para sequências biológicas bacterianas pouco conservadas, sendo a sequência biológica preferencialmente um promotor. É também um objeto da presente invenção um sistema que compreende o método para predizer as características biológicas bacterianas descritas acima.

Preferencialmente, o sistema pode ser entendido como, mas não limitado a, qualquer software e/ou middleware compreendendo o método anteriormente descrito.

É também um objeto da presente invenção um aparelho compreendendo o método para prever as características biológicas bacterianas descritas acima.

Preferencialmente, o aparelho pode ser entendido como, mas não limitado a, qualquer hardware or computador compreendendo o método para predizer as características biológicas bacterianas descritas acima.

É um objeto da presente invenção um método, para predizer e/ou reconhecer e/ou classificar características biológicas bacterianas compreendendo:

a) treinar/aprendizagem da NN para sequências "X" de motivos de reconhecimento de sítios de ligação pouco conservados;

b) extrair regras para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

c) substituir os valores protótipo da extração das regras da NN por um número inteiro para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

d) analisar as sequências para sequências "X" de motivos de reconhecimento de sítio de ligação pouco conservados;

e) pontuar a sequência (> valor de corte); e

f) verificar se é um promotor da família fator sigma "X",

onde X significa a família de uma determinada sequência.

Motivos de Reconhecimento de Sítios de Ligação Pouco Conservados

Os motivos de reconhecimento de sítios de ligação pouco conservados da presente invenção compreendem qualquer sequência biológica (DNAs, RNAs, sequências de proteínas, entre outros) com motivos pouco conservados capazes de reconhecer qualquer sítio de ligação biológico e/ou químico. De preferência, os motivos de reconhecimento de sítios de ligação pouco conservados e/ou sequências biológicas pouco conservadas são promotores, preferencialmente promotores. Em uma modalidade preferencial, os fatores σ compreendem σ24, σ28, σ32, σ38, σ54 e σ70.

Preferencialmente, os motivos de reconhecimento de sítios de ligação pouco conservados e/ou sequências biológicas pouco conservadas são sítios de ligação para RNAP, a região -35 e -10 em relação aos nucleotídeos do TSS.

Os motivos de reconhecimento de sítios de ligação pouco conservados e/ou sequências biológicas pouco conservadas podem ser usados simultaneamente para treinar a rede NN e, depois de obtidas as regras NN, eles também podem ser usados como um alvo a ser previsto/classificado/ reconhecido em sequências biológicas.

Exemplo 1 – BacPP: Bacterial Promoter Prediction – Uma ferramenta para a predição e reconhecimento de promotores

Métodos

1.1 Registro de dados

Sequências de promotores de *Escherichia coli* obtidos a partir do banco de dados RegulonDB (Gama-Castro et al, 2008), em sua versão disponível em abril de 2009, foram usados como exemplos positivos para o treinamento NN. Um total de 1.034 sequências, subdivididas de acordo com seu fator σ foram empregadas (Tabela 1). Como exemplos negativos de treinamento NN, sequências aleatórias com uma probabilidade de 0,28 para nucleotídeos adenina (A) e timina (T) e uma probabilidade de 0,22 para nucleotídeos citosina (C) e guanina (G) foram gerados. O mesmo número de exemplos positivos e negativos foram utilizados nas simulações realizadas.

Tabela 1. Número de sequências empregadas na simulação para cada fator σ.

Fator σ	No. de sequências de promotores
σ24	69
σ28	21
σ32	71
σ38	99
σ54	38
σ70	740

Sequências de promotores de outras *Enterobacteriaceae* foram obtidas da literatura disponível, já que as únicas bases de dados da web disponíveis são para a *E. coli* e *Bacillus subtilis*. Assim, um total de 82 sequências pertencentes a *Citrobacter*, *Enterobacter*, *Klebsiella*, *Proteus*, *Salmonella*, *Shigella*, *Yersinia genera* foram obtidos e empregados (Aldridge et al, 2006; Beach e Osuna, 1998; Castellanos et al, 2009; Ging e Inoye , 1986; Hu et al, 2000; Ibanez-Ruiz et al, 2000; Kutsukake et al, 1990; Mares et al, 1992; Maxson e Darwin, 2006; Penfound e Foster, 1999; Perez e Groisman, 2009; Ramírez-Santos et al, 2001; Skovierova et al, 2006; Smith e Somerville, 1997;

Sulavik et al, 1997; Toru et al, 1993; Yang et al, 2008; Wösten e Groisman, 1999).

1.2 Simulação da Rede Neural

Simulações NN foram realizadas para cada família σ. Os nucleotídeos foram codificados usando um conjunto de quatro dígitos binários como dado por (Brunak et al, 1991): A = 0100, T = 1000, C = 0001 e G = 0010. Uma sequência de entrada foi classificada como promotor presumível se sua saída de valores estabelecidos fosse entre 0,5 e 1,0. Caso contrário, considerou-se como um não-promotor.

As simulações foram realizadas em R environment (R Development Core Team, 2008). Nós escolhemos o algoritmo de propagação reversa (BP) com uma validação cruzada k-fold. Esta escolha foi feita a fim de obter resultados estatisticamente válidos. Nesta técnica, o conjunto de dados foi dividido em subconjuntos k. A cada iteração, um dos subconjuntos k foi usado como o conjunto de teste e os outros foram colocados juntos para formar um conjunto de treinamento. Assim, o erro médio em todos os ensaios k foi calculado (Polat e Gunes, 2007). No estudo apresentado os valores k, determinados pelo número de sequências promotoras disponíveis, foram: 10 para promotores σ70, 2 para promotores σ28 e σ54, 3 para promotores σ24, σ32 e σ38. Esses números para o valor de k foram determinados pelo número de sequências promotoras disponíveis.

Os resultados foram avaliados comparando sua: precisão (A), especificidade (S) e sensibilidade (SN). Estes parâmetros foram calculados utilizando as seguintes fórmulas:

Fórmula 1:

$$A = \frac{TP + TN}{TN + TP + FN + FP}$$

Fórmula 2:

$$S = \frac{TN}{TN + FP}$$

Fórmula 3:

$$SN=\frac{TP}{TP+FN}$$

onde, TP (verdadeiro positivo) são sequências de promotores classificadas como promotores; TN (verdadeiro negativo) são sequências aleatórias reconhecidas como não-promotoras; FP (falso positivo) são sequências aleatórias classificadas como promotoras e FN (falso negativo) são os promotores classificados como sequências não-promotoras.

O NN é aplicável em uma variedade de problemas, mas o processo de aprendizagem é complexo (Andrews et al, 1995). Como NN aprende a classificar uma determinada sequência como promotora ou não-promotora pode ser entendido através da extração de regras. Assim, uma explicação de como cada decisão NN é feita aumenta o conhecimento sobre essas sequências (Odajima et al, 2008). Neste trabalho, foram extraídas as regras com base no valor dos neurônios escondidos por uma metodologia denotada FAGNIS, de acordo com Cechin (1998). A extração de regras foi desenvolvida no R Environment. Esta técnica consiste em segmentar uma função sigmóide em três regiões (ver Figura 1). Para cada entrada, verifica-se em que região do sigmóide os neurônios escondidos estavam aptos. O número máximo de combinações é 3^n, onde n é um número de neurônios na camada oculta. No entanto, todas as combinações possíveis não ocorrem e, somente as combinações mais frequentes são consideradas. Essas combinações são a melhor representação dos dados de entrada.

Portanto, os resultados são convenientemente apresentados por um protótipo de regra, que definimos como modelo médio do conjunto de dados de entrada. A regra pode ser escrita como uma equação linear: "Se x = ou ~ protótipo, então y = constante da equação linear + (coeficientes da equação linear)". Aqui, x é um exemplo de entrada, y corresponde à saída NN e os coeficientes da equação linear são os nucleotídeos da sequência.

1.3 Implementação do BacPP

A ferramenta BacPP foi implementada na linguagem de programação Python (van Rossum, 2010). Uma visão geral desta abordagem é dada na Figura 1. A ideia global desta ferramenta foi ponderar a pontuação obtida pelos protótipos de extração de regras NN para cada promotor de fator σ, e usá-los como modelos para determinar e classificar os promotores pela sua família sigma. Diversos pesos foram avaliados. Os pesos foram definidos usando números inteiros entre -10 e +10. Para um dado nucleotídeo, se a pontuação do protótipo está acima de 0,3 ou 0,2, os valores foram substituídos por um número positivo ou negativo, respectivamente. Se a pontuação do protótipo está entre em 0,29 e 0,2, esses protótipos de valores foram substituídos por zero. As melhores ponderações são apresentadas na Tabela 2.

Tabela 2. A melhor ponderação definida para os resultados da classificação do BacPP

Pontuação do protótipo de regra NN	No. Inteiro o substituindo
Acima de 0.6	+6
0,5 a 0,59	+4
0,4 a 0,49	+2
0,3 a 0,39	+1
0,2 a 0,29	0
0,1 a 0,19	-1
Abaixo de 0.1	-3

Resultados

Na simulação NN, a arquitetura que melhor classifica o conjunto de sequências de entrada para cada promotor σ é apresentada na Tabela 3. Um maior número de neurônios na camada oculta não aumentou significativamente a precisão, a especificidade ou a sensibilidade.

Tabela 3. A melhor arquitetura para toda simulação de fator de família σ

Fator σ	No. de neurônios na camada de entrada	No. de neurônios na camada oculta	No. de neurônios na camada de saída
σ24	324	4	1
σ28	324	2	1
σ32	324	2	1
σ38	324	2	1
σ54	324	2	1
σ70	324	5	1

Usando a melhor arquitetura para cada fator de família σ, o NN alcançou uma precisão média de 71,67%, uma especificidade de 71,08%, e uma sensibilidade de 72,98%, com baixa variação entre os fatores σ (Tabela 4). A semelhança entre a especificidade e valores de sensibilidade dentro de cada σ, é um indicativo da consistência do processo de aprendizagem NN. Os valores obtidos para a precisão, especificidade e sensibilidade para σ70 são baixos, mas comparáveis com os previamente reportados utilizando metodologias NN (Burden et al, 2005).

Tabela 4. Comparação entre precisão, especificidade e sensibilidade entre os fatores σ

Fator Sigma	Precisão (%)	Especificidade (%)	Sensibilidade (%)
σ24	71.6	69.1	73.9
σ28	70.2	66.6	73.8
σ32	72.4	72.4	72.4
σ38	67.7	68.5	66.8
σ54	73.5	73.2	73.8
σ70	77.0	76.7	77.2

Os protótipos NN, extraídos do NN treinado usando o algoritmo FAGNIS, para cada família de fator σ são mostrados como um gráfico de frequência (Figura 1).

O protótipo obtido para σ24 não mostrou qualquer motivo altamente

conservado, mas como a maioria dos promotores, é caracterizado por uma alta prevalência de nucleotídeos AT (Lewin, 2008). Até o momento, nenhum motivo conservado tem sido descrito para sequências de promotores σ24. Por outro lado, protótipos de promotores σ28 (Figura 3) apresentaram dois motivos conservados, um entre -15 e -7, TGCCGATAA, e uma outra entre -33 e -25, TAAAGTTT, que correspondem aqueles descritos anteriormente (Song et al, 2007).

O protótipo obtido para os promotores σ32 foram caracterizados pela presença de dois motivos parcialmente conservados. Um motivo -7 a -15 com uma sequêncial consensual CYCYAWWWW, e uma sequência YTKRWWW - 28 a -35. De acordo com o código IUPAC, as letras Y, W, K, R representam: C ou T, A ou T, G ou T, A ou G, respectivamente. Estes motivos são semelhantes aos descritos por Wang e deHaseth (2003). Por outro lado, nenhum motivo conservado típico foi evidenciado em promotores σ38. No entanto, uma região rica em W pôde ser identificada nos primeiros 11 nucleotídeos, com um T altamente conservado a -7. Promotores σ38 de pouca conservação e motivos ricos em W foram evidenciados por Typas et al (2007).

O protótipo obtido para promotores σ54 mostrou dois motivos conservados: (1) uma sequência consensual WWCGTT entre -10 e -15, e (2) uma sequência ACGGT entre -22 e -26. Estes motivos são semelhantes aos descritos por Barrios et al (1999). Protótipos de promotores σ70 foram caracterizados por um alto conteúdo A/T/W (88%) em comparação com os outros promotores, que variaram entre 30% (σ32) e 58% (σ38). Uma alta frequência de dinucleotídeos AA, AT e TT foi relatada por Kanhere e Bansal (2005) em promotores σ70 de *E. coli*. No entanto, os típicos motivos conservados a -10 e -35 não foram evidentes no protótipo extraído do NN. Este fato pode ser devido à variação no número de nucleotídeos entre o +1 e o primeiro motivo, bem como entre os motivos dos promotores σ70 (Shultzaberger et al, 2006).

Em geral, as regras do protótipo extraído do NN mostraram os motivos conservados descritos anteriormente para cada família σ, indicando que a aprendizagem NN tem fundamento biológico.

Considerando a eficiência do aprendizado NN, as regras foram ponderadas e usadas para desenvolver uma ferramenta de predição para as sequências de promotores bacterianos, separados pela família σ. Este programa foi nomeado BacPP.

Como pode ser observado na Tabela 5, BacPP mostrou maior precisão, especificidade e sensibilidade do que os NN originais. Usando BacPP, as mais altas precisões foram obtidas para σ54, σ28 e σ32, todos acima de 90%. Esses promotores da família σ exibiram motivos altamente conservados, fato que pode explicar a eficiência da previsão. Por outro lado, a menor exatidão foi obtida para σ70, os promotores σ menos conservados. Embora os promotores σ70 sejam os mais abundantes em genomas bacterianos, eles têm mais desvios que promotores consensuais canônicos (Janga e Collado-Villes, 2007).

Tabela 5 – A melhor arquitetura para toda simulação de fator de família σ

Factor σ	Precisão (%)		Especificidade (%)		Sensibilidade (%)	
	BacPP	NNPP	BacPP	NNPP	BacPP	NNPP
σ24	**86.9**	67.2	**95.6**	60.8	**78.2**	50.7
σ28	**92.8**	68.2	**90.4**	**75**	**95.2**	50
σ32	**91.5**	68.4	**92.9**	60.5	**90.1**	64.7
σ38	**89.3**	68.2	**83**	62.6	**93.9**	64.6
σ54	**97**	67.8	**100**	60	**94.11**	48.5
σ70	**83.6**	74.4	**85.4**	68.7	**81.8**	80

Para avaliar a eficiência do BacPP, o mesmo conjunto de promotores e sequências aleatórias foi utilizado para os promotores de previsão usando BacPP e NNPP (Burden et al, 2005), um outro algoritmo baseado em aprendizagem de máquina. Como pôde ser observado na Tabela 5, BacPP mostrou maior precisão, especificidade e sensibilidade, com valores na ordem de 90% e 65%, para BacPP e NNPP, respectivamente. Todos os parâmetros foram maiores, mesmo para promotores σ70, utilizados na modelagem NNPP.

Em outra simulação, BacPP foi usado em um cruzamento-teste para avaliar a especificidade da classificação dos promotores por fator σ. Por exemplo, BacPP para σ70 (testador) foi usado contra sequências de promotor σ24, σ28, σ32, σ38 e σ54 (testados). Em todos os casos, a maior especificidade foi obtida para o próprio fator σ, indicando que o BacPP classifica de forma eficiente sequências de promotores por família σ. No entanto, os valores obtidos quando um dado modelo de fator σ foi aplicado em outras sequências sigma foram superiores aos obtidos com sequências aleatórias, indicando que existe alguma conservação entre sequencias promotoras independente do fator σ. Este fato explica a eficiência relativa de modelos de predição baseados em σ70 na identificação de outras sequências sigma (Gordon et al, 2003; Burden et al, 2005).

Uma comparação entre BacPP com a máquida de aprendizagem baseada em programas relatada anteriormente mostrou que a precisão média de BacPP (90,2%), especificidade (91,0%) e sensibilidade (89,5%) são comparáveis com as abordagens mais eficientes publicadas até o momento. Oppon (2000) relatou um programa de predição de promotores em rede neural (NNPP) que exibiu 60% de especificidade e 50% de sensibilidade para promotores σ70. Este programa foi melhorado na sua especificidade (Burden et al, 2005). Uma abordagem baseada em SVM usando alinhamento da sequência Kernel obteve uma precisão de 84%, especificidade de 84% e sensibilidade de 82% (Gordon et al, 2003). Polate e Günes (2007), com pelo menos uma máquina de vetores de suporte, mas usando apenas 56 sequências obtiveram uma precisão de 84,6%, sensibilidade de 90,9% e especificidade de 80%.

A metodologia NN usando informações de dinucleotídeos como dados de entrada alcançou uma precisão de 96%, especificidade de 98% e sensibilidade de 93% para promotores σ70 usando exemplos negativos com 60% de nucleotídeos AT (Rani et al, 2007). Embora altamente eficiente, esta abordagem é específica apenas para sequências sigma ricas em AT como σ70. Em outro trabalho, Rangannan e Bansal (2007) utilizaram a estabilidade do

DNA para prever sequências promotoras e obter uma precisão de apenas 52,2% com sensibilidade de 98%.

Quando avaliados contra um conjunto de 82 sequências de promotores de outras espécies de enterobactérias, o BacPP exibiu uma precisão de 80,5%, com sensibilidade de 86% e especificidade de 75%, indicando que BacPP pode ser eficientemente usado para predição de promotores e classificação em diferentes bactérias Gram-negativas.

Nesta invenção, apresentamos uma nova abordagem para a predição de sequências biológicas e classificação com base na ponderação do protótipo promotor obtido a partir das regras extraídas de NNs, como descrito anteriormente. Ao separar as sequências de promotor de acordo com seu fator σ, temos demonstrado que a predição e os conhecimentos sobre os promotores podem ser melhorados. A precisão obtida para promotores σ24, σ28, σ32, σ38, σ54 e σ70 de *E. coli* foram 86,9%, 92,8%, 91,5%, 87,8%, 97,0% e 83,6%, respectivamente. Quando aplicada a um conjunto de promotores de diferentes enterobactérias, a precisão de BacPP foi de 76%, indicando que este método pode ser expandido para outras espécies de bactérias. Diferentemente das ferramentas/métodos anteriormente descritos na literatura, BacPP permite não só a identificação de promotores de bactérias, mas também a sua classificação de acordo com seu fator σ.

Os técnicos no assunto prontamente apreciarão esta invenção, e entenderão que outras modalidades devem ser consideradas como dentro do escopo da invenção e das reivindicações anexadas.

Reivindicações

1. Método para predizer e/ou reconhecer e/ou classificar sequências biológicas, **caracterizado** pelas ditas sequências biológicas serem preditas, reconhecidas e/ou classificadas por regras extraídas de processo de aprendizagem de redes neurais para sequências biológicas pouco conservadas, o processo sendo realizado por um computador e compreendendo as etapas de:

a) treinar/aprendizagem da rede neural para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados, obtidos de uma base de dados, sequências em que a posição relativa dos motivos são pouco conservados;

b) extrair regras para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

c) substituir os valores protótipo da extração das regras da NN por um número inteiro para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

d) analisar cada uma das sequências para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

e) pontuar cada uma das sequências e determinar se cada uma das sequências é maior que o valor de corte; e

f) verificar se cada uma das sequências é um promotor da família fator sigma "X",

onde X significa a família de uma determinada sequência, X sendo mais de 1 e,

em que a arquitetura do processo de aprendizado da rede neural compreende entre 2 e 5 neurônios em uma camada escondida da rede neural,

portanto, permitindo a predição e/ou reconhecimento e/ou classificação da sequência.

2. Método, de acordo com a reivindicação 1, **caracterizado** pelos

motivos de reconhecimento de sítio de ligação pouco conservados serem sequências relacionadas a promotores.

3. Método de acordo com a reivindicação 1ou 2, **caracterizado** pelos motivos de reconhecimento de sítio de ligação consensual pertencerem a família de fatores σ24, σ28, σ32, σ38, σ54, σ70.

4. Método de acordo com qualquer uma das reivindicações 1 a 3, **caracterizado** pela arquitetura do processo de aprendizado da rede neural compreender cerca de 324 neurônios na camada de entrada.

5. Método de acordo com qualquer uma das reivindicações 1 a 6, **caracterizado** pela arquitetura do processo de aprendizado da rede neural compreender cerca de 1 neurônio na camada de saída.

6. Método de acordo com qualquer uma das reivindicações 1 a 5, **caracterizado** por um total de 1034 sequências relacionadas com o fator σ serem empregadas para treinar a rede neural.

7. Método de acordo com qualquer uma das reivindicações 1 a 6, **caracterizado** pelas simulações da rede neural serem realizadas para cada família de sequências, considerando-se como um promotor um valor estabelecido de saída entre 0,5 e 1,0.

8. Método de acordo com qualquer uma das reivindicações 1 a 7, **caracterizado** por usar algoritmo de propagação reversa com uma validação k-fold-cross.

9. Método de acordo com qualquer uma das reivindicações 1 a 8, **caracterizado** pelas regras serem extraídos utilizando a metodologia FAGNIS.

10. Método de acordo com qualquer uma das reivindicações 1 a 9, **caracterizado** por uma ferramenta para predição e reconhecimento do promotor estar implementada, ponderando a pontuação obtida pelas regras protótipos das redes neurais.

11. Método de acordo com qualquer uma das reivindicações 1 a 10, **caracterizado** pela ferramenta ser BacPP.

12. Método de acordo com qualquer uma das reivindicações 1 a 11,

caracterizado pelos pesos serem definidos usando números inteiros entre -10 e +10.

13. Método de acordo com qualquer uma das reivindicações 1 a 12, **caracterizado** pela precisão, especificidade ou sensibilidade da rede neural estar entre 65-80%.

14. Método de acordo com a reivindicação 13, **caracterizado** pela precisão, especificidade e/ou sensibilidade estar entre 75-100%.

15. Aparelho para predizer e/ou reconhecer e/ou classificar sequências biológicas, **caracterizado** pelo dito aparelho compreender meios para a predição, reconhecimento e/ou classificação de sequências biológicas de acordo com regras extraídas de processo de aprendizagem de redes neurais para sequências biológicas pouco conservadas compreendendo:

a) meios para o treinamento/aprendizagem da rede neural para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados, isto é, sequências em que a posição relativa dos motivos são pouco conservados;

b) meios para a extração de regras para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

c) meios para a substituição dos valores protótipo da extração das regras da rede neural por um número inteiro para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

d) meios para a análise de cada uma das sequências para sequências “X” de motivos de reconhecimento de sítio de ligação pouco conservados;

e) meios para a pontuação de cada uma das sequências e determinação se cada uma das sequências é maior que o valor de corte; e

f) meios para a verificação se cada uma das sequências é um promotor da família fator sigma "X",

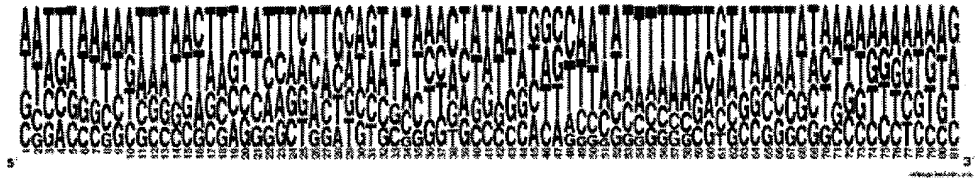
onde X significa a família de uma determinada sequência, X sendo mais de 1,

em que a arquitetura do processo de aprendizado da rede neural

compreende entre 2 a 5 neurônios na camada escondida da rede neural, e

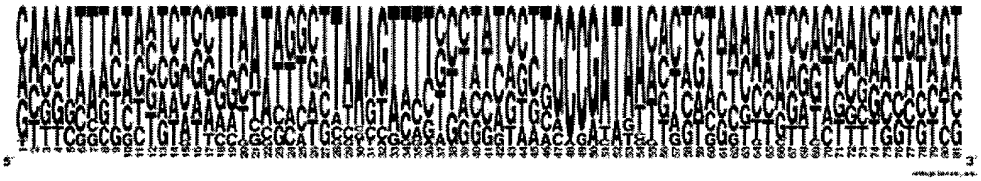
em que o aparelho permite a predição e/ou reconhecimento e/ou classificação da sequência.

Figuras

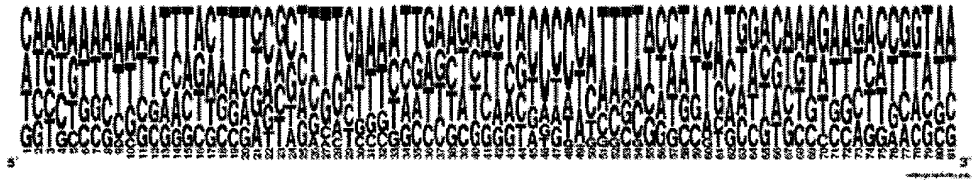


A

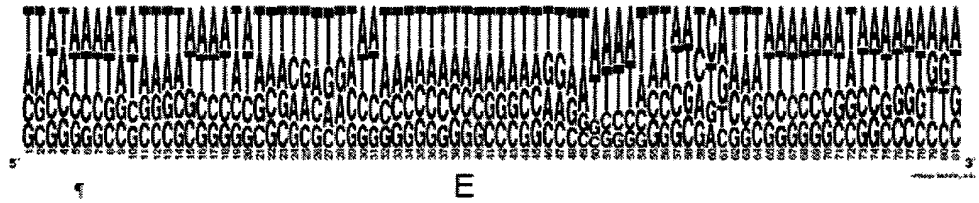
B



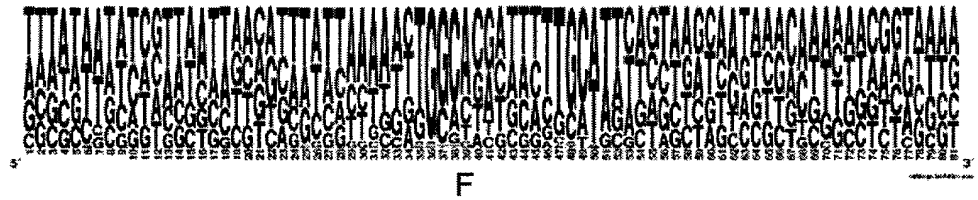
C



D



E



F

Figura 1

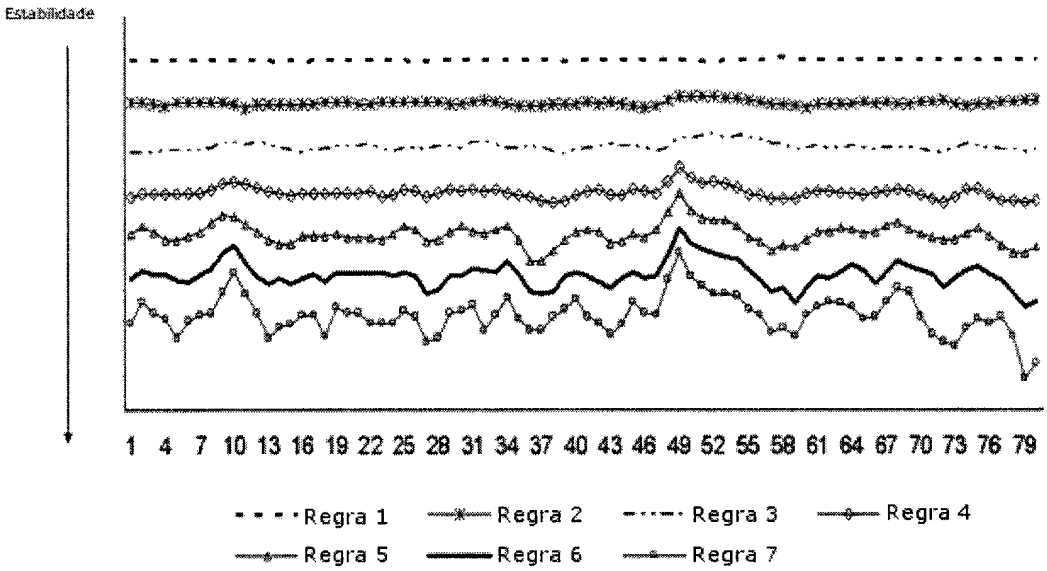


Figura 2

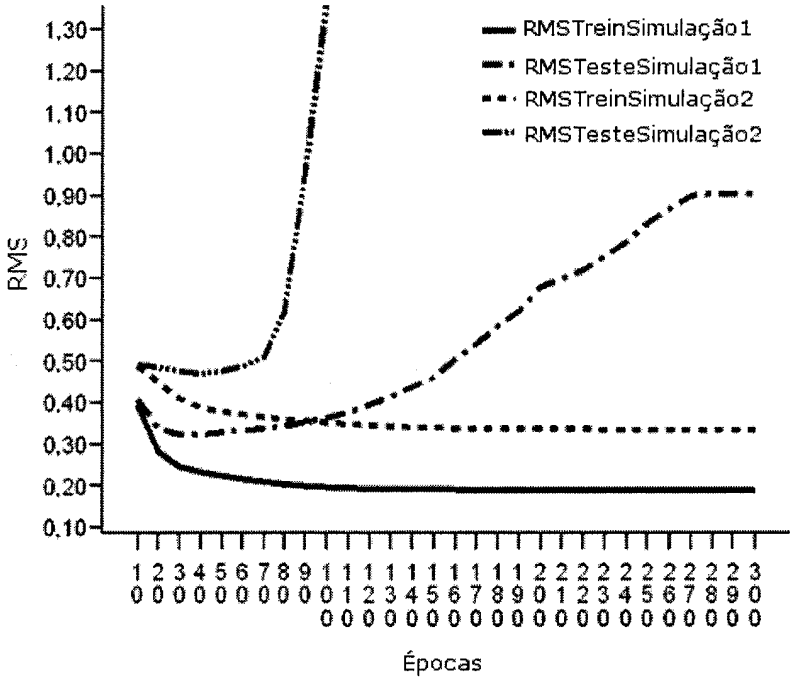


Figure 3