República Federativa do Brasil
Ministério da Indústria, Comércio Exterior e Serviços
Instituto Nacional da Propriedade Industrial

(21) BR 102016023989-3 A2

(22) Data do Depósito: 14/10/2016

(43) Data da Publicação: 02/05/2018

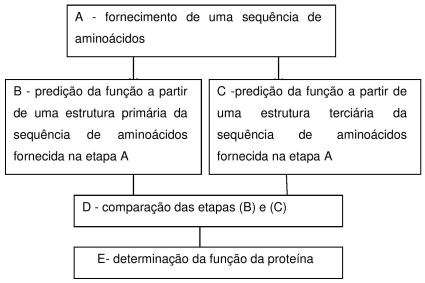
(54) Título: MÉTODO E SISTEMA PARA PREDIÇÃO DE FUNÇÕES DE PROTEÍNAS

(51) Int. Cl.: G06F 19/18

(73) Titular(es): INSTITUTO NACIONAL DE METROLOGIA, QUALIDADE E TECNOLOGIA - INMETRO., UNIVERSIDADE FEDERAL DO RIO DE JANEIRO - UFRJ

(72) Inventor(es): CAIO BULGARELLI; MANUELA LEAL DA SILVA; PAULO MASCARELLO BISCH

(57) Resumo: A presente invenção refere-se a métodos e sistemas relacionados à predição da função de sequência de aminoácidos oriundas de abordagens genômicas, metagenômicas, proteômicas e transcriptômicas em larga escala. O método para predição da função de uma proteína, a partir de uma sequência de aminoácidos, que compreende a predição direta a partir da sequência de aminoácidos e a predição a partir da estrutura tridimensional da referida sequência de aminoácidos com a posterior comparação entre as duas para gerar o resultado.



MÉTODO E SISTEMA PARA PREDIÇÃO DE FUNÇÕES DE PROTEÍNAS

CAMPO DE APLICAÇÃO

[0001] A presente invenção refere-se a métodos e sistemas relacionados à determinação da função de sequência de aminoácidos. Mais especificamente, a presente invenção aborda métodos e sistemas para predição da função de uma proteína a partir da estrutura de sequências genômicas, metagenômicas, proteômicas e transcriptômicas em larga escala.

ESTADO DA TÉCNICA

[0002] As proteínas são grandes moléculas, também chamadas de macromoléculas, que são constituídas essencialmente por aminoácidos que se organizam em uma determinada sequência específica.

[0003] A diferença entre essas estruturas reside, sobretudo, na forma de organização desses aminoácidos, determinado por uma sequência genética que provoca o enovelamento da referida macromolécula em uma estrutura tridimensional específica que determina a atividade e função protéica.

[0004] O número de cadeias proteicas que possui sua função bem caracterizada ainda é pequeno, principalmente quando se trata de longas cadeias de aminoácidos.

[0005] Para determinar ou predizer a função de uma proteína, podem ser utilizadas técnicas experimentais e também buscas por similaridade com sequências conhecidas ou anotadas, que tenham elevado grau de identidade com a sequência de interesse.

[0006] Entre as propostas para determinação da função de uma proteína as técnicas de anotação ou predição funcional de sequências podem ser citadas. Essas técnicas geralmente classificam as estruturas em grupos a partir de similaridades encontradas com proteínas de outras espécies, sendo essa uma rica fonte para a anotação funcional.

[0007] Assim, nesse caso a estratégia tradicional é procurar por similaridades em bancos de dados, por sequências protéicas com funções conhecidas que tenham similaridade com a sequência que se deseja analisar.

[0008] Com essa perspectiva, vários métodos têm sido desenvolvidos na tentativa de obter informação funcional sobre uma sequência de aminoácidos. Os métodos mais tradicionais incluem o alinhamento de sequência e análise de motivos locais da sequência. Nestes métodos, é calculado o grau de identidade da sequência de aminoácidos entre uma sequência experimental de interesse e uma ou mais sequências que tem a função conhecida no banco de dados.

[0009] Métodos de alinhamento, tais como do sistema BLAST (acrônimo em inglês para: *Basic Local Alignment Search Tool*) ou FASTA são utilizados para essa finalidade.

[00010] Embora esse procedimento venha sendo otimizado ao longo do tempo com bancos de dados públicos, por exemplo como os sistemas acima mencionados, ainda persistem limitações que podem comprometer e dificultar a análise em determinadas situações, como a demora de vários dias ou semanas, mesmo quando se utiliza servidores de alto desempenho.

[00011] Além disso, outras implicações que limitam o uso desses sistemas é a ausência de substancial similaridade das sequências e, ainda, novas sequências que muitas vezes não encontram coincidências com sequências conhecidas.

[00012] Outra limitação relata por Hobohm e Sander, é que os métodos puramente de alinhamento de sequências não são capazes de detectar semelhanças estruturais e funcionais quando a identidade de sequência é inferior a 25 % (HOBOHM, Uwe; SANDER, Chris. A sequence property approach to searching protein databases. Journal of molecular biology, v. 251, n. 3, p. 390-399, 1995).

[00013] Em alternativa para superar os inconvenientes da técnica de alinhamento de sequências, foram propostos métodos e sistemas baseados em pequenos fragmentos de sequências, também conhecidos como motivos, para auxiliar na identificação da função da proteína, sendo que ao invés de utilizar as sequências completas ou inteiras, essas técnicas usam apenas partes da cadeia de aminoácidos.

[00014] Assim, como pode ser deduzido das técnicas relatadas acima, a determinação da atividade da proteína é baseada integralmente apenas na estrutura primária da mesma, que fornece informação limitada acerca de sua natureza físico-química e funcionalidade.

[00015] Contudo, a estrutura primária aborda apenas a sequência linear dos aminoácidos que se unem por ligações peptídicas, sem levar em consideração informações das estruturas secundárias e terciárias, as quais abordam a conformação local de algumas regiões da cadeia polipeptídica e o arranjo tridimensional da cadeia polipeptídica com interações entre aminoácidos mais distantes, respectivamente.

[00016] Essa conformação tridimensional é diretamente responsável pelas propriedades biológicas da proteína, tais como atividade enzimática e especificidade de ligação dessa macromolécula.

[00017] Ainda que existam numerosas metodologias que usem a estrutura primária da proteína para inferir sua função, o conhecimento da estrutura tridimensional da mesma é considerado mais assertivo e confiável na averiguação das funções bioquímicas.

[00018] Assim, os métodos baseados unicamente em estrutura primária não são capazes de determinar com alta precisão a função de uma proteína e/ou polipeptídeo.

[00019] Em contrapartida, esse tipo de metodologia necessita de estudo acurado da forma tridimensional, que figura como sua maior desvantagem por necessitar experimentos demorados.

[00020] Assim, nos últimos anos, têm sido feitas novas abordagens para inferir a função de proteína, tal como descreve a patente US7,880,738 que aborda uma ferramenta para visualização e suporte para análise estrutural de moléculas em 3D, listagem de sequência, informações químicas e biológicas, sendo que esse guia ainda descreve utilidades para a análise da estrutura protéica. Contudo, esse sistema demanda atuação constante do técnico para realizar as análises e aferição de cada molécula individualmente e manualmente, não sendo capaz de realizar a determinação automaticamente e em larga escala.

[00021] O documento internacional WO2007140061 descreve um método para determinação de uma estrutura de equilíbrio, tal como estrutura secundária de um polipeptídeo, uma proteína ou uma região de dobragem autônoma, os quais ditam a função e atividade biológica dos mesmos. Essa metodologia não utiliza estruturas terciárias para determinar a função da proteína.

[00022] O documento de patente US2005089878 aborda métodos para detecção de resíduos funcionais na superfície de proteínas que fazem a determinação das contagens de anotação funcionais de uma proteína e a compara com distribuições de pontuações de anotação funcionais semelhantes, derivados de uma pluralidade de proteínas de referência. Tendo em conta essa comparação, um cluster funcional é anotado ou não. Tal como documento anterior, o método proposto por essa descrição também não aborda a determinação da função de uma proteína a partir da análise de uma estrutura terciária.

[00023] A publicação de GABDOULLINE, Razif R. et al. ProSAT: functional annotation of protein 3D structures. Bioinformatics, v. 19, n. 13, p. 1723-1725, 2003, aborda um sistema para efetuar o mapeamento automatizado de anotações funcionais que permite que locais funcionais sejam prontamente identificados. O preceito de referido sistema é mapear em uma estrutura 3D de uma proteína e os aminoácidos responsáveis por executar sua

função, então essas informações são retiradas de uma anotação prévia realizada por meio da sequência de aminoácidos da proteína que o usuário insere no sistema junto com o modelo 3D. Ou seja, ele não realiza nenhum tipo de determinação, somente o mapeamento da mesma.

[00024] O documento *YACHDAV, Guy et al. Predict Protein — an open resource for online prediction of protein structural and functional features. Nucleic acids research, p. gku366, 201*4 evidencia um analisador de sequência, capaz de gerar resultados de características estruturais e funcionais de dada proteína, sendo composto por mais diversas ferramentas. Apesar de o analisador realizar mesmas anotações funcionais, o mesmo não realiza anotação estrutural de seus modelos preditos para tentar validar a anotação a partir das sequências de aminoácidos e também não especifica nenhum mecanismo para análise da qualidade dos modelos gerados. Além disso, os dados gerados pelo *Predict Protein* tem baixo grau de confiabilidade, devido a ausência de aferição do modelo gerado.

[00025] O documento US20040185486 descreve um método para encontrar semelhanças 3D em estruturas de proteínas de uma primeira molécula e de uma segunda molécula, mediante o fornecimento de informações pré-selecionadas em relação à primeira molécula e a segunda molécula e a comparação entre as mesmas usando análise de segmentos contínuos (LCS), análise de distância global (GDT), análise de função de alinhamento local global (LGA_S) e por fim verificando o alinhamento construído e repetindo os passos para encontrar as regiões de semelhanças 3D em estruturas de proteínas. Contudo, para realização desse método é necessário conhecer previamente a estrutura 3D.

[00026] Outro documento do estado da técnica que busca resolver o mesmo problema é o WO0005414, que reivindica um método e sistema para a determinação da função bioquímica da proteína, tendo como base conformação estrutural e domínios da mesma, que são determinados a partir

de ferramentas de ressonância magnética nuclear (RMN). De forma geral, referido método aborda as etapas:

a) identificar um domínio de polipeptídeo que se dobra ou enovela de forma estável, com uma estrutura tridimensional definida;
b) determinar a estrutura tridimensional desse domínio estável;
c) comparar a estrutura da etapa anterior com aquelas contidas em banco de dados de estruturas com funções conhecidas
d) correlacionar as estruturas definidas e definir a função bioquímica da estrutura em análise.

[00027] Essa tecnologia também é dependente de aparelhos acessórios ao método, tal como o RMN, tornando o procedimento mais demorado e oneroso.

[00028] Portanto, as atuais tecnologias de determinação da função de proteínas ainda são complexas, demoradas e dependentes de equipamentos acessórios.

[00029] Consequentemente, existe a necessidade de provisão de uma metodologia de suporte que permita lidar com a crescente quantidade de dados que são obtidos a partir de sequenciadores genéticos, tal como um método automatizado que permita realizar a determinação da função da proteína a partir da estrutura de sequências genômicas, metagenômicas, proteômicas e transcriptômicas em larga escala, com garantias de confiabilidade e que ofereça soluções mais rápidas.

[00030] Destarte, para superar os problemas e deficiências do estado da técnica a presente invenção descreve um método para determinação da função de uma proteína, baseado em uma sequência primária inicialmente fornecida.

[00031] Em continuidade, a presente invenção supera os problemas e desvantagens do estado da técnica ao prover um método e sistema que garante confiabilidade de resultados e que é capaz de realizar determinações em larga escala de forma rápida e eficiente.

SUMÁRIO

[00032] Desse modo a presente invenção proporciona avanços em relação às técnicas convencionais para determinação da função de uma sequência de proteína.

[00033] De forma geral, uma modalidade da presente invenção aborda um método para predição da função de uma proteína, a partir de uma sequência de aminoácidos, que compreende:

(i) a predição direta a partir da sequência de aminoácidos e

(ii) a predição a partir da estrutura tridimensional da referida sequência de aminoácidos e

(iii) a posterior comparação entre as duas para gerar a função da mesma.

[00034] A predição da função da proteína a partir de uma estrutura tridimensional compreende ainda as etapas:

(iia) modelagem;

(iib) filtragem;

(iic) alinhamento estrutural.

[00035] Em outra modalidade a presente invenção aborda um sistema à predição da função de uma proteína, sendo que referido sistema implementa o método da presente invenção de forma totalmente automatizada.

[00036] Assim, a presente invenção proporciona avanços em relação às técnicas convencionais de determinação da função de proteínas, ao realizar a predição também a partir de uma estrutura tridimensional e, sobretudo, ao propor uma etapa de filtragem, a qual é responsável por conferir confiabilidade ao método, além de torná-lo mais rápido e sem a necessidade de intervenção humana.

[00037] Estes objetivos e vantagens da invenção serão imediatamente valorizadas pelos versados na técnica a partir da descrição detalhada das modalidades de invenção que se segue, a qual também será melhor compreendida em conjunto com as figuras e exemplos.

DESCRIÇÃO DAS FIGURAS

[00038] As Figuras anexas são incluídas para fornecer um melhor entendimento da matéria e são aqui incorporadas por constituírem parte desta especificação, bem como ilustram formas de realização da presente invenção e, em conjunto com a descrição e as reivindicações anexas, servem para explicar os princípios da presente invenção.

[00039] A Figura 1 é um diagrama esquemático de uma modalidade da presente invenção que demonstra, de forma geral, o método da presente invenção.

[00040] A Figura 2 é um diagrama esquemático de uma modalidade de realização da presente invenção, que descreve as etapas compreendidas pelo método.

[00041] A Figura 3A-3C são estruturas tridimensionais que demonstram o alinhamento estrutural entre moldes e modelos gerados por uma etapa da presente invenção. 3A mostra um modelo aprovado pela etapa de filtragem do método e 3B e 3C mostram proteínas que tiveram seus modelos reprovados na etapa de filtragem da presente invenção.

[00042] A figura 4 mostra um gráfico de Número de proteínas em função das categorias funcionais para: metabolismo, processo celular e sinalização, processamento e armazenamento de informações, pobremente caracterizadas, fixação de nitrogênio e resistência a antibióticos.

DESCRIÇÃO DETALHADA

[00043] Características e vantagens adicionais da invenção serão apresentadas na descrição detalhada que se segue e em parte serão prontamente evidentes para os peritos na técnica a partir desta descrição e com seus exemplos, ou reconhecidas pelas práticas das formas de realização, tal como aqui descrito, incluindo a descrição detalhada, o resumo, as reivindicações, assim como os desenhos anexos.

[00044] Deve ser entendido que tanto a descrição geral anterior quanto a seguinte descrição detalhada e as formas de realização da invenção com seus exemplos, destinam-se a proporcionar uma visão geral para compreender a natureza e caráter do invento, tal como é reivindicado. As figuras anexas são incluídas para proporcionar uma maior compreensão do invento e constituem uma parte desta especificação.

[00045] Além disso, no intuito de esclarecer e permitir o completo entendimento da presente invenção, os termos usados na descrição seguinte são definidos.

[00046] Desse modo, ao longo desta descrição, o termo "compreende" e suas variações devem ser entendidas como implicando a inclusão de um elemento declarado, ou etapa, ou grupo de elementos, ou etapas, mas não a exclusão de qualquer outro elemento, inteiro ou etapa, ou grupo de elementos, inteiros ou etapas.

[00047] O termo “estrutura primária” deve ser entendido como uma sequência de aminoácidos, de qualquer tamanho, tal como um peptídeo, um polipeptídeo ou uma proteína.

[00048] O termo “estrutura terciária” deve ser entendido como a estrutura tridimensional da sequência de aminoácidos e que carrega consigo a informação de estrutura primária e secundária.

[00049] A presente invenção aborda um método para a predição da função de uma proteína com base no fornecimento somente de uma sequência de aminoácidos, realizando conjuntamente a predição para sequência linear ou estrutura primária e para uma estrutura tridimensional ou estrutura terciária.

[00050] Assim a predição é realizada concomitantemente de duas formas, uma primeira predição é realizada com base na estrutura primária e uma segunda predição é realizada com base em uma estrutura terciária, ambas a partir de uma mesma sequência de aminoácidos que é fornecida.

[00051] Como já destacado, esses dois tipos de predição ocorrem paralelamente e independente um do outro.

[00052] Nesse sentido, referido método compreende as etapas:

(a) fornecimento de uma sequência de aminoácidos;

(b) predição da função a partir de uma estrutura primária da sequência de aminoácidos fornecida na etapa (a);

(c) predição da função a partir de uma estrutura terciária da sequência de aminoácidos fornecida na etapa (a);

(d) comparação das etapas (b) e (c) e

(e) determinação da função da proteína.

[00053] A sequência de aminoácidos, tal como um peptídeo, um polipeptídeo ou uma proteína pode ser obtida a partir de um sequenciador.

[00054] A etapa de predição com base na sequência primária, ou etapa (b) compreende o alinhamento seqüencial e a comparação da sequência a ser analisada com um banco de dados de sequências com funções já conhecidas e assim atribuir a função à proteína.

[00055] A etapa de predição funcional a partir de uma estrutura tridimensional engloba as etapas de:

(i) modelagem,

(ii) filtragem,

(iii) comparação com um banco de dados e

(iv) predição da função a partir da estrutura 3D para posterior comparação com a predição a partir da estrutura primária que é realizada paralelamente.

[00056] Importante destacar que para se inferir uma função à nova sequência, o alinhamento entre as sequências primárias deve ser de alta qualidade, podendo mesmo assim conduzir a predições de baixa confiabilidade, sendo que essa questão pode ser superada mediante o uso de técnica de modelagem comparativa para inferir a função para uma proteína, tal como é proposto por esse invento.

[00057] Dessa forma, a etapa de modelagem trata-se de uma modelagem comparativa que se baseia na construção de modelos 3D e na

identificação de motivos estruturais que auxiliam na determinação da função protéica.

[00058] Desse modo, a modelagem comparativa é a técnica mais bem sucedida de predição de estruturas tridimensionais (3D). Esta abordagem baseia-se no conhecimento de que a conformação estrutural de uma proteína é mais conservada que sua estrutura primária durante o processo evolutivo, e que pequenas mudanças na sequência, em geral, resultam em sutis modificações na estrutura tridimensional.

[00059] Além disso, proteínas homólogas apresentam regiões internas conservadas (principalmente constituídas dos elementos de estrutura secundária, hélices-α e fitas-β) e as principais diferenças estruturais ocorrem nas suas regiões externas, constituídas principalmente por alças (*loops*), que ligam os elementos de estruturas secundárias.

[00060] Como mencionado, a similaridade estrutural entre duas proteínas pode ser sugerida pela comparação de sequências de aminoácidos. Por exemplo, se uma determinada proteína de levedura, que possua sua estrutura tridimensional resolvida experimentalmente, tiver sequência primária similar a uma proteína humana de interesse, é então possível utilizá-la como molde para construir um modelo estrutural para a proteína humana, pelo método de modelagem comparativa.

[00061] Para esse processo de obtenção de um modelo protéico tridimensional, por meio da estratégia da modelagem comparativa, são necessárias basicamente quatro etapas principais:

- identificação de referências de estrutura;;
- alinhamento entre sequências – alvo e molde(s);
- construção do modelo;
- validação do modelo

[00062] A acurácia exigida para um modelo depende do propósito da sua utilização. Se o modelo proposto para a proteína for utilizado para desenho de ligantes, ele deve ter alta precisão do sítio de ligação. Porém, se o

interesse for identificar interfaces de interação com outras moléculas, utilizando padrões de superfícies, ou inferir funções de proteínas para fins de predição funcional, o modelo não precisa ter uma alta precisão. Portanto, modelos tridimensionais de proteínas podem ser utilizados para diversos fins e a acurácia requerida de cada um vai depender da(s) aplicação(ções) que se pretende dar a ele.

[00063] Um dos fatores importantes para a correta predição estrutural das proteínas é o alinhamento sequencial utilizado e a escolha de um bom molde para construção do seu modelo tridimensional.

[00064] Nesse estágio é fundamental realizar um refinamento da busca de uma estrutura tridimensional que será o molde para sequência que se deseja inferir a função. Para isso são identificadas sequências que podem ser utilizadas nas técnicas de modelagem comparativa com a escolha do molde para construção do modelo 3D.

[00065] Portanto, o refinamento na busca da estrutura que será utilizada como molde para a sequência que se pretende determinar a função é fundamental para a etapa de modelagem da estrutura 3D.

[00066] Em uma forma de realização a modelagem comparativa compreende a identificação de sequências que serão utilizadas para a modelagem comparativa e a escolha do molde a ser utilizado na construção da estrutura 3D.

[00067] Em uma forma de realização preferencial da presente invenção a modelagem comparativa pode utilizar um programa denominado BATS (*BLAST Automatic Targeting for Structures*) que identifica as sequências que serão utilizadas e também os moldes aplicáveis às referidas sequências. Além disso, o BATS filtra os resultados provenientes do BLAST, deposita-os em um banco de dados próprio e analisa os resultados do BLAST. Nesta análise é gerada uma pontuação final própria (Sf) que é determinada pelos valores filtrados do BLAST para número de aminoácidos alinhados, identidade e similaridade, e por mais dois índices calculados pelo BATS, o LVI (*Lenght*

Variation Index) e o GRSI (*Gap Relative Strengh Index*), conforme equação (I) evidenciada a seguir:

$$S_f = W_iP_i + W_sP_s + W_nP_n + W_{LVI}P_{LVI} + W_{GRSI}P_{GRSI} \quad (I),$$

em que W_i, W_s, W_n, W_{LVI} e W_{GRSI} são, respectivamente, os pesos de identidade, similaridade, número de aminoácidos alinhados, LVI e GRSI, e P_i, P_s, P_n, P_{LVI} e P_{GRSI} correspondem à pontuação atribuída à porcentagem de identidade, porcentagem de similaridade, número de aminoácidos alinhados, LVI e GRSI, respectivamente.

[00068] O LVI é usado para identificar dentre as sequências similares, aquela que possui um número de aminoácidos mais próximo do número de aminoácidos da sequência alvo, ou seja, aquela que oferece uma melhor cobertura, sendo que LVI ≤ 0,1 é equivalente a uma cobertura ≥ 90%.

[00069] Já o GRSI de cada alinhamento serve para identificar, dentre as sequências similares, as que possuem o menor número de intervalos possível, posicionados de forma pouco dispersa ao longo do alinhamento com a sequência alvo.

[00070] Após a modelagem da estrutura 3D, é realizada a filtragem dos modelos gerados.

[00071] Essa é a etapa fundamental do método da presente invenção, pois os filtros separam os modelos gerados na etapa anterior, de acordo com a qualidade necessária para se inferir função à estrutura tridimensional.

[00072] Assim a etapa de filtragem define quais modelos são bons para seguir para etapa seguinte e aqueles que podem ser descartados. Três filtros são aplicados para analisar os modelos gerados e garantir a qualidade da predição funcional prevista pelo método a partir da estrutura 3D da proteína.

[00073] Essa análise é realizada em três sub-etapas:

- definição de valores de identidade e cobertura,
- avaliação da qualidade estereoquímica e

- avaliação dos valores de alinhamento estrutural entre molde utilizado e modelo gerado.

[00074] As sub-etapas podem ser realizadas em qualquer ordem, sendo que preferencialmente primeiro é realizada a definição de valores de identidade e cobertura, seguidos pela avaliação dos valores de qualidade estereoquímica e, por fim, a avaliação dos valores de alinhamento estrutural entre molde utilizado e modelo gerado.

[00075] Essa filtragem garante que dados em larga escala serão analisados de forma rápida, uma vez que aqueles modelos não interessantes são descartados. Além disso, esse estágio também é responsável por garantir a confiabilidade dos resultados do método da invenção.

[00076] Dessa forma, uma sub-etapa da filtragem baseia-se em critérios de identidade e cobertura do alinhamento sequencial feito entre a proteína alvo e a proteína identificada como sendo o melhor molde a ser utilizado para a construção da estrutura tridimensional da proteína. Esses critérios compõem uma classificação em grupos é dada em função de percentuais de identidade cobertura, conforme tabela a seguir:

[00077] Tabela 1: Classificação gerada para as proteínas que tiveram modelos tridimensionais construídos no seu sistema.

Grupo	Identidade	Cobertura
(Muito Alto) G1	>= 75%	>= 90%
Alto G2	>= 50% e < 75%	>= 90%
Bom G3	>= 50%	>= 70% e < 90%
MédioBom G4	>= 35% e < 50%	>= 70%
MédioBaixo G5	>= 25% e < 35%	>= 70%
Baixo G6	>= 25%	>= 50% e < 70%
Muito Baixo G7	>= 25%	>= 30% e < 50%

[00078] De acordo com essa classificação, apenas as estruturas que apresentem identidade superior ou igual a 25 % e cobertura superior ou

igual a 70% serão submetidas a próxima etapa. Em contrapartida, as estruturas que não satisfizerem essa demanda, não serão consideradas.

[00079] Em continuidade, outra sub-etapa de filtragem é composta pela avaliação da qualidade estereoquímica dos modelos construídos e em uma forma de realização preferida, o método da presente invenção utiliza o Gráfico de Ramachandram para essa finalidade.

[00080] O Gráfico de Ramachandran define os valores permitidos dos ângulos de torção φ e ψ em estruturas tridimensionais indicando os resíduos que se encontram nas regiões mais favoráveis e desfavoráveis, prevendo conflitos estéricos, e orientando na avaliação da qualidade dos modelos teóricos ou experimentais de proteínas.

[00081] Com base nesse gráfico essa sub-etapa realiza a avaliação estereoquímica das estruturas, sendo que os parâmetros de avaliação são de que pelo menos 75% dos aminoácidos da estrutura gerada esteja em regiões favoráveis e no máximo 4% em regiões proibidas do Gráfico de Ramachandran.

[00082] A outra sub-etepa de filtragem é baseada na avaliação dos valores de alinhamento estrutural entre molde utilizado e modelo gerado. Isso pode ser realizado com base na técnica descrita como RMSD.

[00083] O RMSD é a media do desvio médio entre as estruturas das proteínas alinhadas, ou seja, o desvio encontrado entre duas proteínas durante alinhamento estrutural, sendo a referência o molde. O RMSD é calculado conforme equação a seguir:

$$R = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2} \qquad \text{(I)}$$

em que δ é a distância entre os N pares de átomos equivalentes.

[00084] Após a etapa de filtragem e eliminação dos modelos ruins é realizada a predição estrutural por domínios e superfamília de cada estrutura alvo.

[00085] O método, bem como o sistema da presente invenção tem aplicações nas mais diversas áreas para análise de diferentes amostras oriundas das ciências ômicas: genômica, metagenômica, transcriptômica, proteômica etc.

[00086] Conforme já mencionado, uma outra modalidade do presente invento aborda um sistema para predição da função de proteínas.

[00087] Esse sistema compreende meios que são responsáveis pela implementação das etapas descritas pelo método e permite ao usuário que as realize de forma automatizada e eficiente para qualquer número de proteínas.

[00088] De tal modo o sistema compreende meios para identificação de sequência de aminoácidos; meios para predição da função a partir de uma estrutura primária da referida sequência de aminoácidos; meios para predição da função a partir de uma estrutura terciária da referida sequência de aminoácidos e meios para comparação dessas funções preditas.

[00089] Há que se salientar que os meios para predição da função a partir de uma estrutura primária funcionam independentemente dos meios para predição da função a partir de uma estrutura terciária, sendo que ambos podem atuar em paralelo sem que um atrapalhe o funcionamento do outro.

[00090] Os meios para predição da função a partir de uma estrutura primária da referida sequência de aminoácidos são responsáveis por realizar o alinhamento sequencial e a comparação com banco de dados de funções de sequências e os meios para predição da função a partir de uma estrutura terciária da referida sequência de aminoácidos compreendem ainda meios de modelagem, meios de filtragem e bancos de dados de famílias e de domínio de funções.

[00091] **Exemplos**

[00092] Exemplo 1 – Determinação em proteínas da bactéria endofítica *Gluconacetobacter diazotrophicus.*

A bactéria endofítica Gram-negativa *Gluconacetobacter diazotrophicus* é normalmente encontrada em plantas como cana-de-açúcar, abacaxi, café e batata-doce, sendo responsável pela fixação de nitrogênio atmosférico e, pela potência de diferentes benefícios nas plantas hospedeiras (Current Science, 83:137-145, 2002; Proteomics 8:1631-1644, 2008). 3778 sequências protéicas foram preditas pelo genoma da *G. diazotrophicus* PAL5 [Refseq: NC_010125]

[00093] Assim, foi utilizado o método e o sistema da presente invenção para investigar essas 3778 sequências com o intuito de comparar os seus resultados com aqueles obtidos a partir de determinações feitas pelos métodos convencionais, que se baseiam somente em estruturas primárias das proteínas.

[00094] Foram construídos modelos para 1773 sequências de proteínas utilizando a modelagem de estrutura tridimensional e em seguida foram utilizados os filtros descritos anteriormente para a validação dos modelos construídos e seleção daqueles que teriam a predição funcional determinada a partir da estrutura.

[00095] O primeiro filtro selecionou apenas estruturas de acordo com os seguintes parâmetros: e-value > 0.00001, que correspondem a identidade > 25% e cobertura >70%.

[00096] Durante a etapa de validação dos modelos, verificou-se a importância da utilização dos três filtros utilizados pelo método e sistema da presente invenção.

[00097] Com a utilização somente dos valores de cobertura e identidade, 106 proteínas seriam descartadas; com o uso do filtro de qualidade estereoquímica, 223 proteínas seriam descartadas pelos parâmetros de qualidade definidos no Gráfico de Ramachandran e se fosse usado o filtro de alinhamento estrutural entre molde e modelo, 220 proteínas teriam sido descartadas.

[00098] Porém, com a utilização da etapa de filtragem da presente invenção em que os três parâmetros descritos no parágrafo anterior são usados em conjunto e, preferencialmente, em uma sequência específica, foram então descartadas 383 proteínas, sendo que algumas foram descartadas por somente um dos métodos.

[00099] As Figuras 3A-C mostram modelos de boa qualidade e a importância da utilização de mais de um filtro para validação dos modelos construídos na etapa de modelagem. Nessas figuras, a estrutura em preto mostra os modelos construídos pelo método e sistema da presente invenção, enquanto que as estruturas em cinza mostram as proteínas obtidas a partir do banco de dados PDB (Protein Data Bank).

[000100] Na Figura 3A, observamos o modelo gerado para a proteína gi162146620 com boas características estruturais e que foram validadas pelos 3 filtros utilizados. Em B, o modelo gerado para a proteína de gi162148534. Em C, modelo gerado para a proteína de gi162146689. Os valores de cada filtro para os modelos gerados A, B e C são mostrados conforme Tabela 1 a seguir:

[000101] Tabela1: Resultados das etapas de filtragem para três modelos de estruturas tridimensionais geradas pela etapa de modelagem comparativa da presente invenção.

Filtro	Modelos Gerados		
	A	B	C
Identidade (I) e Cobertura (C) (%)	35% > I > 50% C > 70%	I >25%; 70%> C >50%	I >25%; 50%> C >30%
Validação estereoquímica (%)	91,4% em regiões favoráveis e 0% nas regiões proibidas	78.4% em regiões favoráveis e 1.4% nas regiões proibidas	43.7% em regiões favoráveis e 10.3% nas regiões

			proibidas
Alinhamento estrutural entre molde e modelo (A de RMSD)	0,884 A	11,349 A	1.525 A

[000102] É importante observar que devido às etapas de filtragem que compõem o método da presente invenção, os modelos B e C foram considerados inaptos a ter a função protéica determinada e, portanto, foram descartados pela etapa de filtragem definida pelo método da presente invenção, sendo que apenas o modelo A foi considerado apto para a predição de sua função. Muito embora o modelo B estivesse dentro dos parâmetros definidos como aceitáveis para qualidade estereoquímica o mesmo não apresentou os mesmos resultados para o alinhamento estrutural entre o modelo gerado e molde, enquanto que o modelo C não mostrou um valor aceitável na validação estereoquímica, apresentando valores fora dos padrões estabelecidos no presente método para regiões favoráveis e regiões desfavoráveis.

[000103] Sem a utilização dessa etapa de filtragem os modelos B e C teriam sua função determinada, mas essa resposta não teria confiabilidade que é conseguida pelo método e sistema da presente invenção, graças à etapa de filtragem.

[000104] Exemplo 2: Categorias Funcionais

[000105] Foram separados 1390 modelos com boa qualidade estrutural que passaram pela etapa de filtragem em 2 grupos: (1) proteínas que já possuíam uma função predita somente utilizando estrutura primária e (2) proteínas preditas como hipotéticas no banco de dados.

[000106] Dessa forma, com os dados oriundos de (1), foi possível traçar um paralelo entre as predições funcionais oriundas das duas metodologias (predição a partir da estrutura primária e predição a partir da

estrutura terciária), onde 66% das proteínas obtiveram atribuição da função da proteína alvo igual a função já predita pela anotação convencional.

[000107] Já com os dados oriundos de (2), foi possível realizar a predição de função utilizando exclusivamente a abordagem estrutural para 86 proteínas, das 96 pertencentes a esse grupo, demonstrando a importância do método e sistema da presente invenção como ferramenta na genômica estrutural.

[000108] Na figura 4 é possível observar as categorias funcionais preditas para essas 86 proteínas, que anteriormente eram caracterizadas apenas como proteínas hipotéticas pelo banco de dados (que utiliza apenas a abordagem sequencial para inferir função).

REIVINDICAÇÕES

1 **–** Método para predição da função de proteínas, caracterizado por compreender as etapas:

(a) fornecimento de uma sequência de aminoácidos;

(b) predição da função a partir de uma estrutura primária da sequência de aminoácidos fornecida na etapa (a) ;

(c) predição da função a partir de uma estrutura terciária da sequência de aminoácidos fornecida na etapa (a)

(d) comparação das etapas (b) e (c)

(e) determinação da função da proteína.

2 - Método, de acordo com a reivindicação 1, caracterizado pela etapa (b) compreender o alinhamento seqüencial, a comparação com um banco de dados de sequências e a obtenção de modelo tridimensional.

3 – Método, de acordo com as reivindicações 1 e 2, caracterizado pela etapa (c) compreender ainda:

(i) modelagem de estrutura tridimensional;

(ii) filtragem dos modelos da etapa (i);

(iii) comparação com banco de dados para atribuição de função.

4 – Método, de acordo com as reivindicações 3, caracterizado pela etapa (i) de modelagem compreender a identificação da sequência de aminoácidos, o alinhamento dessa sequência, construção e a validação de modelo tridimensional.

5 – Método, de acordo com as reivindicações 1 a 4, caracterizado pela etapa (ii) de filtragem compreender:

- definição de valores de identidade e cobertura da estrutura tridimensional gerada,

- avaliação da qualidade estereoquímica e

- avaliação dos valores de alinhamento estrutural entre molde utilizado e modelo gerado.

6 – Método, de acordo com a reivindicação 5, caracterizado pelas sub etapas de filtragem serem realizadas em qualquer ordem, sendo preferencialmente realizada: a definição de valores de identidade e cobertura, em seguida avaliação da qualidade estereoquímica e avaliação dos valores de alinhamento estrutural.

7 – Método, de acordo com as reivindicações 1 a 6, caracterizado pela sub-etapa de definição de valores de identidade e cobertura selecionar apenas estrutura com identidade igual ou superior a 25% e cobertura igual ou superior a 70%.

8 – Método, de acordo com as reivindicações 1 a 7, caracterizado pela sub-etapa de avaliação estereoquímica utilizar o Gráfico de Ramachandram.

9 – Método, de acordo com as reivindicações 7 e 8, caracterizado pela avaliação estereoquímica selecionar estrutura com pelo menos 75% de aminoácidos em regiões favoráveis e no máximo 4% de aminoácidos em regiões proibidas.

10 – Método, de acordo com as reivindicações 1 e 9, caracterizado pela sub-etapa (iii) de alinhamento estrutural utilizar a técnica de RMSD.

11 – Método, de acordo com as reivindicações 1 a 10, caracterizado por analisar sequências genômicas, metagenômicas, proteômicas e transcriptômicas.

12 – Sistema para predição da função de proteínas, caracterizado por compreender:

- meios para identificação de sequência de aminoácidos;

- meios para predição da função a partir de uma estrutura primária da referida sequência de aminoácidos;

- meios para predição da função a partir de uma estrutura terciária da referida sequência de aminoácidos;

- meios para comparação de funções preditas.

13 - Sistema, de acordo com a reivindicação 12, caracterizado pelos meios para predição da função a partir de uma estrutura primária da referida

sequência de aminoácidos realizarem o alinhamento sequencial e comparação com banco de dados de funções de sequências.

14 – Sistema, de acordo com as reivindicações 12 e 13, caracterizado pelos meios para predição da função a partir de uma estrutura terciária da referida sequência de aminoácidos compreenderem ainda meios de modelagem, meios de filtragem e bancos de dados de famílias e de domínio de funções.

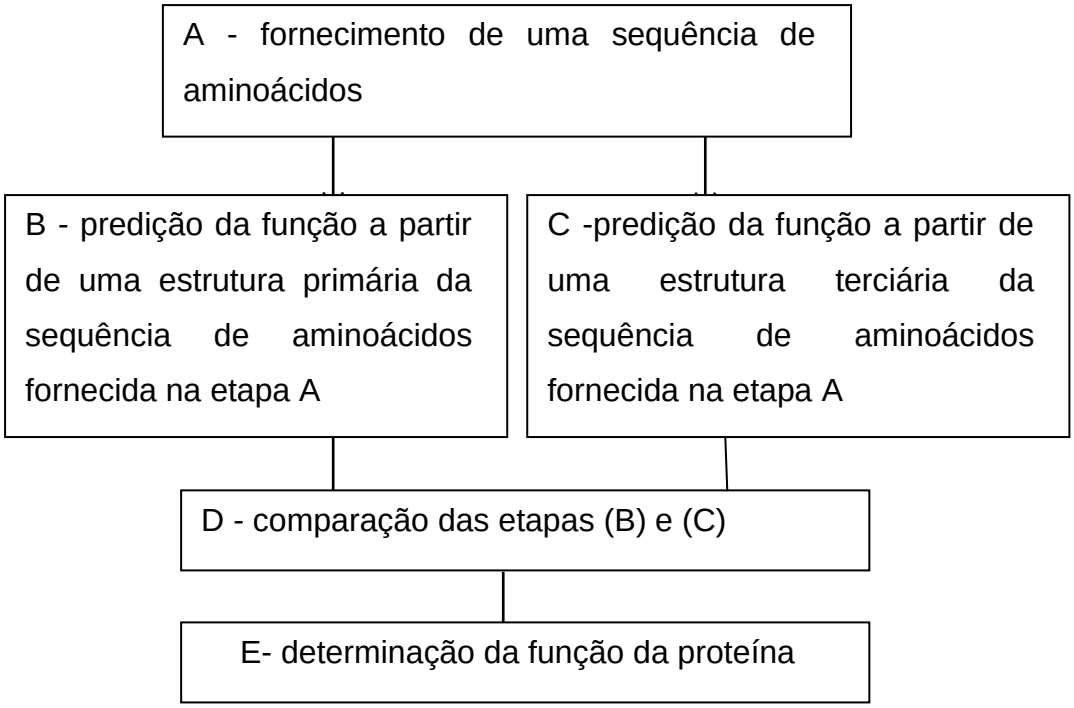

A - fornecimento de uma sequência de aminoácidos
B - predição da função a partir de uma estrutura primária da sequência de aminoácidos fornecida na etapa A
C -predição da função a partir de uma estrutura terciária da sequência de aminoácidos fornecida na etapa A
D - comparação das etapas (B) e (C)
E- determinação da função da proteína

Figura 1

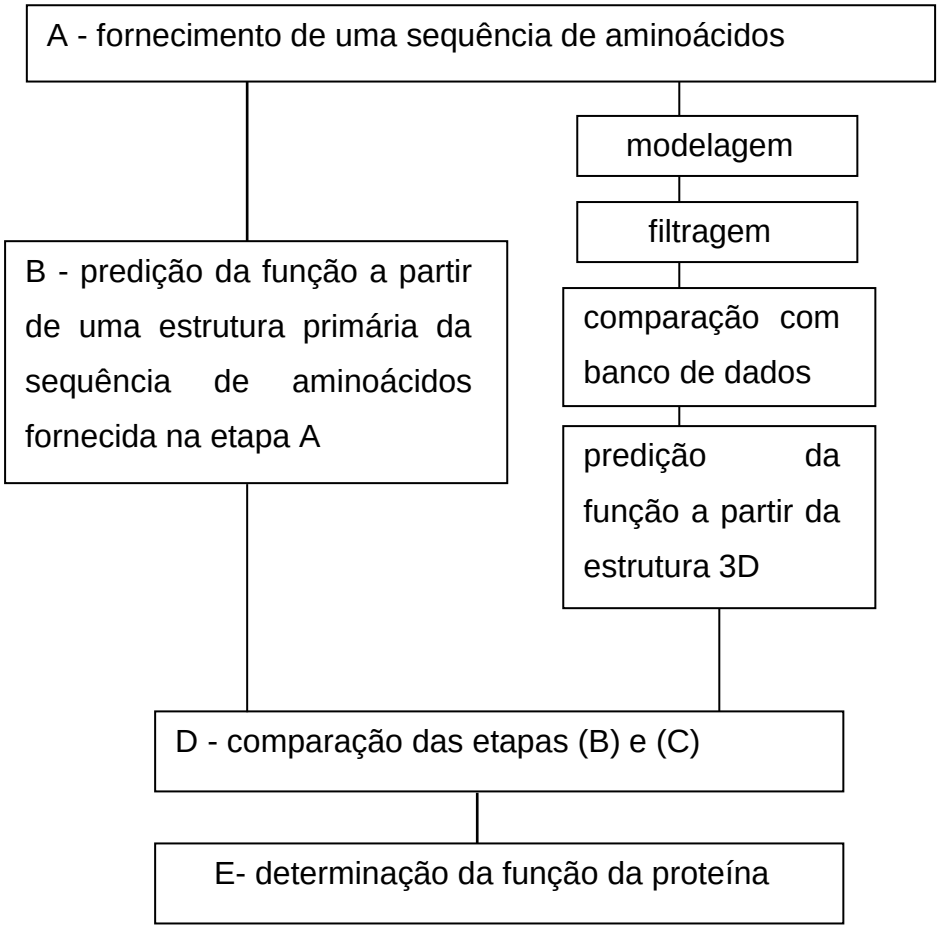

A - fornecimento de uma sequência de aminoácidos
modelagem
filtragem
B - predição da função a partir de uma estrutura primária da sequência de aminoácidos fornecida na etapa A
comparação com banco de dados
predição da função a partir da estrutura 3D
D - comparação das etapas (B) e (C)
E- determinação da função da proteína

Figura 2

Figura 3A

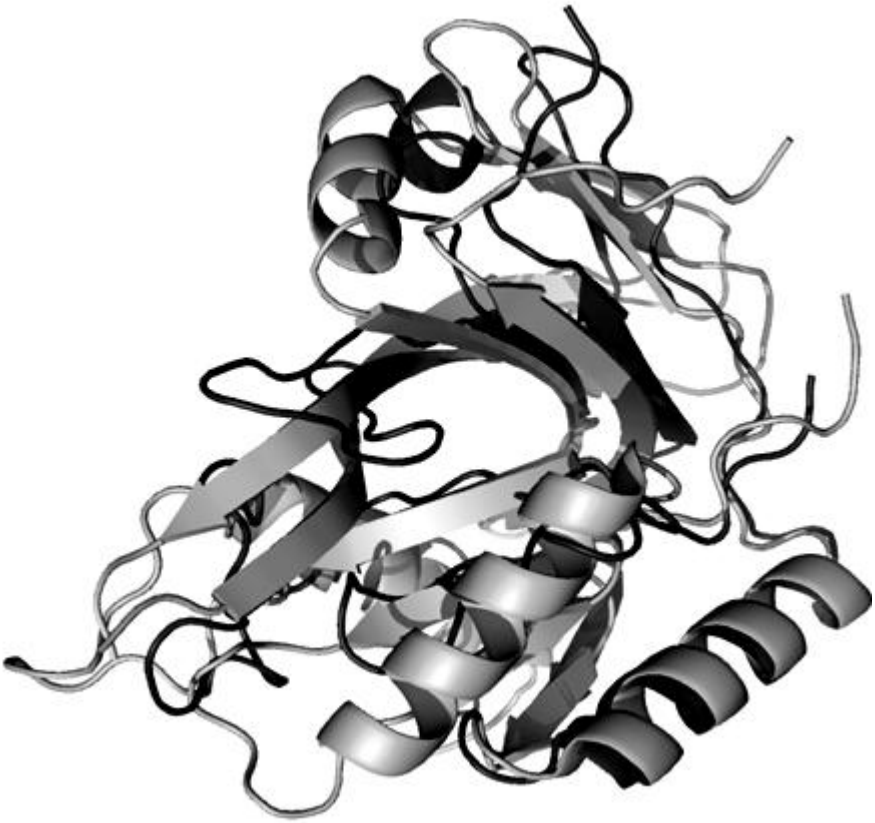


Figura 3B

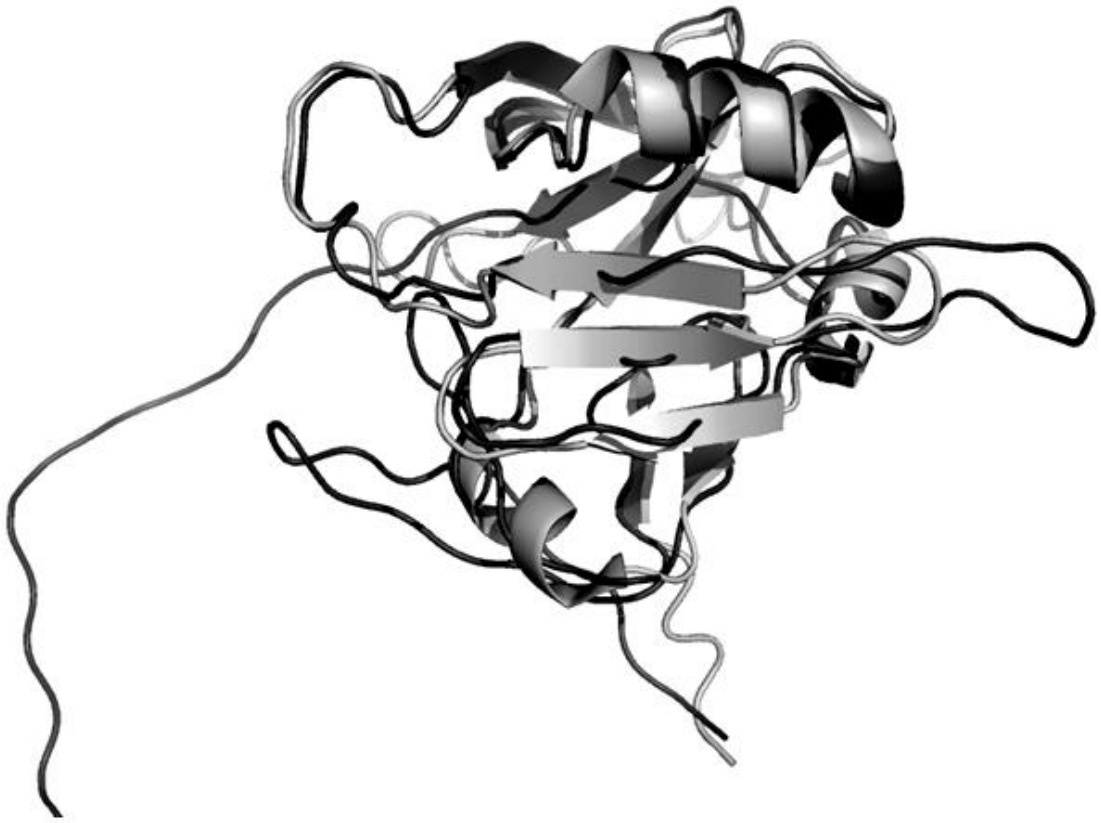


Figura 3C

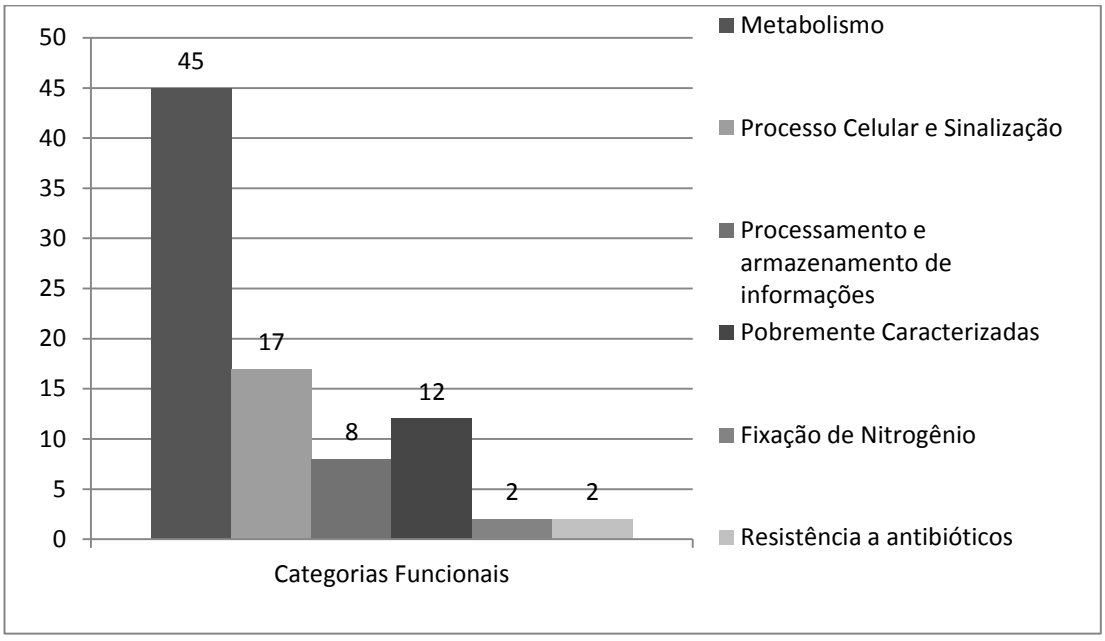


Figura 4

RESUMO

MÉTODO E SISTEMA PARA PREDIÇÃO DE FUNÇÕES DE PROTEÍNAS

A presente invenção refere-se a métodos e sistemas relacionados à predição da função de sequência de aminoácidos oriundas de abordagens genômicas, metagenômicas, proteômicas e transcriptômicas em larga escala. O método para predição da função de uma proteína, a partir de uma sequência de aminoácidos, que compreende a predição direta a partir da sequência de aminoácidos e a predição a partir da estrutura tridimensional da referida sequência de aminoácidos com a posterior comparação entre as duas para gerar o resultado.