

UNIVERSIDADE FEDERAL DO PAMPA

ANNA CAROLINA BOECK

**CONSTRUÇÃO DE UMA PLATAFORMA PARA DETERMINAÇÃO *IN SILICO* DA
ALERGENICIDADE E ANTIGENICIDADE DE PROTEÍNAS**

São Gabriel

2013

ANNA CAROLINA BOECK

**CONSTRUÇÃO DE UMA PLATAFORMA PARA DETERMINAÇÃO *IN SILICO* DA
ALERGENICIDADE E ANTIGENICIDADE DE PROTEÍNAS**

Trabalho de Conclusão de Curso apresentado à comissão avaliadora de graduação em Biotecnologia da Universidade Federal do Pampa *campus* São Gabriel, RS, Brasil, para obtenção do título de Bacharel em Biotecnologia.

Orientador: Juliano Tomazzoni Boldo

**São Gabriel
2013**

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

B669c Boeck, Anna Carolina
Construção de uma Plataforma para Determinação in silico da
Alergenicidade e Antigenicidade de Proteínas / Anna Carolina
Boeck.
59 p.

Trabalho de Conclusão de Curso(Graduação)-- Universidade
Federal do Pampa, BACHARELADO EM BIOTECNOLOGIA, 2013.
"Orientação: Juliano Tomazzoni Boldo".

1. OGM. 2. Alergenicidade. 3. Antigenicidade . 4. Predição
in silico. 5. Bioinformática. I. Título.

ANNA CAROLINA BOECK

**CONSTRUÇÃO DE UMA PLATAFORMA PARA DETERMINAÇÃO *IN SILICO* DA
ALERGENICIDADE E ANTIGENICIDADE DE PROTEÍNAS**

Trabalho de Conclusão de Curso apresentado à banca avaliadora de graduação em Biotecnologia da Universidade Federal do Pampa *campus* São Gabriel, RS, Brasil, para obtenção do título de Bacharel em Biotecnologia.

Área de concentração: Ciências Biológicas

Trabalho de Conclusão de Curso defendido e aprovado em: 18 de outubro de 2013.

Banca examinadora:

Prof. Dr. Juliano Tomazzoni Boldo

Orientador

Universidade Federal do Pampa - UNIPAMPA

Prof. Dr. Paulo Marcos Pinto

Universidade Federal do Pampa - UNIPAMPA

Prof. Dr. Filipe de Carvalho Victoria

Universidade Federal do Pampa - UNIPAMPA

Para minha mãe, Arcia Boeck, e meu irmão,
Leonardo Boeck, meus mentores e heróis.

AGRADECIMENTO

À minha mãe, Arcia, pelos bons exemplos de uma vida toda, pela incansável luta tanto na minha formação quanto na dos meus irmãos, pelo amor incondicional e por, simplesmente, ser minha mãe.

Aos meus irmãos, Leonardo e Eric, pela proteção, carinho e companheirismo. Léo, obrigada pelo auxílio com este trabalho, ele só foi possível porque tu estiveste ao meu lado, serei eternamente grata.

Ao Prof. Juliano Boldo, queridíssimo orientador, pela oportunidade, pelos ensinamentos e por acreditar que eu seria capaz.

Ao Prof. Paulo pela paciência e pelos longos diálogos que sempre renderam bons conselhos.

À Prof. Scheila e ao Prof. Daniel, da Universidade de Caxias do Sul, pela disponibilidade e receptividade para com este trabalho.

Ao Anderson, pela amizade de sempre e pela hospitalidade durante minha passagem pela UCS.

Aos meus amigos de Santa Maria, em especial, Leticia e Janaína, por aceitarem e entenderem que nem sempre eu pude estar presente. Obrigada pelo apoio, pela alegria dos finais de semana em casa e desculpem minha ausência. Parafraseando Fabrício Carpinejar “Amigo é o que fica depois da ressaca. É glicose no sangue.”.

Aos amigos que fiz ao longo desta jornada em São Gabriel. Priscila, Andressa e Diogo vocês foram essenciais.

Aos colegas e também amigos, Jeferson, Mauro e Nathane, pela amizade e pela oportunidade de poder conviver com vocês.

Obrigada a todos que, mesmo não estando citados aqui, tanto contribuíram para a conclusão desta etapa e para a Anna Carolina que sou hoje.



RESUMO

Com o avanço no sequenciamento de diferentes genomas, a construção de organismos geneticamente modificados (OGM), especialmente vegetais, tem aumentado. Uma das preocupações envolvendo a expressão de diferentes proteínas heterólogas é a possibilidade do desenvolvimento de reações alérgicas ou antigênicas nos consumidores finais. Assim, a WHO (*World Health Organization*) e a FAO (*Food And Agriculture Organization Of The United Nations*) propuseram testes *in silico* para a determinação de índices de alergenicidade e antigenicidade de proteínas passíveis de serem utilizadas na construção de OGM. Com o objetivo de congregiar diversos algoritmos de predição que permitissem a análise *in silico* de proteínas e definissem seu potencial alergênico ou antigênico, construiu-se uma plataforma, chamada Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas (PA³P). Nesta plataforma, reuniu-se os bancos de dados PIR, SwissProt/TrEMBL e GenBank e as *suites* AlgPred e Allermatch. Primeiramente houve a identificação das linguagens de programação de cada banco de dados, com posterior pesquisa através dos *sites* visando identificar suas interfaces e a forma como as consultas são realizadas. Alguns *sites* têm interface através do REST, enquanto outros não oferecem nenhum tipo de informação. Identificadas as interfaces, buscou-se alternativas na linguagem de programação Hypertext Preprocessor (PHP versão 5.4.3). Tal linguagem foi escolhida por apresentar muitas bibliotecas, uma comunidade de desenvolvimento ativa, adequada documentação e por ser de livre acesso. Utilizou-se o WampServer (versão 2.2) que permite criar aplicações *web* com Apache2 (versão 2.2.22), PHP e um banco de dados MySQL (versão 5.5.24). Para os *sites* com REST utilizou-se o conjunto de operações definidas por HTTP. Para aqueles que não ofereciam nenhuma interface a maneira encontrada foi realizar o envio remoto dos formulários de consulta via cURL, respeitando as configurações originais dos *sites*. Para as interfaces de entrada e saída de dados foi utilizado HyperText Markup Language (HTML). As saídas foram tratadas via expressões regulares para que obtivéssemos apenas o resultado uma vez que alguns *sites* oferecem diferentes formatos de resposta ao *web service*. Para a validação de PA³P, construiu-se um banco de sequências contendo 30 proteínas sabidamente alergênicas, 30 proteínas sabidamente não alergênicas e 30 proteínas cujos índices de alergenicidade e antigenicidade eram desconhecidos. As sequências foram submetidas para análise tanto na PA³P quanto nos sites individualmente. A validação demonstrou que nossos dados são condizentes com os dados gerados quando as sequências são submetidas aos algoritmos originais, porém com tempo reduzido. Assim, pelo uso da referida plataforma, o

usuário tem uma considerável economia de tempo para executar todas as análises necessárias. O uso de PA³P é considerado confiável como indicado pela validação desta ferramenta. Trabalhos futuros incluem a persistência dos dados e a construção de uma análise estatística, além da adição de novos algoritmos de análise e interface gráfica.

Palavras-chave: OGM. Alergenicidade. Antigenicidade. Predição *in silico*. Bioinformática.

ABSTRACT

With the advancement in sequencing of different genomes, the construction of genetically modified organisms (GMO), especially crop plants, has increased. One of the concerns involving the expression of different heterologous proteins is the possibility of development of allergic or antigenic reactions on final consumers. Thus, the WHO (World Health Organization) and the FAO (Food And Agriculture Organization Of The United Nations) proposed *in silico* tests to determine levels of antigenicity and allergenicity of proteins that could be used in the construction of GMO. Aiming to congregate several prediction algorithms that allow the *in silico* analysis of proteins and define their allergenic or antigenic potential, a platform called Platform for Analysis of Allergenicity and Antigenicity of Protein (PA³P) was built. In this platform the databases PIR, SwissProt/TrEMBL and GenBank and their AlgPred and Allermatch *suites* were congregated. Firstly, the identification of programming languages for each database was conducted with subsequent search through the sites aiming to identify their interfaces and how queries are performed. Some sites presented interface using REST, while others don't provide any information. Once the interfaces were identified, alternatives in the programming language Hypertext Preprocessor (PHP version 5.4.3) were sought. Such language was chosen because it has many libraries, a community of active development, adequate documentation and to be open access. The WampServer (version 2.2) was used since it allows creating *web* applications with Apache2 (version 2.2.22), PHP and a MySQL database (version 5.5.24). For the *sites* with REST the set of operations defined by HTTP was used. For those that offered no interface, the most adequate operation was performing the remote sending of enquiry forms via cURL, respecting the original settings of the *sites*. For the interfaces of input and output data HyperText Markup Language (HTML) was used. The outputs were treated via regular expressions to get only the desired results, once some sites offer different answer formats to the Webservice. To validate PA³P a sequence database was constructed, which contains 30 knowingly allergenic protein sequences, 30 non allergenic protein sequences, and 30 sequences from proteins whose levels of allergenicity and antigenicity were unknown. The sequences were submitted for analysis on both PA³P and sites individually. The validation process showed that our data are consistent with the data generated when the sequences were subjected to the algorithms individually, but with reduced analysis time. Thus, by using the referred platform, the user needs considerably less time when performing all necessary tests. The use of PA³P is considered reliable as

indicated by the validation of this tool. Future works includes data persistence, the inclusion of statistical analysis, the addition of new algorithms of analysis, and graphical interface.

Keywords: GMO. Allergenicity. Antigenicity. *In silico* prediction. Bioinformatics.

APRESENTAÇÃO

No item **INTRODUÇÃO**, consta revisão da literatura sobre os temas trabalhados neste Trabalho de Conclusão de Curso (TCC).

A metodologia realizada e os resultados obtidos que fazem parte deste TCC estão apresentados sob a forma de manuscrito, que se encontra no item **MANUSCRITO**. No mesmo constam as seções: Introdução, Material e Métodos, Resultados e Discussão e Referências Bibliográficas.

O item **CONSIDERAÇÕES FINAIS**, encontrado no final deste trabalho apresenta interpretações e comentários gerais sobre os resultados do manuscrito apresentado neste trabalho.

Nas **REFERÊNCIAS** mencionam-se somente as citações que aparecem nos itens **INTRODUÇÃO** e **CONSIDERAÇÕES FINAIS** deste TCC.

LISTA DE FIGURAS

Figura 1 - Principais técnicas e ferramentas da Engenharia Genética.	14
Figura 2 - Diferença entre um OGM e um melhoramento genético clássico.	15
Figura 3 - Árvore de decisão da FAO.	17
Figura 4 - <i>Pipeline</i> de análise da plataforma.	19

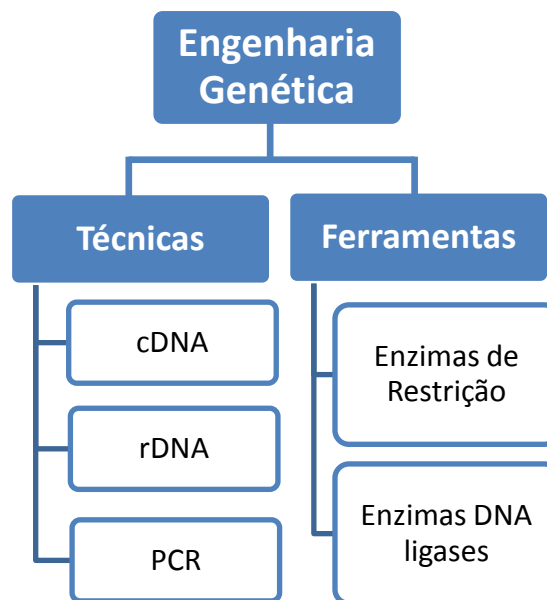
SUMÁRIO

1	INTRODUÇÃO	14
2	OBJETIVOS.....	20
	2.1 Objetivo Geral.....	20
	2.2 Objetivos Específicos	20
3	MANUSCRITO	21
4	CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS	42
5	REFERÊNCIAS	43
	APÊNDICES	45
	APÊNDICE A – Código Fonte Referente ao AlgPred Compilado na Plataforma	46
	APÊNDICE B – Código Fonte Referente ao Allermatch Compilado na Plataforma.....	47
	APÊNDICE C – Código Fonte Referente ao Blastp Compilado na Plataforma.....	48
	APÊNDICE D – Código Fonte Referente à Mensagem de Erro Compilado na Plataforma	49
	APÊNDICE E – Código Fonte Referente ao Index Compilado na Plataforma.....	50
	APÊNDICE F – Código Fonte Referente ao PeptideMatch Compilado na Plataforma	53
	APÊNDICE G – Código Fonte Referente ao UniProt Compilado na Plataforma.....	54
	APÊNDICE H – Código Fonte Referente à Página de Respostas Compilado na Plataforma	56

1 INTRODUÇÃO

O termo engenharia genética refere-se ao uso de técnicas capazes de inserir fragmentos de DNA, geralmente genes, de um organismo a outro, mudando, portanto, a constituição genética do receptor (SHRADER-FRECHETTE, 2005). As principais técnicas e ferramentas utilizadas na engenharia genética são apresentadas na Figura 1. Dentro deste panorama, a utilização de tais técnicas permite a construção de Organismos Geneticamente Modificados (OGM), comumente chamados de transgênicos. Diferentemente das técnicas tradicionais de melhoramento genético, sejam elas cruzamentos pré-definidos (Figura 2A) ou a utilização de raios UV, a engenharia genética permite o controle total da alteração pretendida (Figura 2B). Apenas o gene de interesse é inserido no organismo e nenhuma outra alteração pode ser observada (HUG, 2008). Assim, podem ser construídos OGM expressando proteínas que confiram vantagens econômicas, como vegetais mais resistentes a pragas ou a herbicidas (VON KRIES & WINTER, 2011).

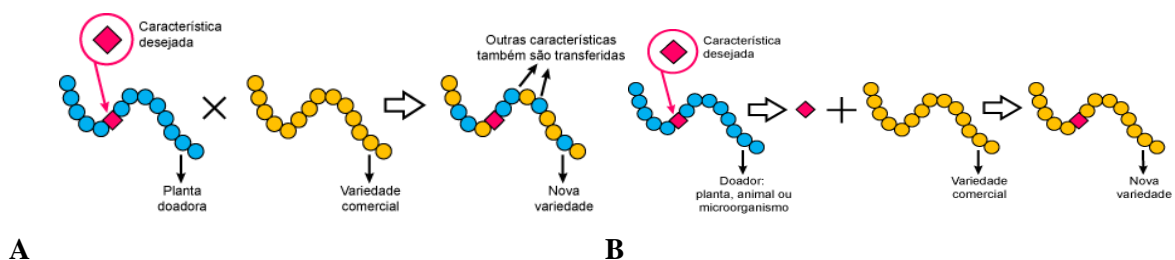
Figura 1 - Principais técnicas e ferramentas da Engenharia Genética.



Fonte: adaptado de <http://oportaldomundobiologico.blogspot.com.br/2011/04/engenharia-genetica-tecnicas.html>. Acesso em 10/09/2013.

Os maiores impasses para a ampla comercialização de alimentos transgênicos são o desenvolvimento de testes padronizados para garantir sua segurança e a confiabilidade do consumidor nos produtos finais, uma vez que a modificação genética de tais organismos leva a expressão de proteínas antes ausentes. A FAO (*Food and Agriculture Organization of the United Nations*) define biossegurança como um conjunto de medidas destinadas a evitar os riscos, para a saúde humana e para a conservação ambiental, advindos da prática de diferentes tecnologias aplicadas em pesquisa e práticas comerciais. Para minimizar um possível impacto negativo devem ser analisados os possíveis efeitos sobre a segurança alimentar bem como aqueles relacionados com o meio ambiente (FAO/WHO, 2001). Os principais perigos potenciais dos OGM podem estar associados com toxicidade, alergenicidade e antigenicidade, alterações nutricionais e efeitos antinutrientes e possibilidade remota de transferência horizontal de genes (COSTA et al., 2011) sendo que todos estes perigos devem ser testados antes da liberação do organismo transgênico e monitorados após a liberação para comercialização. No Brasil, a Lei de Biossegurança (Lei nº 11.105 de 24 de março de 2005) foi aprovada para normalizar a pesquisa, o manuseio e o uso de transgênicos.

Figura 2 - Diferença entre um OGM e um melhoramento genético clássico. Em (A) tem-se um esquema de como ocorre no melhoramento genético clássico e em (B) como um organismo geneticamente modificado é produzido.



Fonte: Adaptado de <http://ciencia.hsw.uol.com.br/transgenicos3.htm>. Acesso em 10/09/2013.

O termo alergenicidade refere-se às características daquele que é alergênico. Os alérgenos são proteínas que possuem a capacidade de induzir respostas alérgicas pela eliciação de anticorpos imunoglobulina E (IgE) (MOHABATKAR et al., 2013), ou seja, proteínas alergênicas são responsáveis pela indução da produção de anticorpo IgE e causam a liberação de mediadores inflamatórios (MARI et al., 2009). Tais proteínas apresentam peso molecular que varia de 10 a 70 kDa e exibem características de solubilidade e estabilidade ao

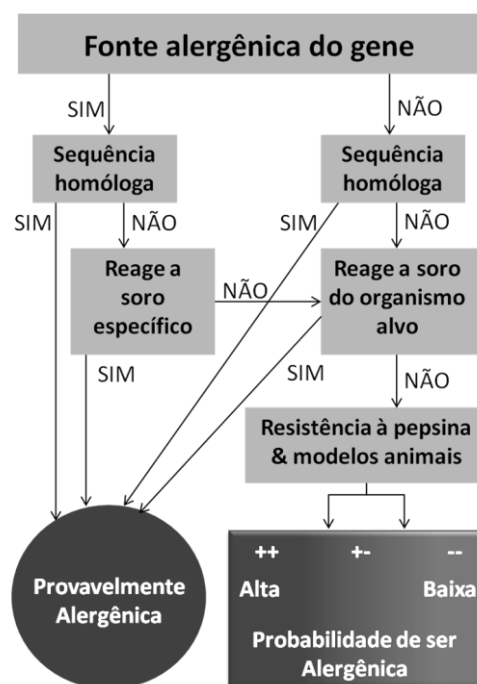
calor (ZORZET, GUSTAFSSON, HAMMERLING, 2002). A identificação de epítomos elicitores de IgE é fundamental em alérgenos alimentares (SHARMA, GAUR, ARORA, 2013). O potencial de alergenicidade de uma proteína pode ser previsível por meio de testes *in silico* (MOHABATKAR et al., 2013), *in vivo* e *in vitro*. Contudo, não há relatos de um experimento universal e confiável para a avaliação de produtos alergênicos (ZORZET, GUSTAFSSON, HAMMERLING, 2002).

A antigenicidade refere-se à capacidade de um agente ou fração do mesmo estimular a formação de anticorpos. Os anticorpos são proteínas produzidas em resposta à exposição a antígenos (ABBAS, LICHTMAN, PILLAI, 2008). As moléculas de IgE, ligadas a receptores, funcionam como receptores naturais para antígenos. As substâncias liberadas em resposta a ligação anticorpo-antígeno são as responsáveis pelos sintomas das reações alérgicas (ALBERTS, BRAY, LEWIS, 2010).

A introdução de proteínas recombinantes em alimentos (alimentos geneticamente modificados), medicamentos e outros produtos podem expor os consumidores a novos riscos (MARI et al., 2009). Assim, a avaliação do potencial de alergenicidade e antigenicidade das proteínas codificadas pelos genes inseridos é imprescindível, uma vez que não é facilmente previsível, sem estudos específicos. Para resolver esta questão a FAO e a *World Health Organization* (WHO), em 2001, determinaram alguns testes que devem ser realizados para identificar características que definem o potencial alergênico e antigênico de uma nova proteína introduzida em alimentos geneticamente modificados. Também foi criada, pela FAO (2001) uma árvore de decisão para a avaliação do potencial de alergenicidade de proteínas introduzidas no OGM (KLETER & KUIPER, 2002). De acordo com a árvore de decisão da FAO (Figura 3), primeiramente, deve-se comparar a estrutura da nova proteína com as estruturas de alérgenos conhecidos através do alinhamento das sequências. O conhecimento acerca da origem do gene é de extrema importância, pois caso esta seja de um alérgeno conhecido a sua reação com soros de pacientes que são alérgicos à fonte específica deve ser testada. Testes adicionais como, por exemplo, a digestão *in vitro* com pepsina, podem ser necessários. A FAO/WHO (2001) indica que testes adicionais devem ser realizados quando não houver nenhuma homologia de sequência com um alérgeno conhecido. O teste de resistência à pepsina é realizado quando o resultado para o teste de reação a soro específico é negativo (FAO/WHO, 2001). O teste de digestão *in vitro* com pepsina é interessante considerando-se que a maioria dos alérgenos alimentares são estáveis à digestão (KLETER &

KUIPER, 2002). Entretanto, deve-se considerar que a promoção ou a redução da degradação de proteínas pode estar relacionada à presença de inibidores de proteases ou outras substâncias. Se qualquer uma destas etapas produzir um resultado positivo, o OGM deve ser considerado passível de ser alergênico (FAO/WHO, 2001). Contudo, a verificação destas características isoladas não é suficiente e por vezes apresenta resultados duvidosos, como falsos positivos ou proteínas com alergenicidade comprovada negligenciadas. Assim, a utilização de vários algoritmos de avaliação torna-se necessária (KLETER & KUIPER, 2002).

Figura 3 - Árvore de decisão da FAO.



Fonte: Adaptado de Kleter & Kuiper (2002).

Neste contexto, alguns algoritmos foram desenvolvidos e hoje são amplamente utilizados, o que torna as avaliações mais robustas. Contudo, a análise por meio destes algoritmos é dificultada por dois problemas principais: (i) a dificuldade de acesso aos *sites* que hospedam tais algoritmos, muitas vezes com interface pouco amigável; e, (ii) resultados em gráficos ou tabelas não autoexplicativos e de qualidade gráfica baixa, comprometendo a análise realizada pelo leigo em bioinformática. Assim, o desenvolvimento de uma plataforma que congregue tais análises, somadas aos testes sugeridos pela FAO/WHO, pouparia tempo de análise por proteína avaliada e ofereceria ao usuário um resultado mais compreensível e confiável.

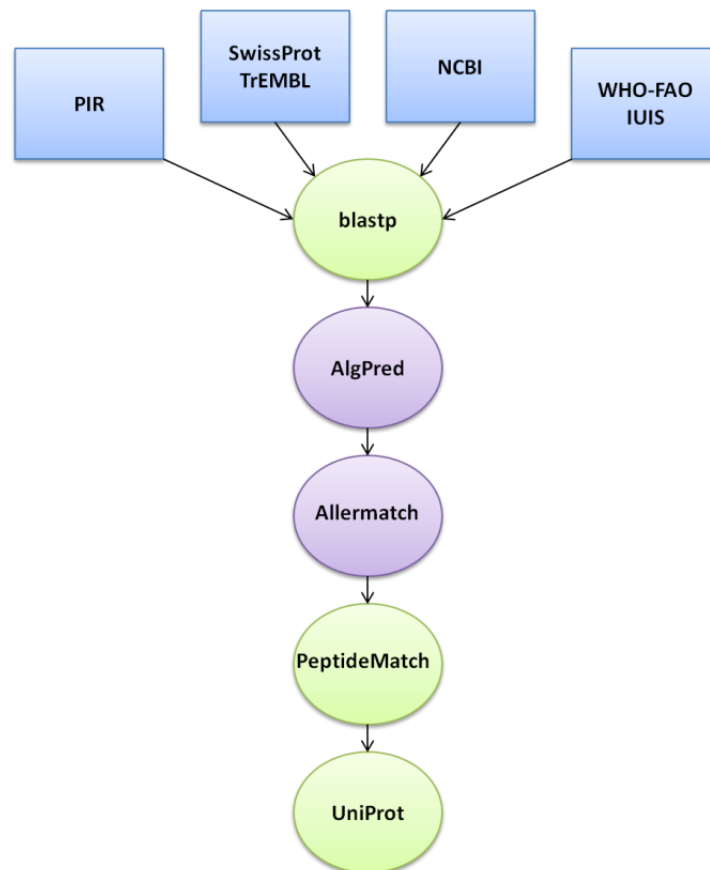
Vários destes algoritmos são baseados em alinhamento de sequências. O alinhamento é a operação primitiva mais importante para a bioinformática sendo a base para outras operações mais complexas (TICONA, 2003). Por meio desta técnica podem ser encontradas medidas de similaridade e/ou identidade com proteínas reconhecidas como alergênicas. Para o Comitê *CODEX Alimentarius* (órgão criado pela FAO/WHO para desenvolver normas alimentares visando à proteção da saúde do consumidor) uma proteína é considerada como potencialmente alergênica quando possuir 35% de identidade em uma janela de 80 aminoácidos ao longo de toda a sequência proteica. Para a FAO e a WHO seis a oito aminoácidos contíguos (identidade) indicam uma proteína com potencial alergênico.

Softwares como FASTA, CLUSTALW, BLASTP e PSI-BLAST são as ferramentas mais populares para o alinhamento de sequências (LIEW, YAN, YANG, 2005). Estes alinhamentos são baseados em bancos de dados disponíveis publicamente. Banco de dados é uma coleção de dados inter-relacionados que organiza e estrutura as informações. Grande parte dos bancos de dados estão vinculados ao SGBD (Sistema de Gerenciamento de Bancos de Dados), sendo os principais gerenciadores de bancos de dados o MySQL, PostgreSQL, ORACLE e o Microsoft SQL Server. Os principais bancos de dados utilizados em bioinformática são Genbank, EBI, PIR (*Protein Information Resource*), Swiss-Prot, KEGG, entre outros (PROSDOCIMI et al., 2002).

A *suite* Allermatch dispõe de um banco de dados com 730 sequências de proteínas alergênicas, sendo 396 sequências provenientes do banco de dados da WHO-IUIS, 98 sequências oriundas do banco de dados Swiss-Prot e as demais (236 sequências) são comuns aos dois bancos de dados (FIERS et al., 2004). Segundo Saha & Raghava (2006), a *suite* AlgPred possui um banco de dados com 2890 proteínas alergênicas. O banco de dados UniProtKB possui duas seções, UniProtKB/Swiss-Prot e UniProtKB/TrEMBL. O primeiro é formado por sequências anotadas manualmente e o segundo consiste de anotação automática (The UniProt Consortium, 2007). Dados do UniProt indicam que o banco de dados Swiss-Prot apresenta 540.958 sequências, sendo que 70% delas têm sua existência inferida a partir de homologias (dados disponíveis em <http://web.expasy.org/docs/relnotes/relstat.html> acessado em 30/09/2013). Por outro lado, o TrEMBL apresenta, aproximadamente, 42.821.879 sequências sendo que mais de 70% destas são preditas (dados disponíveis em <http://www.ebi.ac.uk/uniprot/TrEMBLstats/> acessado em 30/09/2013).

Posto isso, este trabalho buscou construir uma plataforma – chamada PA³P (Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas; acesso <http://www.pa3p.zz.mu>) – que contribuirá para embasar a decisão de utilizar ou não determinada proteína na construção de OGM com finalidade nutricional. A plataforma conta com três seções: (i) as análises sugeridas pela FAO/WHO; (ii) alinhamentos em diferentes bancos de dados que possuem sequências de proteínas comprovadamente ou potencialmente alergênicas e antigênicas; e (iii) análises por diferentes algoritmos de predição por busca de motivos e epítomos (Figura 4).

Figura 4 - *Pipeline* de análise da plataforma. Os quadrados azuis representam os bancos de dados públicos de onde as sequências de interesse são extraídas. As elipses lilás referem-se às *suites* compiladas na plataforma e as verdes, as ferramentas também compiladas.



Fonte: Do autor (2013, dados não publicados).

2 OBJETIVOS

2.1 Objetivo Geral

Construção de uma plataforma como ferramenta de pesquisa conectada a diversos bancos de dados, permitindo a análise do potencial alergênico e antigênico de diferentes proteínas na construção de transgênicos.

2.2 Objetivos Específicos

- Compilar as informações e tecnologias dos diferentes *sites* de pesquisa e bancos de dados públicos em um único local;
- Permitir a análise *in silico* de proteínas e definir seu potencial alergênico ou antigênico;
- Reduzir consideravelmente o tempo gasto por compilar todas as análises necessárias em uma única plataforma;

3 MANUSCRITO

CONSTRUÇÃO DE UMA PLATAFORMA PARA DETERMINAÇÃO *IN SILICO* DA ALERGENICIDADE E ANTIGENICIDADE DE PROTEÍNAS

Anna Carolina Boeck¹, Paulo Marcos Pinto², Scheila de Ávila e Silva³, Milena Schrenkel Homirch⁴, Juliano Tomazzoni Boldo⁵.

Com o avanço no sequenciamento de diferentes genomas, a construção de organismos geneticamente modificados (OGM), especialmente vegetais, tem aumentado. Uma das preocupações envolvendo a expressão de diferentes proteínas heterólogas é a possibilidade do desenvolvimento de reações alérgicas ou antigênicas nos consumidores finais. Com o objetivo de congregiar diversos algoritmos de predição que permitissem a análise *in silico* de proteínas e definissem seu potencial alergênico ou antigênico, construiu-se uma plataforma, chamada Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas (PA³P – www.pa3p.zz.mu). Nesta plataforma, reuniu-se os bancos de dados PIR, SwissProt/TrEMBL e GenBank e as *suites* AlgPred e Allermatch. A partir da identificação das linguagens de programação de cada banco de dados e suas interfaces buscaram-se alternativas na linguagem de programação PHP. Utilizou-se o WampServer (versão 2.2) que permite criar aplicações web com Apache2 (versão 2.2.22), PHP (versão 5.4.3) e um banco de dados MySQL (versão 5.5.24). Para as interfaces de entrada e saída de dados foi utilizado HTML. As saídas foram tratadas via expressões regulares. A validação mostrou que nossos dados são condizentes com os dados gerados quando as sequências são submetidas aos algoritmos originais isoladamente. Desta forma, pelo uso da Plataforma, o usuário tem uma considerável economia de tempo para executar todas as análises necessárias. O uso de PA³P é considerado confiável como indicado pela validação desta ferramenta.

Palavras-chave: Alergenicidade. Antigenicidade. OGM. Predição *in silico*.

¹ Universidade Federal do Pampa – UNIPAMPA – São Gabriel

² Universidade Federal do Pampa – UNIPAMPA – São Gabriel

³ Instituto de Biotecnologia, Universidade de Caxias do Sul – UCS

⁴ Universidade do Vale do Rio dos Sinos - UNISINOS

⁵ Universidade Federal do Pampa – UNIPAMPA – São Gabriel

Introdução

Os genomas de diversos organismos foram parcialmente ou completamente sequenciados e anotados na década passada. A informação acumulada tem sido utilizada não somente para anotar novas sequências como também para construir bancos de dados cada vez mais específicos. A partir da anotação de sequências, informações acerca de funções biológicas são extraídas (FARIA-CAMPOS et al., 2006).

Com base nos conhecimentos acerca da função biológica de determinada sequência, esta pode vir a ser isolada e, através da engenharia genética, é possível que seja introduzida em um organismo receptor, conferindo novas características. Como resultado, ter-se-á um Organismo Geneticamente Modificado (OGM). A tecnologia de modificação genética é muito mais específica quando comparada ao melhoramento clássico, uma vez que apenas a característica de interesse será introduzida (HJÄLTÉN et al., 2013).

Nas décadas de 80 e 90, inúmeras sequências foram geradas através de técnicas de clonagem molecular (CHAPMAN, 2007). Desta forma, diversos alérgenos tiveram sua função biológica atribuída baseada em sua homologia com proteínas de função conhecida. Entretanto, diferentemente da sua provável função biológica, apenas o conhecimento da sequência e a identificação de sua homologia com outras sequências de proteínas não é suficiente para determinar o possível potencial alergênico (CHAPMAN, 2007).

Devido à introdução de alimentos derivados de cultivos geneticamente modificados no mercado, a comunidade científica, órgãos reguladores e associações internacionais intensificaram as discussões sobre os procedimentos de avaliação de risco para identificar o potencial de alergenicidade das proteínas recentemente introduzidas (FAO/WHO, 2001).

Para que novas proteínas, oriundas de modificações genéticas, possam estar presentes em alimentos é necessário que seu potencial alergênico seja avaliado (FIERS et al., 2004). A predição e avaliação do potencial alergênico é essencial tanto para a avaliação da segurança do consumidor final como é importante para outros fatores ambientais (WANG et al., 2013). Contudo, tal potencial não é facilmente previsível uma vez que depende de diversos fatores como, por exemplo, a variabilidade da resposta de IgE específicas (SOUZA JUNIOR & MARTINS, 2003).

A FAO (*Food And Agriculture Organization Of The United Nations*) juntamente com a WHO (*World Health Organization*), em 2001, definiram que uma proteína seria

possivelmente alergênica quando apresentasse: (i) seis ou mais aminoácidos contíguos; ou, (ii) mais de 35% de identidade em uma janela de 80 aminoácidos (FAO/WHO, 2001). Tais características isoladas não são suficientes para determinar um possível potencial alergênico. Conforme descrito por Stadler & Stadler (2003), quando utilizada a metodologia proposta pela FAO/WHO, 1 a cada 200 resultados positivos são verdadeiramente positivos, ou seja, representam um alérgeno.

Para cumprir a exigência de avaliação do potencial de alergenicidade de novas proteínas, diversas abordagens computacionais foram desenvolvidas (WANG et al., 2013). Tais abordagens investigam: (i) a fonte do gene; (ii) a homologia da sequência com alérgenos conhecidos; (iii) as reações de associação com IgEs de fonte sorológica de indivíduos alérgicos; e, (iv) propriedades físico-químicas da proteína codificada pelo gene introduzido (SOUZA JUNIOR & MARTINS, 2003).

Dentro do contexto das novas abordagens computacionais que avaliam o potencial de alergenicidade e antigenicidade, construiu-se uma plataforma chamada Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas (PA³P – disponível em <http://www.pa3p.zz.mu>). Esta plataforma congrega as análises sugeridas pela FAO/WHO, alinhamentos em diferentes bancos de dados que possuem sequências de proteínas comprovadamente ou potencialmente alergênicas e antigênicas e análises por diferentes algoritmos de predição por busca de motivos e epítomos.

Nesta plataforma, foram compilados as ferramentas *protein-protein* BLAST (*Basic Local Alignment Search Tool*) utilizando-se da matriz BLOSUM62, *position-specific iterative* BLAST, *Protein Information Resource* (PIR), *Universal Protein Resource* e as *suites Prediction of Allergenic Proteins and Mapping of IgE epitopes* (AlgPred) e Allermatch.

A ferramenta BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) é uma ferramenta de busca por similaridades de sequências, desenvolvida por Altschul et al. (1990). Uma sequência proteica ou nucleotídica é comparada com outras sequências, de um banco de dados, para identificar as regiões de alinhamento local e reportar os alinhamentos de acordo com o *score* (BERGMAN, 2007). Para Altschul et al. (1990), as informações geradas pela identificação da homologia da sequência são o primeiro indício relativo a função biológica desta sequência. BLAST é composto por outros cinco subprogramas, entre eles blastp e psi-blast, os quais foram utilizados neste trabalho. O *protein-protein* BLAST (blastp) realiza pesquisas proteína-proteína. O *position-specific iterative* BLAST (psi-blast) pode detectar

relações distantes entre proteínas através de iterações com as sequências detectadas no blastp e o banco de dados (BERGMAN, 2007).

A *suite* AlgPred (<http://www.imtech.res.in/raghava/algpred/>) é uma ferramenta *web* desenvolvida para a predição de proteínas alergênicas e para o mapeamento de epítomos de IgE em proteínas alergênicas (SAHA & RAGHAVA, 2006). Outra ferramenta *web* é a Allermatch.

A *suite* Allermatch (<http://allermatch.org>) é uma ferramenta para a predição do potencial de alergenicidade de proteínas e peptídeos de acordo com as recomendações da FAO/WHO (35% de similaridade ou pequenos segmentos de, ao menos, 6 aminoácidos em trechos em uma janela de 80 aminoácidos). A sequência de aminoácidos é comparada com as sequências de proteínas alergênicas conhecidas, obtidas a partir de bancos de dados (FIERS et al., 2004).

A *Protein Information Resource* (PIR - <http://pir.georgetown.edu/>) é uma ferramenta bioinformática de apoio à genômica, à proteômica e aos estudos científicos. Originalmente este recurso foi criado para auxiliar a identificação e interpretação das informações oriundas das sequências proteicas. O banco de dados PIR-NREF possui sequências advindas do PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept e PDB (WU et al., 2003).

O banco de dados UniProt (*Universal Protein Resource* - <http://www.uniprot.org/>) é uma central de sequências de proteínas e de anotação funcional. O UniProt é composto por quatro componentes principais: (i) UniParc; (ii) UniProtKB; (iii) UniRef (The UniProt Consortium, 2007); e, recentemente, (iv) UniMES. A PA³P utiliza do UniProtKB (UniProt Knowledgebase) que possui duas partes: UniProtKB/TrEMBL e UniProtKB/Swiss-Prot.

A construção da plataforma PA³P (Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas) visa contribuir para embasar a decisão de utilizar ou não determinada proteína na construção de OGM com finalidade nutricional. A plataforma conta com três seções: (i) as análises sugeridas pela FAO/WHO; (ii) alinhamentos em diferentes bancos de dados que possuem sequências de proteínas comprovadamente ou potencialmente alergênicas e antigênicas; e, (iii) análises por diferentes algoritmos de predição por busca de motivos e epítomos. O objetivo deste trabalho foi construir um gerador de relatórios multianálise como ferramenta de pesquisa conectada a diversos bancos de dados, permitindo a análise *in silico* do potencial alergênico e antigênico de diferentes proteínas na construção de

transgênicos. Como resultado, obtivemos a redução considerável do tempo gasto por compilar todas as análises necessárias em uma única plataforma.

Material e Métodos

1. Coleta de Dados

Partiu-se do princípio de que cada *site* tem seu formato próprio, ou seja, são linguagens de programação heterogêneas. Desta forma, o primeiro passo foi a identificação do tipo e se era ofertado alguma interface e, ainda, como as consultas são realizadas. Este processo foi realizado mediante consultas aos *sites* de interesse. Identificadas as interfaces buscou-se alternativas na linguagem de programação PHP (*Hypertext Preprocessor*) para construção da PA³P.

2. Linguagens de Programação

2.1 PHP

A tecnologia PHP foi criada por Rasmus Lerdorf em 1994. Esta é uma linguagem de *script open source* de uso geral, que pode ser incorporada dentro do HTML (*HyperText Markup Language*) (MARRA, 2009; ARNTZEN et al., 2013). A escolha pela linguagem PHP deve-se ao fato de esta não apresentar nenhum custo de licença, ou seja, por ser um *software* livre. Outras vantagens da linguagem incluem: criptografia de dados; definição de interfaces para *webservices*; manipulação de arquivos XML; e, documentação oficial em português, dentre outras (MELO & NASCIMENTO, 2007). Determinados bancos de dados utilizados neste trabalho apresentam uma interface através do REST (*Representational State Transfer*), a base para a arquitetura *web* moderna. Segundo Melo e Nascimento (2007), PHP é considerada uma das linguagens de scripts mais utilizadas para os ambientes de sistemas para a internet (Figura 1).

Figura 1 - Trecho de código em PHP.

```

1 <?php
2
3 /*
4 Página inicial
5 Contém o formulário de consulta e o botão de submit
6 */
7
8 /*
9 Declaração dos arquivos com as funções necessárias as consultas
10 */
11 require 'Blastp.php';
12 require 'psiBlast.php';
13 require 'Algpred.php';
14 require 'Allermatch.php';
15 require 'PeptideMatchURL.php';
16 require 'Uniprot.php';
17 require 'writeResult.php';
18 require 'testeURL.php';
19
20 //função chamada após clicar no botao submit
21 if (isset($_POST['seq'])) {
22     $RIDblastp = "";
23     $RIDpsiBlast = "";
24     $JobID = "";
25     $seq = $_POST['seq'];
26     //verifica se a sequência foi colada, ou seja, comprimento > 0
27     if(strlen($seq)>0) {

```

Fonte: Do autor (2013, dados não publicados).

2.2 HTML

A linguagem de programação HTML foi utilizada na entrada e na saída dos dados. Documentos construídos por esta linguagem são interpretados diretamente pelos navegadores *web* do usuário (MARRA, 2009) (Figura 2).

Figura 2 – Código em HTML.

```

1 <!DOCTYPE html><html>
2 <head>
3 </head>
4 <body>
5 <table>
6 <tr>
7 <br>
8 Sequences producing High-scoring Segment Pairs:          Score  P(N)      N
9
10 SP:C01A2_BOVIN P02465 Collagen alpha-2(I) chain OS=Bos ta... 1119  2.2e-123  2
11 <br>L8HQF7_BOSMU L8HQF7 Collagen alpha-2(I) chain OS=Bos g... 1119  3.7e-121  2
12 SP:C01A2_CANFA O46392 Collagen alpha-2(I) chain OS=Canis ... 1078  4.7e-119  2
13 <br>F1PHY1_CANFA F1PHY1 Collagen alpha-2(I) chain OS=Canis... 1078  4.7e-119  2
14 <br>M3WVN3_FELCA M3WVN3 Uncharacterized protein OS=Felis c... 1080  6.1e-119  2
15 <br>F1SFA7_PIG F1SFA7 Uncharacterized protein OS=Sus scrof... 1073  1.6e-118  2
16 <br>G1MH95_AILME G1MH95 Uncharacterized protein OS=Ailurop... 1079  6.1e-117  2
17 <br>G3TIC0_LOXAF G3TIC0 Uncharacterized protein OS=Loxodon... 1058  7.8e-117  2
18 <br>B9VR89_EQUAS B9VR89 Collagen alpha-2 type I chain OS=E... 1057  2.6e-116  2
19 <br>F6RUA6_HORSE F6RUA6 Uncharacterized protein OS=Equus c... 1057  2.6e-116  2
20 </tr>
21 </table>
22 </body>
23 </html>

```

Fonte: Do autor (2013, dados não publicados).

3. Construção da Plataforma

Utilizou-se o WampServer (versão 2.2) que permite criar aplicações *web* com Apache2 (versão 2.2.22), PHP (versão 5.4.3) e um banco de dados MySQL (versão 5.5.24). Todo trabalho foi desenvolvido em um computador com sistema operacional Windows 7 de 64 Bits.

Alguns bancos de dados possuem interface através do REST. Para estes utilizou-se o conjunto de operações definidas no protocolo HTTP, respeitando as configurações originais dos bancos de dados.

REST é um conjunto coordenado de restrições arquiteturais que se comporta da seguinte maneira: uma rede de páginas da *web* forma uma máquina de estado virtual, permitindo que o usuário possa progredir através do aplicativo selecionando um *link* ou enviando dados escriturais, resultando em uma transição para o próximo estado da aplicação através da transferência de uma representação desse estado para o usuário (FIELDING & TAYLOR, 2002). Para os *sites* com arquitetura através do REST, utiliza-se o protocolo HTTP (*HyperText Transfer Protocol Secure*). O protocolo HTTP é do tipo *request/response*, sendo utilizado pela *World Wide Web* desde 1990.

Para os que não ofereciam nenhuma interface a maneira encontrada consiste em realizar o envio remoto dos formulários de consulta via cURL – POST (Figura 3). O método POST é projetado para permitir um método único para as seguintes funções: anotação dos recursos existentes, proporcionar um bloco de dados e estender um banco de dados. A função primordial executada pelo método POST é determinada pelo servidor, e geralmente é dependente da Request-URI (FIELDING et al., 1999).

Figura 3 – Método POST.

```

1 <?php
2 /*
6
7 function getAlgpred($seq){
8
9     //Inicia a biblioteca curl, ajustando o endereço alvo e demais parâmetros
10    $curl = curl_init();
11    curl_setopt($curl, CURLOPT_URL, "http://www.imtech.res.in/cgibin/algpred/algpred_main_temp.pl");
12    curl_setopt($curl, CURLOPT_RETURNTRANSFER, true);
13    // Metodo POST - Depende do site, ou POST ou GET
14    curl_setopt($curl, CURLOPT_POST, true);
15    // Seguir qualquer redirecionamento que houver na URL
16    curl_setopt($curl, CURLOPT_FOLLOWLOCATION, true);
17

```

Fonte: Do autor (2013, dados não publicados).

Para as interfaces de entrada e de saída de dados utilizou-se a linguagem de programação HTML. As saídas foram tratadas via expressões regulares (Figura 4) para que se

obtenha apenas o resultado pretendido uma vez que alguns *sites* oferecem diferentes formatos de resposta ao *WebService*.

Figura 4 – Tratamento de saídas.

```

30      // Define quais informacoes serao enviadas pelo POST ($dados)
31      curl_setopt($curl, CURLOPT_POSTFIELDS, $dados);
32      //Executa o envio
33      $resultado = curl_exec($curl);
34      $string_to_find = "Prediction by";
35      $string_onde_procurar = $resultado;
36      $localizacao = strpos($string_onde_procurar, $string_to_find);
37      $filtrado = substr($string_onde_procurar, $localizacao);
38      //Fecha a conexao
39      curl_close($curl);
40⊕    /*
44      wr("alg.html", $filtrado);
45      //retorna valor alertando que acabou esta consulta
46      return 1;

```

Fonte: Do autor (2013, dados não publicados).

4. Hospedagem

A hospedagem da plataforma foi realizada no Hostinger (<http://www.hostinger.com.br>), pois este oferece uma hospedagem gratuita com PHP. O plano gratuito de hospedagem oferta 2000 MB de espaço em disco, 10mbps de velocidade de rede do servidor e são fornecidas as estatísticas do *site*. Contudo, não há a garantia da disponibilidade do *site* devido a instabilidade do servidor.

5. Validação da Plataforma

A análise da confiabilidade se fez necessária para que esta ferramenta fosse validada. Para avaliar a confiabilidade desta nova ferramenta foram realizados testes com proteínas. Para tanto, foram eleitas 90 proteínas do banco de dados de proteínas do *National Center for Biotechnology Information* (<http://www.ncbi.nlm.nih.gov/>). Construiu-se um banco de sequências contendo 30 proteínas sabidamente alergênicas, 30 proteínas sabidamente não alergênicas e 30 proteínas cujos índices de alergenicidade e antigenicidade eram desconhecidos.

A metodologia definida para a escolha das proteínas foi possuir a palavra *allergen* em sua nomenclatura (grupo de proteínas alergênicas), ser uma proteína imunogênica humana

(grupo de proteínas sabidamente não alergênicas) e ser uma proteína cujos índices de alergenicidade e antigenicidade eram desconhecidos (grupo de proteínas desconhecidas).

Todas as proteínas selecionadas foram submetidas à análise tanto na Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas (PA³P) quanto às análises individualmente.

O ambiente computacional que se utilizou para teste foi um computador com memória RAM de 4GB, processadores Intel Centrino e Intel Core 2 Duo 2,0 GHz e sistema operacional Windows 7.

Resultados e Discussão

Consulta na plataforma

A Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas (PA³P) pode ser acessada em <http://www.pa3p.zz.mu> (Figura 5). Para realizar uma consulta na plataforma, deve-se inserir uma sequência proteica, a exemplo da Figura 6, no campo destinado para tal.

Figura 5 – Página inicial da plataforma.



Fonte: Do autor (2013, dados não publicados).

Figura 6 - Sequência de Blag g 4 isoalergen 1 de *Blattella germanica* em formato FASTA.

Bla g 4 isoalergen 1 [Blattella germanica]

GenBank: ACF53836.1

[GenPept](#) [Graphics](#)

```
>gi|194350815|gb|ACF53836.1| Bla g 4 isoalergen 1 [Blattella germanica]
MCITGVILFAVLAVCATDTLANEDCFRHESLVPLNDYKKFIGTWVIAAGTSEALTQYKCWNDLFFFNNAL
VSKYTDSKGKNRTTIRGRTKFEGNKFTIDYDDEGKAFSAPYSVLATDYDNYAIVEGCPAAANGHVIVVQL
RLTLRSFHPEQGDKEALQHYTVHQVNQHKAIEEDLKHFNLYEDLHSTCH
```

Fonte: Do autor (2013, dados não publicados).

Caso ocorra de o usuário não inserir a sequência corretamente e clicar no botão *submit*, um aviso é mostrado (Figura 7).

Figura 7- Aviso para que uma sequência seja inserida.



Favor colar a sequência de consulta!



Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas

PA³P

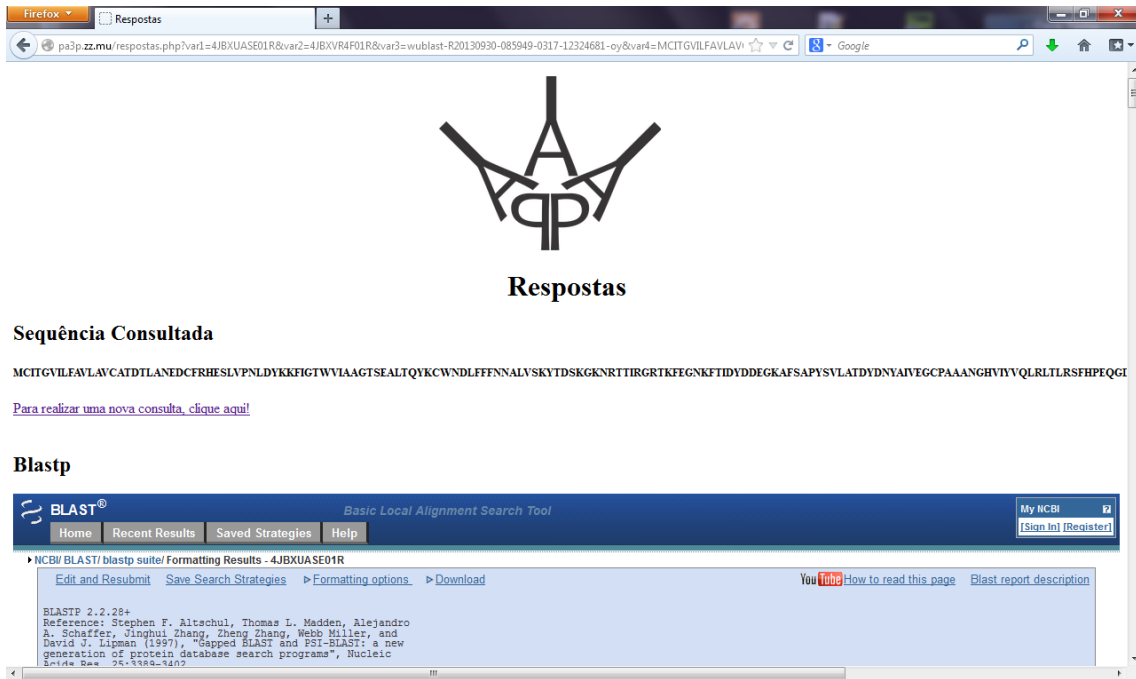
Insira a Sequência

Fonte: Do autor (2013, dados não publicados).

Respostas Impressas ao Usuário

O protocolo HTTP é do tipo *request-response* no modelo cliente-servidor, ou seja, quando o usuário (neste caso, o cliente) clicar no botão *submit* ele estará realizando uma requisição HTTP para o servidor. O servidor então retornará uma resposta ao usuário. A resposta (Figura 8) a requisição do usuário à plataforma, ou, simplesmente, a saída de dados da referida plataforma, é por meio de um documento em HTML (Figura 9).

Figura 8 - Padrão de resposta impressa à requisição do usuário.



Fonte: Do autor (2013, dados não publicados).

Figura 9 - Trecho do código em HTML que imprime a página de respostas visualizada pelo usuário.

```

1
2
3 <html>
4 <head>
5 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
6 <blockquote>
7 </blockquote>
8 <blockquote>
9 <blockquote>
10 <blockquote>
11 <blockquote>
12 <blockquote>
13 <blockquote>
14 <blockquote>
15 <blockquote>
16 <blockquote>
17 <p>
18 </p>
19 </blockquote>
20 </blockquote>
21 </blockquote>
22 </blockquote>
23 </blockquote>
24 </blockquote>
25 </blockquote>
26 </blockquote>
27 </blockquote>
28 </blockquote>
29 </blockquote>
30 </blockquote>
31 <title> Respostas </title>
32 </head>
33 <body>
34 <h1 align="center">Respostas</h1>
35 <h2 align="left">Sequência Consultada</h2><h5 align="left">MCITGVILFAVLAVCATDILANEDCFRHSLSVPLNDYKXKFGITWVIAAGTSEALTQYKCWNDLFFNNALVSKYDTSKGNRTTIRGRITKFEKGFIDYDDEGKAFSAPYVSLATDYDNYAIVEGCPAAANGHVIVYQLRLILRSFHPEQGI
36 </div>
37 <a href="http://pa3p.zz.mu/">Para realizar uma nova consulta, clique aqui!</a>
38 </div>
39 <div>
40 <div>
41 <h2 align="left">Blastp</h2><iframe id="ncbiBlastp" src="http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&RID=4JBXUASE01R&FORMAT=Align
42 </div>

```

Fonte: Do autor (2013, dados não publicados).

Pipeline de Análise da Plataforma

Seguindo o *pipeline* de análise da plataforma, a primeira etapa da consulta consiste no alinhamento da sequência de interesse. O alinhamento é a operação primitiva mais importante para a bioinformática (TICONA, 2003). É por meio do alinhamento que podem ser identificadas as medidas de similaridade/identidade com proteínas sabidamente alergênicas (SOUZA JUNIOR & MARTINS, 2003). Na PA³P, está presente o *software* BLAST para o alinhamento de sequências. Especificamente, estão presentes *blastp* e *psi-blast*.

Os dez principais alinhamentos obtidos pela consulta da Bla g 4 são apresentados na Figura 10 e na Figura 11. Tais resultados referem-se ao alinhamento realizado por intermédio da PA³P.

Figura 10 - Principais alinhamentos produzidos pela PA³P utilizando o *blastp*.

Sequences producing significant alignments:	Score (Bits)	E Value
gb ACF53836.1 Bla g 4 isoallergen 1 [Blattella germanica]	401	1e-140
gb ACF53837.1 Bla g 4 isoallergen 2 [Blattella germanica]	339	4e-116
gb ACJ37389.1 Bla g 4 allergen, partial [Blattella germanica]	338	1e-115
sp P54962.1 BLG4_BLAGE RecName: Full=Allergen Bla g 4; AltNam...	336	3e-115
pdb 3EBK A Chain A, Crystal Structure Of Major Allergens, Bla...	328	3e-112
gb ABP04043.1 Bla g 4 allergen [Blattella germanica]	327	2e-111
ref XP_004923377.1 PREDICTED: lopap-like [Bombyx mori]	53.1	7e-06
gb ABN11964.1 hypothetical protein [Maconellicoccus hirsutus]	48.5	5e-04
ref XP_001951488.1 PREDICTED: apolipoprotein D-like [Acyrtho...	44.3	0.009
gb EHJ763999.1 Lopap [Danaus plexippus]	44.3	0.010

Fonte: Do autor (2013, dados não publicados).

Figura 11 - Alinhamento de maior *score*.

```

ALIGNMENTS
>gb|ACF53836.1| Bla g 4 isoallergen 1 [Blattella germanica]
Length=191

Score = 401 bits (1031), Expect = 1e-140, Method: Compositional matrix adjust.
Identities = 191/191 (100%), Positives = 191/191 (100%), Gaps = 0/191 (0%)

Query 1 MCITGVILFAVLAVCATDTLANEDCFRHESLVPNLDYKKFIGTWVIAAGTSEALTQYKWCW 60
Sbjct 1 MCITGVILFAVLAVCATDTLANEDCFRHESLVPNLDYKKFIGTWVIAAGTSEALTQYKWCW 60

Query 61 NDLEFFNNALVSKYTDKSGKNRTTIRGRTKFEKNKFTIDYDDEGKAFSAPYSVLATDYDN 120
Sbjct 61 NDLEFFNNALVSKYTDKSGKNRTTIRGRTKFEKNKFTIDYDDEGKAFSAPYSVLATDYDN 120

Query 121 YAIVEGCPAAANGHVIVQLRLTLRSFHPEQGDKEALQHYTVHQVNQHKKAI EEDLKHFN 180
Sbjct 121 YAIVEGCPAAANGHVIVQLRLTLRSFHPEQGDKEALQHYTVHQVNQHKKAI EEDLKHFN 180

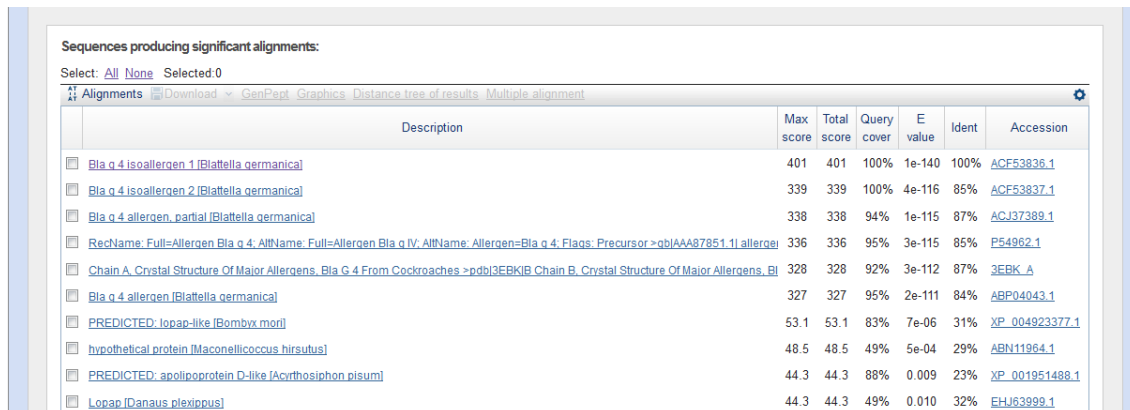
Query 181 LKYEDLHSTCH 191
Sbjct 181 LKYEDLHSTCH 191

```

Fonte: Do autor (2013, dados não publicados).

Estes resultados são os mesmos apresentados quando é realizada uma consulta diretamente no *site* da ferramenta (Figura 12 e Figura 13). O formato como os dados são apresentados ao usuário, é a única diferença que ocorre quando confrontados os resultados obtidos pela PA³P e pelo *site* individualmente.

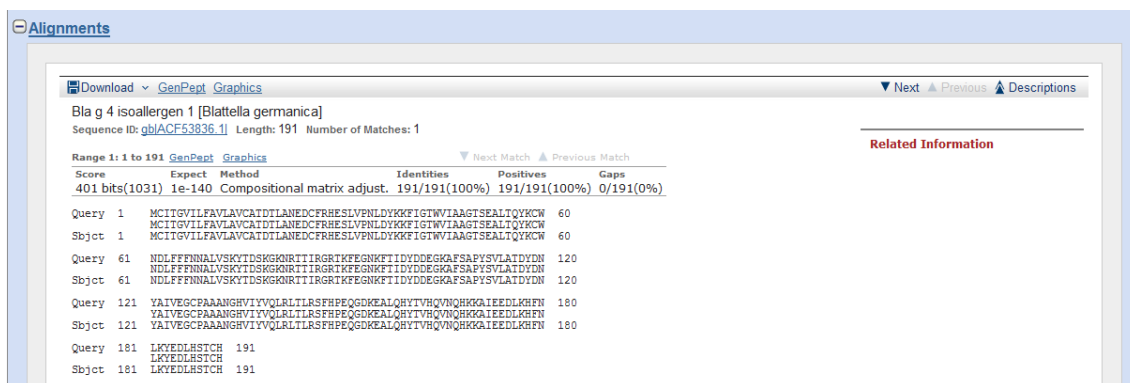
Figura 12 - Principais alinhamentos produzidos via blastp.



Description	Max score	Total score	Query cover	E value	Ident	Accession
Bla q 4 isoallergen 1 [Blattella germanica]	401	401	100%	1e-140	100%	ACF53836.1
Bla q 4 isoallergen 2 [Blattella germanica]	339	339	100%	4e-116	85%	ACF53837.1
Bla q 4 allergen, partial [Blattella germanica]	338	338	94%	1e-115	87%	ACJ37389.1
RecName: Full=Allergen Bla q 4; AltName: Full=Allergen Bla q IV; AltName: Allergen=Bla q 4; Flags: Precursor >gb AAA87851.1 allergei	336	336	95%	3e-115	85%	P54962.1
Chain A, Crystal Structure Of Major Allergens, Bla G 4 From Cockroaches >pdb 3EBKIB.Chain B, Crystal Structure Of Major Allergens, Bl	328	328	92%	3e-112	87%	3EBK_A
Bla q 4 allergen [Blattella germanica]	327	327	95%	2e-111	84%	ABP04043.1
PREDICTED: leopap-like [Bombyx mori]	53.1	53.1	83%	7e-06	31%	XP_004923377.1
hypothetical protein [Maconellicoccus hirsutus]	48.5	48.5	49%	5e-04	29%	ABN11964.1
PREDICTED: apolipoprotein D-like [Acyrthosiphon pisum]	44.3	44.3	88%	0.009	23%	XP_001951488.1
Lopap [Danaus plexippus]	44.3	44.3	49%	0.010	32%	EHJ63999.1

Fonte: Do autor (2013, dados não publicados).

Figura 13 - Destaque para o alinhamento de maior *score*.



Score	Expect	Method	Identities	Positives	Gaps
401 bits (1031)	1e-140	Compositional matrix adjust.	191/191 (100%)	191/191 (100%)	0/191 (0%)

```

Query 1  MCITGVILFAVLAVCAIDTLANEDCFRHSLSVPLDYKGFIGTWIAAGTSEALTQYKCN 60
Sbjct 1  MCITGVILFAVLAVCAIDTLANEDCFRHSLSVPLDYKGFIGTWIAAGTSEALTQYKCN 60

Query 61  NDLEFFNNALVSKYTDGKGRNRTIRGRITKFEKGFITIDYDEGKAFSAPYSVLATVDN 120
Sbjct 61  NDLEFFNNALVSKYTDGKGRNRTIRGRITKFEKGFITIDYDEGKAFSAPYSVLATVDN 120

Query 121 YAIVEGCPAAANGHVIVQLRLILRSFHFPEQSDKEALQHYTVHQVNHKKAIEEDLKHFN 180
Sbjct 121 YAIVEGCPAAANGHVIVQLRLILRSFHFPEQSDKEALQHYTVHQVNHKKAIEEDLKHFN 180

Query 181 LKYEDLHSTCH 191
Sbjct 181 LKYEDLHSTCH 191
  
```

Fonte: Do autor (2013, dados não publicados).

De acordo com Alberts, Bray e Lewis (2010), *e value* é um valor que representa quantas vezes se esperaria que um alinhamento ao acaso viesse a ocorrer. O valor de *e* é inversamente proporcional a significância da semelhança, ou seja, quanto menor o valor de *e*, mais significativa é a semelhança. Já para os valores de *score*, quanto mais alto este for, melhor é a semelhança. O *score* considera penalidades tanto para substituições quanto para *gap*.

A diferença observada entre a interface dos resultados apresentados pela PA³P e pelo blastp é consequência do uso de expressões regulares para filtragem dos dados. A saída dos dados da plataforma foi filtrada para que apenas os resultados pretendidos fossem mostrados e também para proporcionar ao usuário uma interface mais amigável. A plataforma apresenta apenas os dez primeiros alinhamentos para que desta forma não sejam demonstrados alinhamentos redundantes e/ou nem tão significativos. O mesmo ocorre com o psi-blast (dados não apresentados).

Os resultados obtidos pela *suite* AlgPred indicam que a Blag g 4 isoallergen 1 possui potencial alergênico (Figura 18 e Figura 19). O algoritmo utilizado nesta análise é o SVM (*Support Vector Machines*). O SVM é capaz de suportar um grande conjunto de dados.

Figura 14- Resultado da consulta, por meio da plataforma, à *suite* AlgPred.

Potential ALLERGEN
Score= 1.6268957 [Threshold= -0.4]
Positive Predictive Value= 85.64% Negative Predictive Value= 67.96%

Fonte: Do autor (2013, dados não publicados).

Figura 15 - Resultado obtido quando a sequência foi submetida na *suite* AlgPred individualmente.

AlgPred: Prediction of Allergenic Proteins and Mapping of IgE Epitopes

Name of sequence	Protein
Length of Sequence	191
Preicted On	Thu Sep 26 13:22:09 2013

Prediction by SVM method based on amino acid composition

Potential ALLERGEN
Score= 1.6268957 [Threshold= -0.4]
Positive Predictive Value= 85.64% Negative Predictive Value= 67.96%

Fonte: Do autor (2013, dados não publicados).

O valor do parâmetro *Positive Predictive Value* refere-se à proporção na qual os resultados positivos são verdadeiramente positivos. Assim como a taxa referente ao *Negative Predictive Value* identifica a taxa de resultados negativos que são diagnosticados como tal. Também são consideradas a especificidade e a sensibilidade neste teste estatístico. Ambos os parâmetros são descritos por Altman & Bland (1994).

A consulta à *suite* Allermatch utiliza o algoritmo baseado em “janelas” de 80 aminoácidos baseado na sequência de interesse com *cutoff* de 35% como é indicado pela FAO/WHO (Figuras 16 e 17), utilizando o banco de dados que mescla o banco de dados UniProt e o banco de dados WHO-IUIS.

Figura 16 - Consulta a *suite* Allermatch utilizando a plataforma.

Database : UniProt and WHO-IUIS

Hit No	Db	Allermatch Id	Best hit Identity	No of hits ident > 35.00	% of hits ident > 35.00	Full Identity	External link	Species Name	Detailed Information
*1	*2	*3	*4	*5	*6	*7	*8	*9	*10
1	?	al_Bla_g_4	88.75	112	100.00	84.71 / 170	P54962^S	Blattella germanica	<input type="button" value="Go"/>

Fonte: Do autor (2013, dados não publicados).

Figura 17 - Resultado referente a consulta ao *site* da Allermatch individualmente.

Database : UniProt and WHO-IUIS

Hit No	Db	Allermatch Id	Best hit Identity	No of hits ident > 35.00	% of hits ident > 35.00	Full Identity	External link	Species Name	Detailed Information
*1	*2	*3	*4	*5	*6	*7	*8	*9	*10
1	?	al_Bla_g_4	88.75	112	100.00	84.71 / 170	P54962^S	Blattella germanica	<input type="button" value="Go"/>

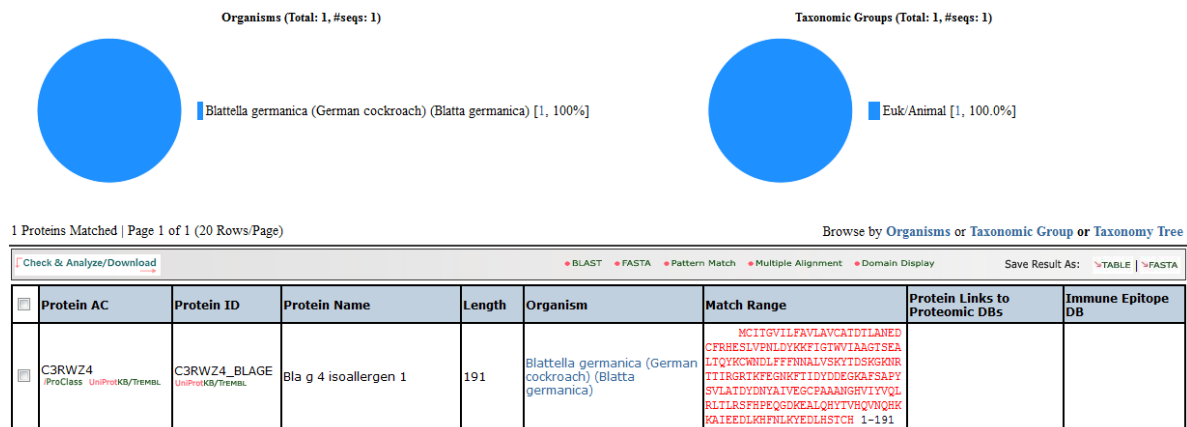
Fonte: Do autor (2013, dados não publicados).

Os resultados provenientes do Allermatch são organizados de acordo com o número do *hit*, ou seja, o melhor *hit* vem em primeiro lugar. A identidade do melhor *hit* também é especificada de acordo com o percentual de aminoácidos idênticos na “janela” de 80 aminoácidos.

O PeptideMatch realiza buscas por *matches* no banco de dados UniProtKB. *Match*, de acordo com Prosdocimi et al. (2002), é cada unidade pareada no alinhamento de duas sequências de proteínas. Chama-se *mismatch* as unidades não pareadas e as penalidades por *gaps* inseridos é considerada (PROSDOCIMI et al., 2002). Os resultados de uma análise no PeptideMatch incluem: (i) um gráfico à esquerda que mostra o número de organismos em que

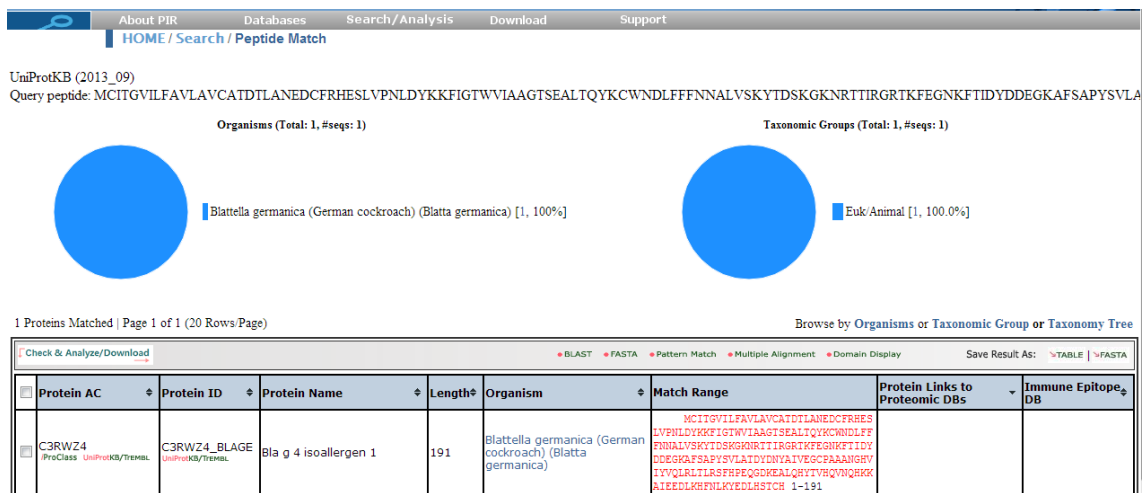
a sequência consultada foi encontrada; (ii) um gráfico à direita que refere-se aos grupos taxonômicos, destacando-se o número e a porcentagem de proteínas encontradas em cada um; e, (iii) uma tabela que resume todas as informações referentes à sequência consultada, destacando-se a extensão do *match* encontrado (Figuras 18 e 19).

Figura 18 – Resultados impressos em PA³P para a consulta ao PeptideMatch.



Fonte: Do autor (2013, dados não publicados).

Figura 19 – Resultados referentes a consulta junto ao PeptideMatch.



Fonte: Do autor (2013, dados não publicados).

Ainda há a possibilidade de realizar um blastp utilizando o banco de dados UniProtKB no UniProt. Este alinhamento é mais robusto quando comparado ao alinhamento realizado no blastp, pois no blastp somente há a opção de utilizar a seção UniProtKB/SwissProt. Quando

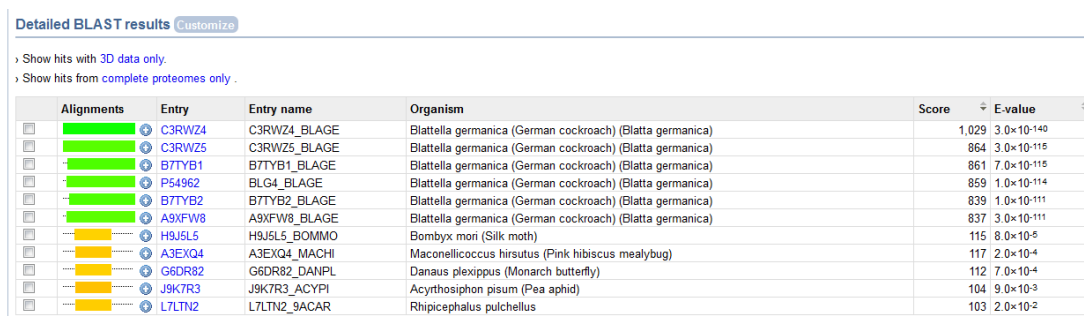
realizado no UniProt, o blastp utiliza-se das duas seções do UniProtKB (SwissProt e TrEMBL). Os resultados são mostrados na Figura 20 e na Figura 21.

Figura 20- Blastp realizado no UniProt utilizando o banco de dados UniProtKB através da plataforma.

```
Sequences producing High-scoring Segment Pairs: Score P(N) N
C3RWZ4_BLAG E C3RWZ4 Bla g 4 isoallergen 1 OS=Blattella... 1029 7.9e-102 1
C3RWZ5_BLAG E C3RWZ5 Bla g 4 isoallergen 2 OS=Blattella... 867 1.2e-84 1
B7TYB1_BLAG E B7TYB1 Bla g 4 allergen (Fragment) OS=Blattella... 861 5.0e-84 1 SP:BLG4_BLAG E P54962 Allergen Bla g 4 (Fragment) OS=Blattella... 859 8.2e-84 1
B7TYB2_BLAG E B7TYB2 Bla g 4 allergen variant 1 (Fragment) OS=Blattella... 839 1.1e-81 1
A9XFW8_BLAG E A9XFW8 Bla g 4 allergen (Fragment) OS=Blattella... 837 1.8e-81 1
H9J5L5_BOMMO H9J5L5 Uncharacterized protein OS=Bombyx... 124 6.3e-06 1
G6DR82_DANPL G6DR82 Lopap OS=Danaus plexippus GN=KGM_1... 120 0.00069 1
A3EXQ4_MACHI A3EXQ4 Putative uncharacterized protein (... 116 0.0058 1
J9K7R3_ACYPI J9K7R3 Uncharacterized protein OS=Acyrtosiphon... 109 0.020 1
```

Fonte: Do autor (2013, dados não publicados).

Figura 21 - Resultado obtido diretamente do *site* do UniProt para blastp com o banco de dados UniProtKB.



Alignment	Entry	Entry name	Organism	Score	E-value
	C3RWZ4	C3RWZ4_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	1,029	3.0×10 ⁻¹⁴⁹
	C3RWZ5	C3RWZ5_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	864	3.0×10 ⁻¹¹⁵
	B7TYB1	B7TYB1_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	861	7.0×10 ⁻¹¹⁵
	P54962	BLG4_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	859	1.0×10 ⁻¹¹⁴
	B7TYB2	B7TYB2_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	839	1.0×10 ⁻¹¹⁴
	A9XFW8	A9XFW8_BLAG E	Blattella germanica (German cockroach) (Blattella germanica)	837	3.0×10 ⁻¹¹¹
	H9J5L5	H9J5L5_BOMMO	Bombyx mori (Silk moth)	115	8.0×10 ⁻⁶
	A3EXQ4	A3EXQ4_MACHI	Maconellicoccus hirsutus (Pink hibiscus mealybug)	117	2.0×10 ⁻⁴
	G6DR82	G6DR82_DANPL	Danaus plexippus (Monarch butterfly)	112	7.0×10 ⁻⁴
	J9K7R3	J9K7R3_ACYPI	Acyrtosiphon pisum (Pea aphid)	104	9.0×10 ⁻³
	L7LTN2	L7LTN2_SACAR	Rhhipicephalus pulchellus	103	2.0×10 ⁻²

Fonte: Do autor (2013, dados não publicados).

Validação

A validação desta nova ferramenta foi imprescindível tanto para comprovar sua eficácia quanto para reiterar sua funcionalidade. Os dados anteriormente apresentados referem-se ao processo de validação da plataforma e podem ser observados na Tabela 1. Pelo processo de validação, os resultados obtidos na PA³P tem sua veracidade comprovada.

Tabela 1 – Resumo do processo de validação da Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas

ID de acesso	Grupo	PA ³ P			Análises Individuais		
		blastp <i>e value</i>	AlgPred Decisão	UniProt <i>Score</i>	blastp <i>e value</i>	AlgPred Decisão	UniProt <i>Score</i>
AAT00594.1	Alergênicas	0.0	Potencial Alergênico	1544	0.0	Potencial Alergênico	1544
P12547.2	Alergênicas	0.0	Potencial Alergênico	2043	0.0	Potencial Alergênico	2043
AAD13650.1	Alergênicas	0.0	Alergênica	1391	0.0	Alergênica	1391
ACF53836.1	Alergênicas	$1 e^{-140}$	Potencial Alergênico	1029	$1 e^{-140}$	Potencial Alergênico	1029
NP_776945.1	Alergênicas	0.0	Alergênica	7584	0.0	Alergênica	7584
3LHP_M	Não Alergênicas	$3 e^{-76}$	Não Alergênica	488	0.0	Não Alergênica	488
NP_001186758.1	Não Alergênicas	0.0	Não Alergênica	2020	0.0	Não Alergênica	2020
NP_001171802.1	Não Alergênicas	$4 e^{-136}$	Não Alergênica	1012	$4 e^{-136}$	Não Alergênica	1012
NP_001230072.1	Não Alergênicas	0.0	Não Alergênica	2186	0.0	Não Alergênica	2186
Q538Z0.1	Não Alergênicas	$2 e^{-30}$	Não Alergênica	284	0.0	Não Alergênica	284
NP_001011575.1	Não Identificadas	0.0	Não Alergênica	2843	0.0	Não Alergênica	2843
AAX33330.1	Não Identificadas	$3 e^{-29}$	Alergênica	292	$3 e^{-29}$	Alergênica	292
AAA75609.1	Não Identificadas	$2 e^{-162}$	Não Alergênica	1179	$2 e^{-162}$	Não Alergênica	1179
ADC84211.1	Não Identificadas	0.0	Não Alergênica	2012	0.0	Não Alergênica	2012
ABA19277.1	Não Identificadas	$7 e^{-14}$	Não Alergênica	193	$7 e^{-14}$	Não Alergênica	193

Fonte: Do autor (2013, dados não publicados).

Alternativa ao uso da PA³P

Uma alternativa ao uso desta metodologia de predição *in silico* é o acompanhamento das recomendações da FAO/WHO. O pesquisador, primeiramente, deve obter a sequência de aminoácidos de todas as proteínas comprovadamente ou supostamente alergênicas a partir de bancos de dados como, por exemplo, o SwissProt e o GenBank para construir um banco de dados em formato FASTA. Outro banco de dados deveria ser construído contendo sequências de 80 aminoácidos cada, derivados da sequência da proteína que está sendo analisada. Por fim, deve comparar cada uma das sequências do banco de dados (que contém as sequências de 80 aminoácidos) com todas as sequências do banco de dados composto de todas as proteínas comprovadamente ou supostamente alergênicas. Com base nos resultados dos alinhamentos, uma proteína seria considerada possivelmente alergênica quando houvesse mais de 35% de identidade em uma janela de 80 aminoácidos ou houvesse identidade de 6 aminoácidos contíguos (FAO/WHO, 2001). Contudo, apenas a identidade entre as sequências não é suficiente para atestar o potencial alergênico de determinada proteína. Visando agregar valor as predições *in silico* é que as ferramentas que foram compiladas neste trabalho foram criadas.

Considerações Finais

Pelo exposto e pelo conhecimento disponível na literatura, pode-se concluir que o uso da Plataforma de Análise da Alergenicidade e Antigenicidade de Proteínas é considerado satisfatório uma vez que permite que análises *in silico* sejam realizadas e que potenciais alergênicos ou antigênicos sejam definidos. O tempo é outro fator importante, pois o uso de PA³P reduz consideravelmente o tempo necessário para executar todas as análises. Desta forma, a referida plataforma contribuirá para embasar a decisão de utilizar ou não determinada proteína na construção de um OGM com finalidade nutricional. Trabalhos futuros incluem a persistência dos dados para aperfeiçoar o desempenho da PA³P; a inclusão de outros algoritmos de predição; a construção de uma análise estatística por meio de inteligência artificial; e, a conexão com bancos de estruturas tridimensionais de proteínas.

Referências

ALBERTS, B.; BRAY, D.; LEWIS, J.; **Biologia Molecular da Célula**. Porto Alegre: Artmed. 1396 páginas. 2010.

ALTMAN, D.G.; BLAND, J.M.; **Diagnostic Tests 2: Predictive Values**. BMJ, v. 309, p. 102. 1994.

ARNTZEN, T. C. et al; **Manual do PHP**. Disponível em http://www.php.net/manual/pt_BR/index.php acessado em 10/09/2013.

BERGMAN, N.H.; **Comparative Genomics**. Totowa: Human Press. 546 páginas. 2007.

CHAPMAN, M.D.; et al; **Nomenclature And Structural Biology Of Allergens**. Allergy Clinical Immunology, v.199, p.414-420. 2007.

FAO/WHO. **Evaluation of Allergenicity of Genetically Modified Foods**. Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. 27 páginas. 2001.

FARIA-CAMPOS, A. C.; et al.; **Efficient Secondary Database Driven Annotation Using Model Organism Sequences**. In Silico Biology, v.5, p.363-372. 2006.

FIELDING, R.T.; et al.; **Hypertext Transfer Protocol -- HTTP/1.1**. United States: RFC Editor. 1999.

FIELDING, R.T.; TAYLOR, R.N.; **Principled Desing of the Modern Web Architecture**. ACM Transactions on Internet Technology, v. 2, p.115–150. 2002.

FIERS, M.W.; et al.; **Allermatch™, a Webtool for the Prediction of Potential Allergenicity According to Current FAO/WHO Codex Alimentarius Guidelines**. BMC Bioinformatics, v.5, p.1-6.2004

Food and Agriculture Organization of the United Nations – FAO; disponível em <http://www.fao.org/> acessado em 10/09/2013.

HJÄLTÉN, J; et al; **Innate and Introduced Resistance Traits in Genetically Modified Aspen Trees and Their Effect on Leaf Beetle Feeding**. Plos One, v.8, p.1-7. 2013.

MARI, A.; et al.; **Allergen Databases: Current Status and Perspectives**. Current allergy and asthma reports, v.5, p.376-383. 2009.

MARRA, G. M.; **Aplicações PHP Compiladas Utilizando o Roadsend Studio for PHP**. Estudos, v. 36, p. 653-662. 2009.

MELO, A. A.; NASCIMENTO, M.G.F.; **PHP Profissional. Aprenda a Desenvolver Sistemas Profissionais Orientados a Objetos com Padrões de Projeto**. São Paulo: Novatec. 464 páginas. 2007.

SAHA, S.; RAGHAVA, G.P.S.; **AlgPred: Prediction of Allergenic Proteins and Mapping of IgE Epitopes**. Nucleic Acids Research, v.34, p.202-209. 2006.

SOUZA JUNIOR, M.T.; MARTINS, N.F.; **Predição do Potencial de Alergenicidade em OGMs – Estudo de Caso**. Biotecnologia, Ciência e Desenvolvimento, v.30, p.10-15. 2003.

The UniProt Consortium; **The Universal Protein Resource (UniProt)**. Nucleic Acids Research, v. 35, p. 193-197. 2007.

TICONA, W. G. C.; **Aplicação de Algoritmos Genéticos Multi-Objetivo para Alinhamento de Sequências Biológicas**. Dissertação apresentada ao ICMC-USP para obtenção do título de Mestre em Ciências na área de Ciências de Computação e Matemática Computacional. São Carlos, SP. Fev/2003.

WANG, J.; et al.; **Evaluation and Integration of Existing Methods for Computational Prediction of Allergens**. BMC Bioinformatics, v.14, p.. 2013.

World Health Organization – WHO; disponível em <http://www.who.int/en/> acessado em 10/09/2013.

WU, C. H.; et al.; **The Protein Information Resource**. Nucleic Acids Research, v. 31, p.324-247. 2003.

4 CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS

4.1 Considerações Finais

- A plataforma apresenta resultados condizentes com os resultados apresentados pelas análises nos *sites* individuais;
- O tempo necessário para realizar todas as análises é reduzido;

4.2 Perspectivas Futuras

- Persistência dos dados;
- Inclusão de outros algoritmos de predição;
- Construção de uma análise estatística por meio de inteligência artificial;
- Conexão com bancos de estruturas tridimensionais de proteínas;
- Demais necessidades identificadas.

5 REFERÊNCIAS

ABBAS, A.K.; LICHTMAN, A.H.; PILLAI, S.; **Imunologia Celular e Molecular**. Rio de Janeiro: Elsevier. 574 páginas. 2008.

ALBERTS, B.; BRAY, D.; LEWIS, J.; **Biologia Molecular da Célula**. Porto Alegre: Artmed. 1396 páginas. 2010.

ALTSCHUL, S.F.; et al.; **Basic Local Alignment Search Tool**. Journal of Molecular Biology, v.215, p.403-410. 1990.

BRASIL, Lei nº 11.105 de 24 de março de 2005.

COSTA, T. E. M. M.; et al.; **Avaliação de Risco dos Organismos Geneticamente Modificados**. Ciência e Saúde coletiva, v.16, p.327-336. 2011.

FAO/WHO. **Evaluation of Allergenicity of Genetically Modified Foods**. Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. 27 páginas. 2001.

Food and Agriculture Organization of the United Nations – FAO; disponível em <http://www.fao.org/> acessado em 10/09/2013.

HUG, K. **Genetically Modified Organisms: do the Benefits Outweight the Risks?** Medicina. 44(2):87-99. 2008.

KLETER, G. A.; KUIPER, H. J.; **Considerations for the Assessment of the Safety of Genetically Modified Animals Used for Human Food or Animal Feed**. Livestock Production Science, v. 74, p. 275-285. 2002.

LIEW, A. W.C.; YAN, H.; YANG, M.; **Chapter 4: Data Mining for Bioinformatics**. In: CHEN, Y.P.; **Bioinformatics Technologies**. Heidelberg: Springer Berlin. 396 páginas. 2005.

MARI, A.; et al.; **Allergen Databases: Current Status and Perspectives**. Current allergy and asthma reports, v.5, p.376-383. 2009.

MOHABATKAR, H. et al; **Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach**. Medicinal Chemistry, v.9, p.133-137. 2013.

PROSDOCIMI, F. et al; **Bioinformática: Manual do Usuário**. Biotecnologia Ciência e Desenvolvimento v.29, p. 12- 25. 2002.

SHARMA, P; GLAUR, S.N.; ARORA, N.; **In Silico Identification of IgE-binding Epitopes of Osmotin Protein**. Plos One, v.1, p.1-7. 2013.

SOUZA JUNIOR, M. T. S.; MARTINS, N. F.; **Predição do Potencial de Alergenicidade em OGMs – Estudo de Caso**. Biotecnologia Ciência e Desenvolvimento, v.30, p.10-15. 2003.

The UniProt Consortium; **The Universal Protein Resource (UniProt)**. Nucleic Acids Research, v. 35, p. 193-197. 2007.

TICONA, W. G. C.; **Aplicação de Algoritmos Genéticos Multi-Objetivo para Alinhamento de Sequências Biológicas**. Dissertação apresentada ao ICMC-USP para obtenção do título de Mestre em Ciências na área de Ciências de Computação e Matemática Computacional. São Carlos, SP. Fev/2003.

SHRADER-FRECHETTE, K.; **Property Rights and Genetic Engineering: developing nations at risk**. Science and Engineering Ethics, v.1, p. 137-149. 2005.

STADLER, M.B.; STADLER, B.M.; **Alergenicity Prediction by Protein Sequence**. The FASEB Journal, v.17, p.1141-1143. 2003.

VON KRIES, C.; WINTER, G. **The Structuring of GMO Release and Evaluation in EU law**. Journal Biotech, v.4, p.569-581. 2011.

ZORZET, A.; GUSTAFSSON, M.; HAMMERLING, U.; **Prediction of Food Protein Allergenicity: a Bioinformatic Learning Systems Approach**. In Silico Biology, v.2, p.525-534. 2002.

World Health Organization – WHO; disponível em <http://www.who.int/en/> acessado em 10/09/2013.

APÊNDICES

APÊNDICE A – Código Fonte Referente ao AlgPred Compilado na Plataforma

```

<?php
    /*
    Função que faz o POST dos dados no formulário do AlgPred
    Recebe a sequência a ser consultada da página inicial
    */

    function getAlgpred($seq){

        //Inicia a biblioteca cURL, ajustando o endereço alvo e demais
parâmetros
        $curl = curl_init();
        curl_setopt($curl, CURLOPT_URL,
"http://www.imtech.res.in/cgibin/algpred/algpred_main_temp.pl");
        curl_setopt($curl, CURLOPT_RETURNTRANSFER, true);
        // Método POST - Depende do site, ou POST ou GET
        curl_setopt($curl, CURLOPT_POST, true);
        // Seguir qualquer redirecionamento que houver na URL
        curl_setopt($curl, CURLOPT_FOLLOWLOCATION, true);

        /*
        Define um array seguindo o padrão:
        '<name do input>' => '<valor inserido>'
        Depende do site, tem que verificar quais dados são necessários
através de Chrome e Java script
        */
        $dados = array(
            'SEQNAME' => '',
            'SEQ' => $seq,
            'format' => 'nformat',
            'approach' => '3'
        );

        // Define quais informações serão enviadas pelo POST ($dados)
        curl_setopt($curl, CURLOPT_POSTFIELDS, $dados);
        //Executa o envio
        $resultado = curl_exec($curl);
        $string_to_find = "Prediction by";
        $string_onde_procurar = $resultado;
        $localizacao = strpos($string_onde_procurar, $string_to_find);
        $filtrado = substr($string_onde_procurar, $localizacao);
        //Fecha a conexão
        curl_close($curl);
        /*
        Chama a função wr do arquivo "writeResult.php" para criar o arquivo
.html com a resposta
        Parâmetros "nome do arquivo", "resultado"
        */
        wr("alg.html", $filtrado);
        //retorna valor alertando que acabou esta consulta
        return 1;
    }
}

```



APÊNDICE B – Código Fonte Referente ao Allermatch Compilado na Plataforma

```

<?php
    /*
    Função que faz o POST dos dados no formulário do Allermatch
    Recebe a sequência a ser consultada da página inicial
    */
    function getAllermatch($seq){
        //Inicia a biblioteca cURL, ajustando o endereço alvo e demais
parâmetros
        $cURL = curl_init();
        curl_setopt($cURL, CURLOPT_URL,
"http://www.allermatch.org/allermatch.py/search");
        curl_setopt($cURL, CURLOPT_RETURNTRANSFER, true);
        // Método POST - Depende do site, ou POST ou GET
        curl_setopt($cURL, CURLOPT_POST, true);
        // Seguir qualquer redirecionamento que houver na URL
        curl_setopt($cURL, CURLOPT_FOLLOWLOCATION, true);
        echo($cURL);
        /*
        Define um array seguindo o padrão:
        '<name do input>' => '<valor inserido>'
        Depende do site, tem que verificar quais dados são necessários
através de Crhome e Java script
        */
        $dados = array(
            'seq' => $seq,
            'against' => '',
            'method' => 'window',
            'cutOff' => '35',
            'wordlength' => '6',
            'database' => 'UniProt and WHO-IUIS'
        );

        // Define quais informações serão enviadas pelo POST ($dados)
        curl_setopt($cURL, CURLOPT_POSTFIELDS, $dados);

        //Executa o envio
        $resultado = curl_exec($cURL);
        //Fecha a conexão
        curl_close($cURL);
        /*
        Chama a função wr do arquivo "writeResult.php" para criar o arquivo
.html com a resposta
        Parâmetros "nome do arquivo", "resultado"
        */
        wr("aller.html", $resultado);
        //retorna valor alertando que acabou esta consulta
        return 3;
    }

```



APÊNDICE C – Código Fonte Referente ao Blastp Compilado na Plataforma

```

<?php
/*
Função que manda a sequência para o site do NCBI através da URL API
Recebe o RID da consulta
*/
function getNCBIblastp($seq){
    //TAG que será procurada no arquivo de resposta
    echo "inicio";
    $tag = "RID";
    $RID = "";
    /*
    Os dados são passados diretamente pela URL, não precisando fazer
    POST. Não usa a biblioteca cURL
    CMD=Put comando especificado na URL API do NCBI
    fopen - função que abre o endereço alvo e retorna seu conteúdo
    handle - variável que recebe o conteúdo (resposta)
    */
    $handle =
fopen("http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Put&QUERY=" . $seq
."&PROGRAM=blastp&DATABASE=nr&FORMAT_FILE=TEXT", "r");
    /*
    Lê linha a linha a resposta até o final
    Procura em cada linha a TAG "RID"
    Ao encontrar peça a posição e depois extrai apenas a RID limpa
    composta de 11 caracteres $posicao+6, 11
    */
    while ($data = fgets($handle, 1024))
    {
        $texto = strip_tags($data);
        $posicao = strpos ($texto, $tag);
        if($posicao == false) {
            Sleep(0);
        }else{
            //extrai apenas o pedaço da RID da linha da resposta
            $RID = substr($texto, $posicao+6, 11);
            //sai do laço while após encontrar
            break;
        }
    }

    //Retorna a RID para consultar a resposta
    return $RID;
}

```



APÊNDICE D – Código Fonte Referente à Mensagem de Erro Compilado na Plataforma

```
<?php
/*
Arquivo que recebe o erro, e imprime a mensagem
*/
$erro = "";
//Pega a variável com o erro
if (isset($_GET['var1'])) {
    $erro = $_GET['var1'];
}

printResult();

function printResult(){
    echo'
    <html>
    <head>
    <title> Erro </title>
    </head>
    <body>
    <br>
    <div>
    </div>
    <br>
    <div>';
    echo ("<h1 align=\"center\">0(s) site(s) estão indisponível(is) "
.$erro ."</h1>");
    echo'
    </div>
    <div>
    <a href="http://localhost:8080/PAP/index.php">Realizar outra Consulta
- Clique aqui!!</a>
    </div>
    </body>
    </html>';
    return;
}
?>
```

APÊNDICE E – Código Fonte Referente ao Index Compilado na Plataforma

```

<?php

/*
Página inicial
Contém o formulário de consulta e o botão de submit
*/

/*
Declaração dos arquivos com as funções necessárias as consultas
*/
require 'Blastp.php';
require 'psiBlast.php';
require 'Algpred.php';
require 'Allermatch.php';
require 'PeptideMatchURL.php';
require 'Uniprot.php';
require 'writeResult.php';
require 'testeURL.php';

//função chamada após clicar no botão submit
if (isset($_POST['seq'])) {
    $RIDblastp = "";
    $RIDpsiBlast = "";
    $JobID = "";
    $seq = $_POST['seq'];
    //verifica se a sequência foi colada, ou seja, comprimento > 0
    if(strlen($seq)>0) {
        //Remove os espaços em branco da sequência
        $quebra_linha = trim(str_replace("\r\n", "", $seq));
        $seqOK = str_replace(" ", "", $quebra_linha);

        /*
        Testa os sites com o arquivo testeURL.php
        Se algum estiver fora direciona a página de erro e envia a string do
erro
        */
        $valida = verifyURL();
        if (strlen ($valida) > 0){
            header("Location:
http://localhost:8080/PAP/erro.php?var1=".$valida);
        }
        else{
            //Executa as consultas uma a uma,
            $end =0;
            $end = $end + getAlgpred($seqOK);
            if ($end == 1) {
                $end = $end + getAllermatch($seqOK);
                set_time_limit(600);
            }
            if ($end == 4) {
                $end = $end + getPeptideMatch($seqOK);
                set_time_limit(600);
            }
        }
    }
}

```

```

        if ($end == 9) {
            $JobID = str_replace(' ', '', getUniprot($seqOK));
            set_time_limit(600);
        }
        if (strlen ($JobID) >0) {
            $RIDblastp = str_replace(' ', '',
getNCBIblastp($seqOK));
            set_time_limit(600);
        }
        if (strlen ($RIDblastp) > 0) {
            $RIDpsiBlast = str_replace(' ', '',
getNCBIpsiBlast($seqOK));
            set_time_limit(600);
        }
        if (strlen ($RIDpsiBlast) > 0) {
            //redireciona para a página das respostas.php
            //header("Location:
http://localhost:8080/PAP/respostas.php?var1=".$RIDblastp."&var2=".$RIDpsiBlast
."&var3=".$JobID."&var4=".$seqOK);
            header("Location:
http://pa3p.zz.mu/respostas.php?var1=".$RIDblastp."&var2=".$RIDpsiBlast
."&var3=".$JobID."&var4=".$seqOK);
        }
    }
}
else{
    echo ("Favor colar a sequência de consulta!");
}
}
//Cria o formulário de entrada de dados e chama a função acima
echo '
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<blockquote>
    <blockquote>
        <blockquote>
            <blockquote>
                <blockquote>
                    <blockquote>
                        <blockquote>
                            <blockquote>
                                <blockquote>
                                    <blockquote>
                                        <blockquote>
                                            <blockquote>
                                                <blockquote>
                                                    <p>
                                                    </p>
                                                </blockquote>
                                            </blockquote>
                                        </blockquote>
                                    </blockquote>
                                </blockquote>
                            </blockquote>
                        </blockquote>
                    </blockquote>
                </blockquote>
            </blockquote>
        </blockquote>
    </blockquote>
</blockquote>
'

```


APÊNDICE F – Código Fonte Referente ao PeptideMatch Compilado na Plataforma

```

<?php
    /*
    Função que faz a consulta dos dados no PeptideMatch
    Recebe a sequência a ser consultada da página inicial
    */
    function getPeptideMatch($seq){
        //Limpa a variável conteúdo
        $conteudo = "";
        /*
        Os dados são passados diretamente pela URL, não precisando fazer
        POST. Não usa a biblioteca cURL
        fopen - função que abre o endereço alvo e retorna seu conteúdo
        handle - variavel que recebe o conteudo (resposta)
        */
        $handle = fopen("http://proteininformationresource.org/cgi-
bin/peptidematch?peptide=".$seq , "r");
        /*
        Lê linha a linha a resposta até o final
        Coloca cada linha na variável $data através da função fgets
        Coloca todas as linhas na variável $conteudo
        */
        while ($data = fgets($handle, 1024)){
            $conteudo .= $data;
        }
        /*
        Chama a função wr do arquivo "writeResult.php" para criar o arquivo
        .html com a resposta
        Parâmetros "nome do arquivo", "resultado"
        */
        wr("pep.html", $conteudo);
        //retorna valor alertando que acabou esta consulta
        return 5;
    }

```



APÊNDICE G – Código Fonte Referente ao UniProt Compilado na Plataforma

```

<?php
    /*
        Recebe e envia a sequência para o site do Uniprot através de um web service
do site EBI
        Utiliza wublast rest web service que é o mesmo que consultar diretamente o
Uniprot
        Retorna o JobID da consulta
    */
    function getUniprot($seq){
        //Inicia a biblioteca cURL, ajustando o endereço alvo e demais
parâmetros
        $cURL = curl_init();
        //URL do web service do EBI
        curl_setopt($cURL, CURLOPT_URL,
"http://www.ebi.ac.uk/Tools/services/rest/wublast/run/");
        curl_setopt($cURL, CURLOPT_RETURNTRANSFER, true);
        // Método POST - é mandatório
        curl_setopt($cURL, CURLOPT_POST, true);
        // Seguir qualquer redirecionamento que houver na URL
        curl_setopt($cURL, CURLOPT_FOLLOWLOCATION, true);

        /*
        Define um array seguindo o padrão:
        '<name do input>' => '<valor inserido>'
        Depende do site, tem que verificar quais dados são necessários
através de Chrome e Java script
        Tem uma tabela no site EBI com o nome das databases que podem ser
consultados
        Exemplo: 'database' => 'uniprotkb_swissprot'
        */
        $dados = array(
            'email' => 'anna.boeck@gmail.com',
            'program' => 'blastp',
            'database' => 'uniprotkb',
            'stype' => 'protein',
            'sequence' => $seq
        );

        /*
        Define quais informações serão enviadas pelo POST ($dados)
        Recebe o JobID da consulta
        */
        curl_setopt($cURL, CURLOPT_POSTFIELDS, $dados);
        $JobID = curl_exec($cURL);
        curl_close($cURL);

        /*
        De posse do JobID temos que ver o status da consulta
        Às vezes demora, por isso temos que aguardar o status "FINISHED"

```

Faz uma nova consulta através do cURL agora para a URL de status e com o parâmetro JobID

```
*/
//Usa laço DO WHILE, para executar ao menos uma vez
do{
    $status =
"http://www.ebi.ac.uk/Tools/services/rest/wublast/status/" . $JobID;
    $cURLstatus = curl_init();
    curl_setopt($cURLstatus, CURLOPT_URL, $status);
    curl_setopt($cURLstatus, CURLOPT_RETURNTRANSFER, true);
    $status = curl_exec($cURLstatus);
    curl_close($cURLstatus);
    //aguarda 2 segundos e repete a consulta do status caso seja
diferente de "FINISHED"
    sleep(2);
}while (strcmp($status, "FINISHED") != 0);

return $JobID;
}
```



APÊNDICE H – Código Fonte Referente à Página de Respostas Compilado na Plataforma

```
<?php
```

```

    /*
    Recebe os RIDs NCBI e a sequência consultada e monta as respostas
    Cada uma fica em um IFRAME
    */

    /*
    Variáveis para montar os comando GET do NCBI
    CMD=GET especificação da URL API do NCBI
    URL vai no scr do iframe
    */
    $inicio = "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&RID=";
    $fim =
"&FORMAT_OBJECT=Alignment&ALIGNMENT_TYPE=Pairwise&FORMAT_TYPE=PLAIN_TEXT&DESCRIPTI
ONS=10&ALIGNMENTS=10&SHOW_LINKOUT=no";
    $seq="";
    $RIDblastp="";
    $RIDpsiBlast="";
    //$inicio_down - o NCBI nao pisca
    $inicio_down =
"http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?RESULTS_FILE=on&RID=";
    $fim_down =
"&FORMAT_TYPE=Text&FORMAT_OBJECT=Alignment&ALIGNMENT_VIEW=Tabular&CMD=Get";

    /*
    Variáveis para montar os comando do Uniprot
    out - saída em formato texto
    Ha outros formatos na saida - Consultar EBI web service
    URL vai no scr do iframe
    */
    $endereco = "http://www.ebi.ac.uk/Tools/services/rest/wublast/result/";

    $formatoTexto = "/out";
    $JobID="";

    //Captura os parâmetros passados da página inicial (index.php)
    if (isset($_GET['var1'])) {
        $RIDblastp = $_GET['var1'];
    }
    if (isset($_GET['var2'])) {
        $RIDpsiBlast = $_GET['var2'];
    }
    if (isset($_GET['var3'])) {
        $JobID = $_GET['var3'];
    }
    if (isset($_GET['var4'])) {
        $seq = $_GET['var4'];
    }

    //Monta os comandos do NCBI para Blastp e PSIBlast, monta a URL de pesquisa
do Blast

```

```

$cmdGetBlastp = $inicio . $RIDblastp . $fim;
$cmdGetBlastp_down = $inicio_down . $RIDblastp . $fim_down;
$cmdGetpsiBlast = $inicio . $RIDpsiBlast . $fim;
$cmdGetpsiBlast_down = $inicio_down . $RIDpsiBlast . $fim_down;

//Monta os comandos do Uniprot para texto
$UniprotTexto = $endereco . $JobID . $formatoTexto;
//Esta função pega o conteúdo de uma pag. e o joga dentro da variável data.
// Na função print resultado essa função é chamada, passando a url desejada.

$resultado = get_content ($UniprotTexto);

function get_content ($URL){
    $ch = curl_init ();
    curl_setopt ($ch, CURLOPT_RETURNTRANSFER, true);
    curl_setopt ($ch, CURLOPT_URL, $URL);
    $data = curl_exec ($ch);

    curl_close ($ch);
    return $data;
}

$string_to_find = "Sequences producing";
$string_onde_procurar = $resultado;
$localizacao = strpos($string_onde_procurar, $string_to_find);
$filtrado = substr($string_onde_procurar, $localizacao);
$filtrado = str_replace ("TR:", "<br>", $filtrado);
$arquivo_filtrado = fopen ("saidauniprot.html", "w");
fwrite ($arquivo_filtrado, "<!DOCTYPE html>");
fwrite ($arquivo_filtrado, "<html>\n");
fwrite ($arquivo_filtrado, "<head>\n");
fwrite ($arquivo_filtrado, "</head>\n");
fwrite ($arquivo_filtrado, "<body>\n");
fwrite ($arquivo_filtrado, "<table>\n");
fwrite ($arquivo_filtrado, "<tr>\n");
fwrite ($arquivo_filtrado, "<br>\n");
fwrite ($arquivo_filtrado, $filtrado, $filtrado+890);
fwrite ($arquivo_filtrado, "\n</tr>");
fwrite ($arquivo_filtrado, "\n</table>");
fwrite ($arquivo_filtrado, "\n</body>\n");
fwrite ($arquivo_filtrado, "</html>\n");

fclose ($arquivo_filtrado);

//Chama a função que imprime a página de respostas
printResult($cmdGetBlastp, $cmdGetBlastp_down, $cmdGetpsiBlast,
$cmdGetpsiBlast_down, $UniprotTexto, $arquivo_filtrado, $seq);

function printResult($cmdGetBlastp, $cmdGetBlastp_down, $cmdGetpsiBlast,
$cmdGetpsiBlast_down, $UniprotTexto, $arquivo_filtrado, $seq){
    echo'

```

```

        <html>
        <head>
        <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<blockquote>
  <blockquote>
    <blockquote>
      <blockquote>
        <blockquote>
          <blockquote>
            <blockquote>
              <blockquote>
                <blockquote>
                  <blockquote>
                    <blockquote>
                      <blockquote>
                        <blockquote>
                          <blockquote>
                            <blockquote>
                              <blockquote>
                                <p>
                                </p>
                              </blockquote>
                            </blockquote>
                          </blockquote>
                        </blockquote>
                      </blockquote>
                    </blockquote>
                  </blockquote>
                </blockquote>
              </blockquote>
            </blockquote>
          </blockquote>
        </blockquote>
      </blockquote>
    </blockquote>
  </blockquote>
</blockquote>
  <title> Respostas </title>
</head>
<body>
<h1 align="center">Respostas</h1>
<h2 align="left">Sequência Consultada</h2>';
echo("<h5 align=\"left\">\" . $seq .\"</h5>");
echo'
<div>
<a href="http://pa3p.zz.mu/">Para realizar uma nova consulta, clique
aqui!</a>
</div>
<br>
<div>
<h2 align="left">Blastp</h2>';

echo("<iframe id =\"ncbiBlastp\" src=\"\" . $cmdGetBlastp . \"\"
scrolling=\"auto\" frameborder=\"0\" vspace=\"0\" hspace=\"0\" width=\"100%\"
height=\"4500px\">");

$file = fopen ($cmdGetBlastp_down , "r");
echo $cmdGetBlastp_down;
echo $file;
fclose ($file);

//inicio do código que faz não piscar-----
$curl = curl_init();
curl_setopt($curl, CURLOPT_URL, $cmdGetBlastp_down);

```

```

    curl_setopt($curl, CURLOPT_POST, true);
    curl_setopt($curl, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($curl, CURLOPT_HEADER, false);
    curl_setopt($curl, CURLOPT_TIMEOUT, 60);
    curl_setopt($curl, CURLOPT_SSL_VERIFYPEER, false);
    curl_exec($curl);
    curl_close($curl);

//-----
echo("</iframe>");
echo'
</div>
<br>
<div>

<h2 align="left">psi-blast</h2>';

echo("<iframe id = \"ncbipsiBlast\" src=\"\" .\$cmdGetpsiBlast . \"\"
scrolling=\"auto\" frameborder=\"0\" vspace=\"0\" hspace=\"0\" width=\"100%\"
height=\"4250px\">");
echo("</iframe>");
echo'
<br>
<div>
<h2 align="left">AlgPred</h2>
<iframe id = "algpred" src="alg.html" scrolling="auto"
frameborder="0" vspace="0" hspace="0" width="100%" height="200px" >
</iframe>
</div>
<br>
<div>
<h2 align="left">Allermatch</h2>
<iframe id = "allermatch" src="aller.html" scrolling="auto"
frameborder="0" vspace="0" hspace="0" width="100%" height="600px" >
</iframe>
</div>
<br>
<div>
<h2 align="left">Peptidematch</h2>
<iframe id = "pep" src="pep.html" scrolling="auto"
frameborder="0" vspace="0" hspace="0" width="100%" height="600px" >
</iframe>
</div>
<div>
<h2 align="left">UniProt</h2>';

$contentFile = "saidauniprot.html";
readfile( $contentFile );
echo'
</div>
<br>
</body>
</html>';
return;
}

```

