Gislaine da Silva Pimentel Pereira

Abordagem Computacional para a Predição do Secretoma Humano

Tese apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Doutor em Ciências.

Área de Concentração: Genética
Orientador: Prof. Dr. Wilson Araújo da Silva Junior

Ribeirão Preto – SP
2009

Catalogação na Publicação
Serviço de Documentação
Faculdade de Medicina de Ribeirão Preto

Pereira, Gislaine da Silva Pimentel

Abordagem Computacional para a Predição do Secretoma Humano/ Gislaine da Silva Pimentel Pereira; orientador: Wilson Araújo da Silva Junior. - Ribeirão Preto - SP, 2009.
97f.: il.

Tese (Doutorado – Programa de Pós-Graduação em Genética. Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto.

1. Peptídeo sinal - Proteínas secretoras. 2. N-glicosilação - Proteína secretada.

Dedicatória

Dedico este trabalho a meu filho Paulo Roberto, presente de Deus para minha vida. Dedico também a meu marido Paulo por seu incentivo e admiração como esposa e mulher.

A meu pai Sebastião que sempre acreditou em mim e foi um eterno amante do conhecimento, da ciência e da informação e a minha mãe Izabel por seu amor e caráter ensinados desde o berço.

A meu irmão Alexander que acreditou em minha capacidade e nesse momento quero demonstrar minha total gratidão em especial a meu irmão Maikel que sempre me fez sorrir nos momentos mais difíceis de minha vida.

Amo vocês...

Agradecimentos

Em especial expresso minha total gratidão a Deus que sempre me amou, me fortaleceu e nunca me deixou desistir mesmo quando me senti sozinha e sem destino.

Quero agradecer a meu marido Paulo Roberto e com muito amor a meu filho Paulo, meu amigo, companheiro e amante da ciência, pois suportou durante esse período minha ausência, minhas angústias, ansiedades, lágrimas e nervosismo.

Agradeço ao Departamento de Genética pela confiança para que eu pudesse cursar o Doutorado em Ciências e a Fundação Hemocentro de Ribeirão Preto pela oportunidade de estudo e utilização de suas instalações e também ao CTC/CEPID/FAPESP, por me proporcionar condições físicas e financeiras.

Agradeço a meu orientador Prof. Dr. Wilson Araújo da Silva Junior por sua grandiosa paciência e incentivo na realização desta pesquisa e na elaboração deste trabalho.

Em especial agradeço ao Prof. Dr. José César Rosa, por sua confiança, amizade, incentivo e conhecimentos compartilhados, bem como sua colaboração e auxílio.

Agradeço a minha amiga Prof^a. Dr^a. Silvana Giuliatti, por seu incentivo durante todos anos, pois foram valiosos para prosseguir na caminhada.

Agradeço aos colegas Dr^a. Greice Andreotti de Molfetta, ao Mestrando Leonardo Barcelos de Paula e a Dr^a. Helen Julie Laure por todo trabalho feito na parte experimental, e ao Departamento de Biologia Celular Molecular e Bioagentes Patogênicos.

Agradeço a equipe BiT - *Bioinformatic Team* e ao LGMB - Laboratório de Genética Molecular e Bioinformática, em especial aos colegas Adriana, Alynne, Antônio, Brown, Camila, Carla, Cristiane, Dalvinha, Daniel, Francis, Greice, Israel, Meire, Thiago e também a toda equipe de funcionários do Hemocentro.

Agradeço a meus pais Izabel e Sebastião, e a meus irmãos Alex e Maikel por sempre acreditarem na importância do estudo.

"Ainda antes que houvesse dia, Eu sou; e ninguém há que possa fazer escapar das Minhas Mãos; Operando Eu, quem impedirá? "

Isaías

Resumo

As proteínas sintetizadas podem permanecer no citosol ou ter outros destinos diversos, como mitocôndrias, lisossomos, núcleo, tornar-se constituintes da membrana plasmática ou serem secretadas. Este direcionamento das proteínas, existe tanto em eucariotos quanto em procariotos e pode ser feito por meio de sequências sinalizadoras, pela própria estrutura da proteína ou por modificações pós-traducionais. As proteínas que possuem como destino final lisossomos, membrana celular ou meio extracelular, normalmente possuem em sua região amino-terminal, um peptídeo sinal: N-região, H-região, e um segmento de quebra de hélice, C-região. Além destas características, proteínas secretadas passam por um processo de formação chamado glicosilação: nitrogênio-glicosilação, que ocorre no nitrogênio e oxigênio-glicosilação, que ocorre no hidróxi-oxigênio de serina, treonina e cisteína. Um melhor entendimento destas regiões fornecem informações para desenvolvimento de sistemas computacionais para busca de proteínas com características de secretadas da célula. Neste trabalho associamos 38432 proteínas (banco de dados RefSeq *H. sapiens september 2008* NCBI) procurando peptídeo sinal, N-região, H-região, C-região, com os *softwares* SignalP 3.0 (baseado em redes neurais e modelos escondidos de Markov) e Phobius (baseados em modelos escondidos de Markov). Para encontrar a característica de glicosilação, importante em proteínas secretadas, um programa em PERL, foi implementado para localizar os motivos de N-glicosilação Asn-Xaa-Ser, Asn-Xaa-Thr ou Asn-Xaa-Cys. Com base nestes critérios de seleção obtivemos 2944 proteínas (sem isoformas) com características de secretadas pela via secretora que foram utilizadas para validação *in silico*, que consistiu de uma busca na literatura, onde encontramos 695 (25%) proteínas secretadas já descritas. Aos resultados, foi adicionado dados de experimentos de SAGE (*Serial Analysis of Gene Expression*) *tag* do tecido Osteoblasto em tempos de expressão nos tecidos: Mesenquimal T0hs, Mesenquimal T2hs, Mesenquimal T12hs e Osteoblasto T21hs e encontramos 1128 proteínas com peptídeo sinal e *sequon*, e destas 49 proteínas faziam parte das 695 validadas *in silico*. Para validação experimental desta análise de predição *in silico* das 2944, foram feitos estudos experimentais SDS-PAGE e espectrômetria de massa MALDI TOF-TOF utilizando a linhagem celular ***HCC*1954**, onde obtivemos o resultado experimental de 116 proteínas e dentre elas algumas já descritas na literatura como secretadas. Encontramos das 116 proteínas desta linhagem, 11 proteínas buscando nas 2944, que apresentam peptídeo sinal e *sequon* e destas, 9 fazem parte da validação *in silico* com os dados da literatura. As 105 proteínas, que não foram identificadas nesta abordagem, verificamos que não possuem peptídeo sinal e *sequon*. A estratégia de combinação computacional utilizando *softwares* para predição de peptídeo sinal, tornou possível a construção de um catálogo de proteínas secretadas (*H. sapiens*), com o objetivo de fornecer alvos ou fontes terapêuticas em processos apoptóticos associados a anomalias, na perspectiva farmacêutica e busca a dados experimentais.

Abstract

Computational Approach for Prediction of Human Secretome

The proteins can be synthesized in the cytosol or to other destinations different, as mitochondria, lysosomes, nucleus, become constituents of plasma membrane or be secreted. This targeting of proteins, exists both in eukaryotes and in procariotos and can be done through signaling sequences, by the structure of the protein or changes post translations. The proteins that have as final destination lysosomes, cell membrane or extracellular environment, normally they have in their amino-terminal region, a signal peptide: N-region, H-region, and a segment of the brake helix, C-region. Besides these features, secreted proteins go through a process called glycosylation: N-linked glycosylation, which occurs in the nitrogen and oxygen glycosylation, in hydroxyl oxygen of serine, threonine and cysteine. A better understanding of these regions provide information for development of computer systems to search for proteins with characteristics of secreted. In this work proteins associate 38432 proteins (RefSeq (sep 2008 in *H. sapiens* NCBI) looking for signal peptide, N-region, region-H, C-region, using the softwares SignalP 3.0 (based on a combination of several artificial neural networks and hidden Markov models) and Phobius (based on hidden Markov models). To find the glycosylation characteristics, in the secreted proteins, a PERL program, was implemented to find the N-glycosylation motifes: Asn-Xaa-Ser, Asn-Xaa-Thr or Asn-Xaa-CYS. Based on these standard of selection, we find 2944 proteins with characteristics of proteins secreted by secretory pathway, that were used for validation in silico, which consisted of a search in the literature for secretory proteins, where we find 695 (25%) secreted proteins already described from 2944. The results, was added to SAGE experiments (Serial Analysis of Gene Expression) in the tissue Osteoblast in times of expression differentially observed: T0hs mesenchymal, mesenchymal T2hs, T12hs mesenchymal and osteoblast T21hs and found 1128 proteins with signal peptide and sequon, and 49 proteins was into of 695 validated in silico. For experimental validation of this computational analysis, to predict in silico, from 2944, experimental studies were made using SDS-PAGE mass spectrometry and MALDI TOF-TOF, using the cell line ***HCC*1954**, where obtained a result of 116 proteins and some are described as secreted in the literature in the silico data of the 2944 proteins with characteristics of secreted proteins. We found of result obtained from 116 proteins this cell line, 11 proteins in the experimental data from 2944, showing signal peptide and sequon motifes, and 9 are in silico validation. The 105 proteins that were not identified in this approach, see that no have signal peptide and sequon. The strategy of combining bioinformatics using softwares for prediction of signal peptide, made possible the construction of a catalog of proteins secreted in H. sapiens (Human Secretome) with the goal of providing targets or sources therapies in apoptotic processes associated with abnormalities in the pharmaceutical perspective and seeks to experimental data.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas e siglas

sequon	Motivos de glicosilação, sequência de três aminoácidos consecutivos.
7SL RNA	Molécula componente de uma partícula de reconhecimento de sinal.
Ala	Alanina
Asn	Asparagina
ATP	Adenosina Trifosfato
ATP/ADP	Adenosina Trifosfato/Adenosina Difosfato
bias	Usado em estatística, viés para designar qualquer comentário ou análise que seja tendenciosa.
C	Cisteína
C-região	Sítio de Clivagem
CaOV3, OVCAR3 e ES2	Linhagens celular.
CDH1, TIMP1 e TIMP2	Marcadores de proteínas.
CM	Proteínas de Membrana Citoplasmática
CPU	Unidade Central de Processamento
Cyr61	*Cysteine-rich*
Cys	Cisteína
DCs	Células Dendríticas
EMT	Transição Epitélio Mesenquimal
ERRO	Aprendizado Supervisionado por Correção de Erros
EX	Proteínas Extracelular
GB	*Gigabyte*
GLC	Oligossacarídeo comum em eucariotos.
GlcNAc	Oligossacarídeo comum em eucariotos.
Gly	Glicina
GO	*Gene Ontology*
GTP	Guanosina Trifosfato
H-região	Região Hidrofóbica
HD	*Hard Disk*
HepG2	Linhagem de células hepáticas infectadas.
hESCs	Células Embrionárias Humanas
hMADS	Tecido Adiposo Humano Multipotente Derivado de Células Tronco
HMM	Modelo oculto de Markov
HMMs	Modelos Ocultos de Markov

HNEL	motivos C-termial Lys-Asp-Glu-Leu
HTML	*HyperText Markup Language*
IGF	Sinalizador de linhagens.
Ile	Isoleucina
KDEL	motivos C-termial Lys-Asp-Glu-Leu
LDA	Análise Linear Discriminante
Len	Comprimento da sequência de alinhamento.
Leu	Leucina
MALDI	*Matrix-assisted laser desorption/ionization*
Man	Manose. Sacarídeo - Oligossacarídeo comum em eucariotos.
Met	Metionina
MLP	Perceptron Multi-Camadas
mRNA	RNA mensageiro
ms	*Massa sobre carga*
MSCs	Células tronco mesenquimal.
MTH	*Dictyostelium discoideum AX4: 7,8-dihydro-8-oxoguanine triphosphatase*
MySQL	*Structured Query Language*
N	Asparagina
N	Rede Neural
N-lig-Glycan	N-acetylgucosamine
N-região	Região N-terminal da pré-proteína e H-região
OM	Membrana Externa
OST	Complexo Oligossacaril Transferase
OxPAPC	Oxidação lipídica e produtos derivados de *1*-palmitoyl-2-arachidoyl-sn-glicerol-3-phosphorylcholin
P	Prolina
PB	Proteínas Citoplasmáticas
PERL	*Practical Extraction And Report Language*
PERL-CGI	*Practical Extraction And Report Language - Common Gateway Interface*
PERL-GD	Módulo PERL
Phe	Uracila
PHP	*Hypertext Preprocessor*
PI16	Peptidase Inhibitor 16
Pro	Prolina
RAM	Memória de acesso aleatório
RE	Retículo Endoplasmático
RER	Retículo Endoplasmático Rugoso
RNA	Rede Neural Artificial
ROR2	*receptor tyrosine kinase-like orphan receptor 2 precursor [Homo sapiens]*

S	Serina
SAGE	*Serial Analysis of Gene Expression*
Ser	Serina
SGBD	Sistema de Gerenciamento de Banco de Dados
Shell	*POSIX-compliant shell script*
SP	Peptídeo Sinal
SPase I	peptidase sinal I
SPS	*Shotgun Peptide Sequencing*
SRP	Partícula de Reconhecimento de Sinal
T	Treonina
Thr	Treonina
TM	Topologia Transmembrana
TMHMM	*Prediction of Transmembrane Helices in Proteins*
TOF	*time-of-flight*
Trp	Fenilalanina
TXN4	*Thioredoxin domain containing protein 4*
Val	Valina
wt-p53	*regulated tumor cell secretome*
WWW	*World Wide Web*
X	Qualquer aminoácido
Xaa	Qualquer aminoácido

1 Introdução

A complexidade dos processos biológicos sugere que estudos em larga escala, realizados por métodos computacionais, devem centrar-se em categorias específicas da proteína, selecionados com base em critérios biológicos. A estratégia é dividir a totalidade do proteoma de um tipo específico de célula em sub-populações de proteínas que partilham função análoga e características biológicas. Nos organismos multicelulares, proteínas secretoras desempenham papel central na comunicação intercelular. Vários tecidos, além dos órgãos endócrinos, mostram que a atividade secretora é fundamental para a fisiologia de um tecido. Por exemplo, toda a área do metabolismo energético vem sofrendo mudanças pelo reconhecimento de que o tecido adiposo não é simplesmente um local para o armazenamento de gordura, mas sim um órgão endócrino que secreta um grande número de hormônios que tem papel importante na regulação do consumo alimentar, peso corporal, glicose e homeostase de lipídios (HAVEL, 2004).

Atualmente, não há uma lista concreta do número total de proteínas secretoras presentes no proteoma humano. Esta informação é importante para ajudar na caracterização das redes gênicas envolvidas em cada processo biológico, tais como proliferação, diferenciação, migração, etc. Com o sequenciamento completo do genoma humano, várias abordagens computacionais foram desenvolvidas para a predição de componentes funcionais do genoma humano, incluindo genes, regiões reguladoras, regiões repetitivas e etc.

Em um estudo computacional de proteínas secretadas pela via clássica, Klee e Sosa (2007), reviram e avaliaram programas que podem prever proteínas secretoras. Os programas usavam uma variedade de algoritmos e arquiteturas preditivas para inferir sobre a localização da proteína. Na avaliação da capacidade preditiva, eles avaliaram programas antigos e atuais usando sequências anotadas experimentalmente. Os resultados variavam nas exatidões de predição, ilustrando que os programas mais antigos mantiveram as exatidões nos resultados.

A identificação de proteínas secretoras é importante para o desenvolvimento de procedimentos terapêuticos e de diagnósticos. A falta da anotação experimental de proteínas

secretoras, aumenta a importância dos algoritmos de predição. Segundo Klee e Sosa (2007), tais programas fornecem recursos valiosos na localização e investigação da proteína, mas são igualmente valiosos aqueles cujos algoritmos podem ser usados juntamente com mineração da base de dados, assegurando que um perfil mais detalhado da localização da proteína seja obtido. Proteínas que são secretadas da célula para o meio extracelular representam a principal classe de moléculas envolvidas na comunicação intercelular de organismos multicelulares e humanos. Elas tem importância terapêutica suplementar no combate as doenças (GREENBAUM *et al.*, 2001). De acordo com Caneparo *et al.* (2007) e Bonin-Debs *et al.* (2004), o conhecimento do conjunto completo de proteínas secretoras por um determinado tecido (ou seja, seu Secretoma) pode ampliar a rede de relacionamentos com outros tecidos e fornecer elementos importantes para a identificação de alvos terapêuticos.

1.1 Predições Computacionais

Algumas abordagens computacionais incluem uma ampla variedade de métodos utilizados atualmente em análise de proteomas (GREENBAUM *et al.*, 2001). No trabalho intitulado "Sequences and Topology", Gerstein e Honig (2001), mencionam várias aproximações computacionais importantes para a definição de 'omes:

i) Construção de algoritmos para predição de genes, da estrutura proteica. Das interações ou localização gênica, baseada em padrões biológicos de nucleotídeos e aminoácidos, como por exemplo: a definição de proteoma ou *ORFome* (a soma total de *open reading frames* no genoma) usando o algoritmo *Gene Finding* (CLAVERIE, 1997; GUIGÓ *et al.*, 2000; HARRISON *et al.*, 2002; YEH; LIM; BURGE, 2001), definição do *Foldome* (população de produtos gênicos classificados através de sua estrutura terciária) do proteoma (SIMONS; STRAUSS; BAKER, 2001), definição do *Foldome* usando sítios de ligação conhecidos (TEICHMANN; MURZIN; CHOTHIA, 2001) e a determinação do secretoma pela identificação de sequência sinal no proteoma (TJALSMA *et al.*, 2000).

ii) Anotação por homologia inferindo estrutura ou função baseando-se na sequência e informação estrutural de proteínas homólogas (GERSTEIN, 1997; BRENNER, 1999; HEGYI; GERSTEIN, 1999; WILSON; KREYCHMAN; GERSTEIN, 2000; THORNTON, 2001; HEGYI; GERSTEIN, 1999).

iii) Método por associação (*guilt-by-association*) baseado em clusterização, onde funções ou interações são inferidas de dado funcional genômico de *clusters*, como informação de expressão. Por exemplo, funções similares podem algumas vezes ser inferidas por interação com outras proteínas ou perfis de expressão similares (GERSTEIN, 1998; MARCOTTE *et al.*, 1999; ITO *et al.*, 2001).

É importante considerar que a composição do genoma e do proteoma é praticamente estática. Já a do transcriptoma e translatoma é dinâmica em resposta a fatores endógenos e exógenos. No entanto 'omes adicionais descrevem a presença de moléculas que não são codificadas pelo genoma, mas que ainda assim são essenciais, como por exemplo, o *Metabolome* (TWEEDDALE; NOTLEY-MCROBB; FERENCI, 1998). Na tabela 1 há uma lista de vários 'omes já descritos na literatura (GREENBAUM *et al.*, 2001).

Tabela 1: Lista de 'omes descritos na literatura.

Termo	**Descrição**
Genome	A completa informação genética de regiões codificantes e não codificantes de um organismo.
Proteome	As regiões (proteína) codificantes do genoma.
Transcriptome	A população de mRNAs transcritos na célula, ponderado por seus níveis de expressão.
Physiome	Descrição quantitativa da dinâmica fisiológica ou funções de todo o organismo.
Metabolome	O complemento quantitativo de todas as pequenas moléculas presentes na célula em um específico estado fisiológico.
Phenome	Identificação qualitativa de forma e função derivada de genes, mas faltando um quantitativo, definição integrada.
Morphome	A descrição quantitativa da estrutura anatômica, bioquímica e de composição química de um organismo, incluindo o seu genoma, proteoma, células, tecidos e estruturas.
Interactome	Lista de interações entre todas macromoléculas na célula.
Glycome	A população de moléculas de carboidratos na célula.
Secretome	A população de produtos gênicos que são secretados da célula.
Ribonome	A população de regiões codificantes de RNA do genoma.
Orfeome	A soma total de *open reading frames* no genoma sem levar em conta o código; um subconjunto é o proteoma.
Regulome	*Genome-wide* rede regulatória da célula
Cellome	O conjunto de moléculas e suas interações com célula.
Operome	A caracterização de proteínas com desconhecida função biológica.
Transportome	A população de produtos gênicos que são transportados; inclui o secretoma.
Pseudome	O complemento de pseudo-genes no proteoma.
Functome	A população de produtos gênicos classificados por suas funções.
Translatome	A população de proteínas na célula, ponderada pelos níveis de expressão.
Foldome	A população de produtos gênicos classificados por suas estruturas terciárias.
Surfaceome	Identificação do subconjunto de proteínas de membrana externa na superfície celular.
Unknome	Genes de funções desconhecidas.

A elucidação de cada um destes 'omes contribui para o objetivo final da genômica funcional e também a definição do "Functome", que descreve todas as funções que são

atribuídas para cada gene no genoma. Um dos importantes grupos de 'omes que estão em menor quantidade, é o "Secretome", um subconjunto do proteoma que é definido pela ação de proteínas que são secretadas (GREENBAUM *et al.*, 2001).

Antelmann *et al.* (2001), descrevem as primeiras abordagens computacionais para atribuir função as proteínas secretoras. Anteriormente, o grupo utilizou uma metodologia para prever as proteínas secretoras, em *Bacillus subtilis* com base na identificação do peptídeo sinal em proteínas (TJALSMA *et al.*, 2000). Eles também validaram experimentalmente a função secretora por eletroforese em gel 2D e espectrometria de massa. A taxa de validação foi de aproximadamente 50%. Grande parte das proteínas não validadas foi dado ao fato de que as mesmas não apresentaram peptídeo sinal.

Viratyosin *et al.* (2008), também utilizaram em suas pesquisas um método baseado em predição, para identificar proteína de membrana exterior (OM - *outer membrane*) e proteínas extracelular (EX) em *Leptospira interrogans*. A estratégia teve as seguintes etapas: (i) identificar proteínas homólogas às proteínas conhecidas na localização subcelular com dados derivados do consenso de predições computacionais; (ii) incorporar homologia baseado em pesquisa e informação estrutural aumentando a eficiência da anotação e identificação funcional inferindo estrutura e localização; e (iii) desenvolver um classificador específico para proteínas citoplasmáticas (PB) e proteínas de membrana citoplasmática (CM), utilizando análise linear discriminante (LDA). Nesse estudo, os autores identificaram 114 proteínas EX e 63 proteínas OM, das quais 41% tem sequências conservadas ou hipotéticas com estruturas semelhantes as de EX e MO.

Tais trabalhos, proporcionam suporte para predições *in silico* de anotação de genomas, definição de *ORFomes*, definição de *Foldomes*, análise de proteomas e determinação de secretomas. Eles também fornecem valorosa informação para estudos de alvos terapêuticos, descoberta de biomarcadores e desenvolvimento de drogas e vacina. No presente estudo nós usamos uma estratégia para a caracterização de proteínas secretoras com o objetivo central de gerar a primeira lista do secretoma humano. O foco foi identificar proteínas secretadas pela via clássica com base nas regiões específicas N-terminal com peptídeo sinal, N-região, H-região, C-região e motivos de glicosilação caracterizados em proteínas secretadas da célula que cruzam o lúmen do Retículo Endoplasmático Rugoso (RER). Nesta análise usamos a base de dados RefSeq, com o registro de 38432 sequências - *Release September human* 2008 (`ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz`).

As seções seguintes irão apresentar em detalhes as características e os aspectos gerais das proteínas secretadas e em seguida, nas subseções uma síntese das proteínas na via secretora e as características do peptídeo sinal.

1.2 Aspectos Gerais das Proteínas Secretadas

As proteínas secretadas representam um subconjunto do proteoma ativamente envolvido em sinalização intercelular, desenvolvimento embrionário e organogênese (CANEPARO *et al.*, 2007; PICKART *et al.*, 2006). O papel desta classe de proteínas, no estágio inicial da doença, tem sido estudado intensivamente visando alvos terapêuticos e biomarcadores de diagnóstico e prognóstico. Estas são importantes, pois estão presentes em fluidos do corpo, de todos os animais. Um grande número de proteínas secretada foi recentemente caracterizado em *zebrafish* (KLEE, 2008). O termo "Secretoma" foi descrito primeiramente para descrever as proteínas secretoras no *Bacillus subtilis* (TJALSMA *et al.*, 2000). A adaptação subsequente deste termo aos eucariotos serviu para descrever subconjuntos do proteoma, incluindo todas as proteínas processadas pela via secretora clássica e não clássica (SCHATZ; DOBBERSTEIN, 1996; KLEE, 2008).

A localização final das proteínas ocorre por um processo em cascata movido por uma série de eventos que são mediados por pequenos peptídeos, ou motivos, que habilitam a ligação de sítios proteicos específicos de retenção e transporte (MCGEOCH, 1985; DOUDNA; BATEY, 2004). Estes eventos variam dependendo do reino, mas em eucariotos ele ocorre durante a tradução quando as proteínas entram na via secretora sendo co-traduzidas e transportadas dentro do retículo endoplasmático rugoso. Este processo é mediado por um peptídeo sinal N-terminal na sequência da proteína nascente. A sequência do peptídeo possui três regiões com propriedades conservadas: uma região N-terminal de resíduos básicos, uma região central de resíduos hidrofóbicos e uma região C-terminal de resíduos polares. Seguindo a travessia dentro do RE, o peptídeo sinal é clivado da proteína madura. Proteínas que são processadas desta maneira são proteínas classicamente secretadas. Estas proteínas usam sinais secundários, como motivos C-termial "KDEL" e "HNEL" para retenção e direção final RE, Golgi, lisossomo e membrana celular. Considerando que a maior parte das proteínas extracelulares são processadas por meio dessa via, algumas são exportadas da célula por mecanismos independentes e são chamadas não-classicamente secretadas e não são frequentemente bem caracterizadas por programas de predição (KLEE; SOSA, 2007). Neste trabalho o termo secretoma se refere as proteínas processadas classicamente pela via secretora. Proteínas classicamente processadas pas-

sam por um processo composto dos seguintes passos: 1) translocação - cruza o retículo endoplasmático rugoso; 2) N-glicosilação e empacotamento no lúmen RE; 3) saída do RE; 4) modificações no complexo de Golgi; 5) e finalmente a liberação dos grânulos secretores (vesículas secretoras) para o espaço extracelular (SAKAGUCHI, 1997) (Figura 1).

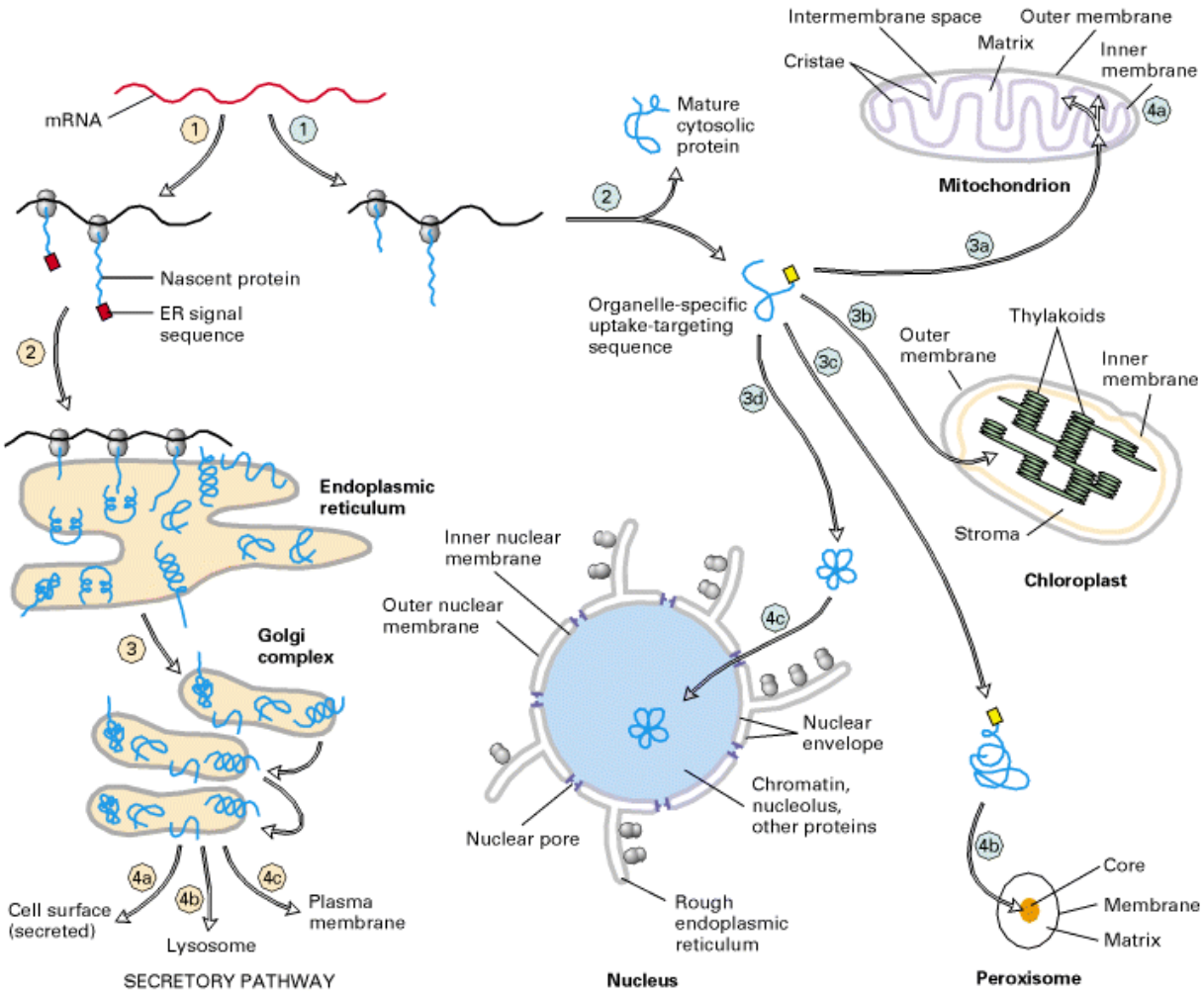


Figura 1: Visão geral da classificação de proteínas nuclear codificadas em eucarioto (LODISH *et al.*, 1999).

1.2.1 Síntese da Proteína na Via Secretora Clássica

A síntese da maioria das proteínas secretadas, inicia nos ribossomos. A presença da sequência de peptídeo sinal (16 a 30 resíduos) direciona o ribossomo para a membrana do retículo endoplasmático rugoso, onde finaliza a síntese proteica para em seguida iniciar seu transporte através do RE. O peptídeo sinal está localizado na região N-terminal da proteína secretada e contém um ou mais aminoácidos com carga positiva seguido por um contínuo trecho de 6 a 12 resíduos hidrofóbicos (Tabela 2) (LODISH *et al.*, 1999; BURKA, 1974; RAPOPORT; WIEDMANN, 1985).

Tabela 2: Sequência de peptídeo sinal do RE em três proteínas eucarióticas. Resíduos hidrofóbicos em negrito; ↓ indica o sítio de clivagem pelo peptídeo sinal.

Proteína - **Sequência de Aminoácido**
Preproalbumin MetLysTrpValThr**PheLeuLeuLeuLeuPheIleSerGlySerAlaPheSer** ↓ Arg...
Pre-IgG light chain MetAspMetArgAlaProAlaGln**IlePheGlyPheLeuLeuLeuLeuPhe**ProGlyThrArgCys ↓ Asp...
Prelysozyme MetArgSer**LeuLeuIleLeuValLeuCysPheLeu**ProLeuAlaAlaLeuGly ↓ Lys...

A sequência sinal N-terminal emerge do ribossomo quando o polipeptídeo é de aproximadamente 70 aminoácidos, pois aproximadamente 30 aminoácidos permanecem no ribossomo. O aminoácido iniciador metionina (**Met**), juntamente com os resíduos hidrofóbicos constituem a região citosólica. A região N, entre o inicializador **Met** e a região H, é de aproximadamente 1 a 5 aminoácidos com carga positiva. A região H é rica em leucinas, alaninas e valinas. Entre a região H e a região C que consiste de 3 a 7 aminoácidos polar não carregados está o sítio de clivagem para a enzima peptidase sinal. Em relação ao sítio de clivagem encontra-se preferencialmente alaninas e outros resíduos com curtas cadeias laterais nas posições -3 e -1 e prolinas nas posições -6 e -4. Para a clivagem ser eficiente a região deve ser pequena e neutra (KAISER *et al.*, 1987; von Heijne, 1985) (Figura 2).

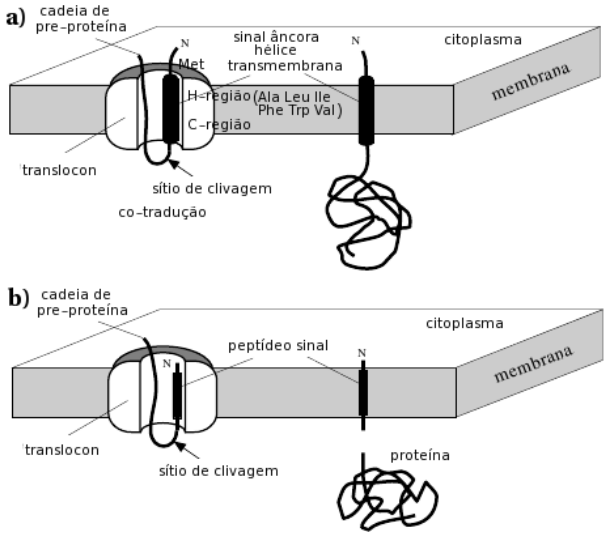


Figura 2: Representação esquemática do processo de clivagem e liberação da proteína secretada.

O processo de clivagem inicia-se com uma ligação de uma partícula de reconhecimento

de sinal (SRP) no peptídeo sinal da proteína nascente (Figura 3 [1]), a SRP é considerada uma ribo-nucleoproteína composta de 6 subunidades e 7SL RNA (WALTER; JOHNSON, 1994) (Figura 3 [1]). Em seguida, o complexo SRP se ligam a subunidade α do receptor SRP na membrana RE (Figura 3 [2]). Nessa ocasião, o SRP é liberado para ligar-se a outra proteína nascente. O peptídeo sinal se liga ao *translocon* e imediatamente, junto com a proteína nascente é passado para o lúmen do RE (Figura 3 [3], [4]). Nesse momento, ocorre a clivagem do peptídeo sinal que é imediatamente degradado (Figura 3 [5]). Nessa etapa, a síntese da proteína nascente é finalizada no lúmen do RE e em leveduras, essa fase é acompanhada pela ligação de chaperonas (Hsc70) e modificações pós-tradução fundamentais para o amadurecimento funcional da proteína madura (Figura 3 [6]).

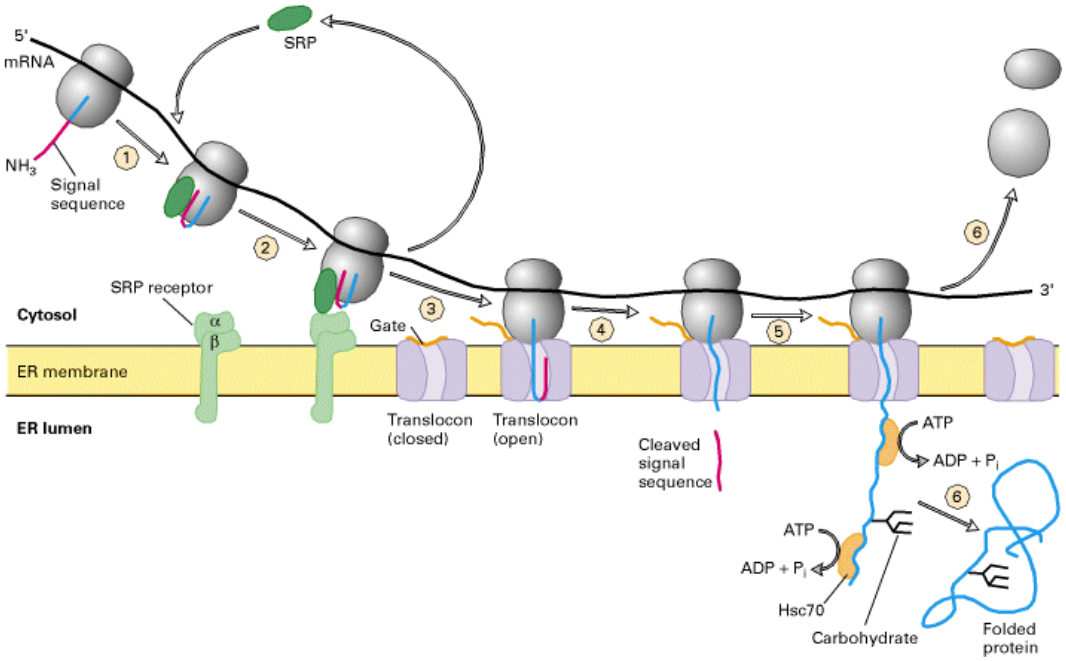


Figura 3: Síntese da proteína a ser secretada no Retículo Endoplasmático Rugoso.

Enquanto cruza o RE a proteína nascente passa por modificações como glicosilação (N-glicosilação) necessária para o empacotamento auxiliado por chaperonas, onde são adicionados resíduos específicos de oligossacarídeo ligados a Asparagina (Asn) (VARKI, 1993; KORNFELD; KORNFELD, 1985). A N-glicosilação desempenha um papel importante em proteínas que serão secretadas, e é muitas vezes necessária para a estabilidade da proteína. Os efeitos da N-glicosilação muitas vezes dependem do número e posição de olissacarídeos presentes na cadeia proteica. Isto é determinado durante a glicosilação onde os oligossacarídeos GLC (WEST, 1986), Man (MACHAMER; ROSE, 1988), GlcNAc (KORNFELD; KORNFELD, 1985), são transferidos para proteína pela enzima *oligossacaril transferase* que é integrante da membrana do retículo endoplasmático. A glicosilação

ocorre enquanto a proteína nascente cruza o lúmen do RE, em motivos de glicosilação (Asn-Xaa-Thr/Ser/Cys) pelo complexo OST (oligossacaril transferase). Esses motivos de glicosilação, são também denominados *sequon* e trata-se de uma sequência de três aminoácidos consecutivos, que pode ligar-se ao sítio de polisacarídeo (açúcar) chamado N-lig-Glycan, átomo de nitrogênio na corrente lateral da Asn. Um *sequon* é Asn-Xaa-Ser ou Asn-Xaa-Thr, onde Xaa é qualquer aminoácido exceto prolina (P). O termo parece ter sido usado primeiramente por Marshall (1974).

As enzimas na superfície luminal adicionam carboidratos à asparagina (Asn), serina (Ser), treonina (Thr), cisteína (Cys) (KALIES; HARTMANN, 1998; DWEK; BROOKS; SCHUMACHER, 2002). O complexo OST está próximo ao canal de transporte e sua estrutura específica reflete em eficiente glicosilação com os resíduos de aminoácido acessíveis para a posição Xaa, exceto para prolina, pois ela na posição Xaa inibe a glicosilação. Aminoácidos aspárticos e glutâmicos na posição Xaa não são favoráveis pelo reconhecimento do complexo OST. Análises estatísticas indicam que resíduos de prolina na posição Xaa próxima a posição de serina/treonina reduz grandemente o grau de glicosilação (GAVEL; von Heijne, 1990). Olissacarídeos N-glicosilados aumentam a solubilidade e estabilidade de muitas proteínas contribuindo para o enovelamento. Quando os carboidratos (*glicans*) faltam nas proteínas que serão secretadas, estas podem não enovelar-se sendo consequentemente degradadas dentro do RE (LOGANATHAN; AICH, 2006). A proteína nascente, continua a prolongar-se e os ribossomos são liberados para a síntese de outra proteína (Figura 3 [6]) (RAPOPORT, 1991; SANDERS *et al.*, 1992; POWERS; WALTER, 1996; MARTOGLIO; DOBBERSTEIN, 1996; BACHER *et al.*, 1996).

O retículo endoplasmático rugoso é uma série interconectada de sacos aplainados em camadas. O RE quebra-se em pequenas vesículas fechadas, denominadas microssomos rugosos. Durante o transporte e co-tradução no lúmen do RER a proteína a ser secretada é clivada da proteína madura pela enzima peptidase sinal. Se corretamente N-glicosilada e empacotada, a proteína entra no *Golgi apparatus* e é incorporada em pequenas vesículas de transporte em fusão cis-Golgi ou com outra forma da membrana como cis-Golgi retículo. Do cis-Golgi, determinadas proteínas são enviadas ao RER por um conjunto diferente de vesículas de transporte. Neste processo chamado migração cisternal ou progressão, um nova pilha cis-Golgi com proteína luminal move-se da posição cis (próxima ao RER) a posição trans (distante do RER), sucessivamente transportando-se para as cisternas medial-Golgi e trans-Golgi (GLICK; MALHOTRA, 1998; LADUNGA, 2000) (Figura 4).

Enquanto isto acontece, membrana e proteínas luminais são constantemente recupe-

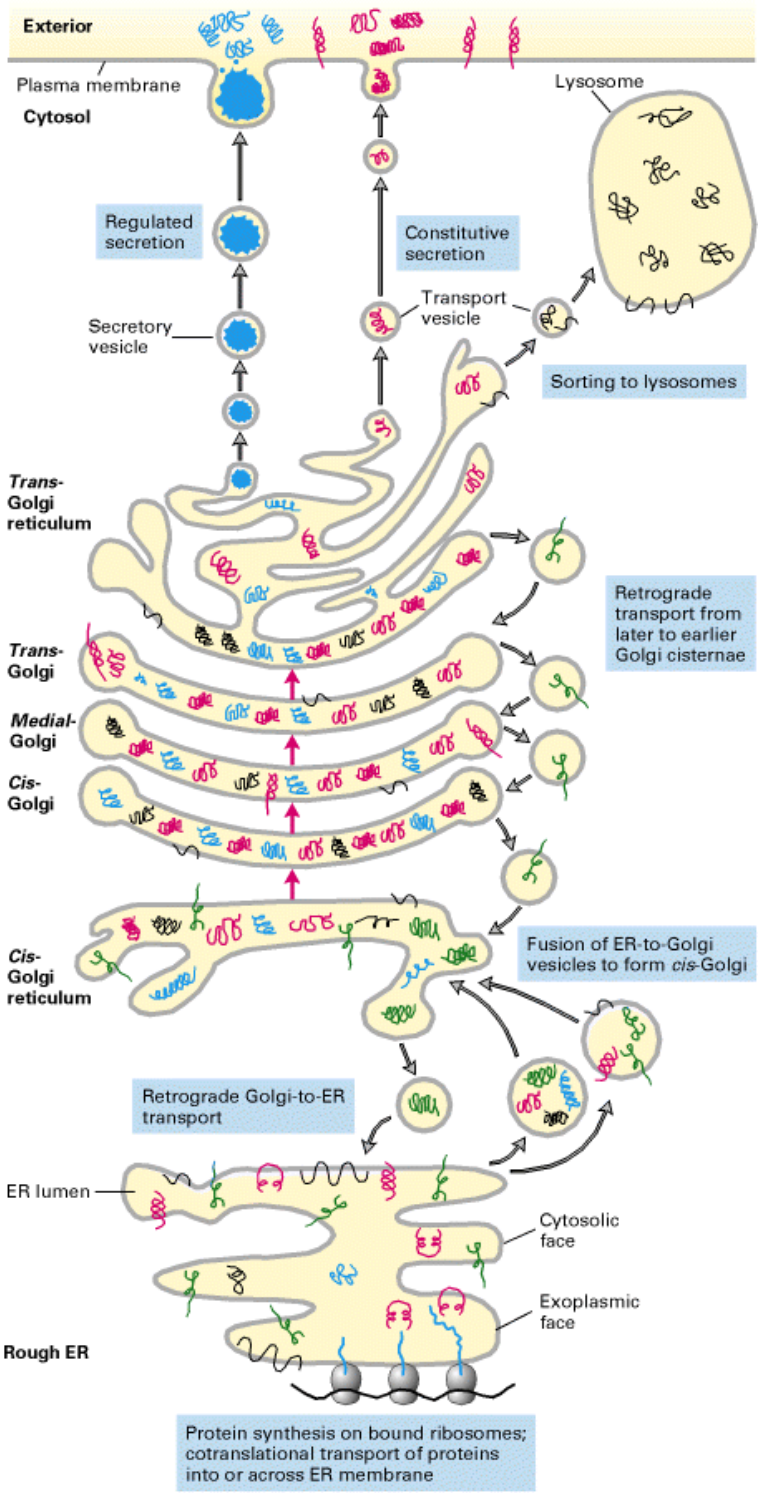


Figura 4: Via secretora da síntese da proteína. Os ribossomos sintetizam proteínas com peptídeo sinal transportando-as ao lúmen, migração cisternal Golgi e agregação aos grânulos secretores para exportar ao espaço extracelular.

radas da cisterna Golgi mais distante para a mais próxima por pequenas vesículas de transporte adquirindo modificações e agregando-se aos grânulos secretores para ser exportada ao espaço extracelular (SAKAGUCHI, 1997; LADUNGA, 2000).

1.2.2 Características do Peptídeo Sinal na Proteína Secretada

A característica mais comum do peptídeo sinal é o pequeno trecho de 7 a 15 aminoácidos hidrofóbicos chamado região hidrofóbica ou H-região. Entre a região N-terminal da pré-proteína e H-região é denominada de N-região e é do tamanho de 1 a 5 aminoácidos que normalmente são de carga positiva. Entre a H-região e sítio de clivagem está a C-região, que consiste de 3 a 7 aminoácidos polar *uncharged* (Figura 5).

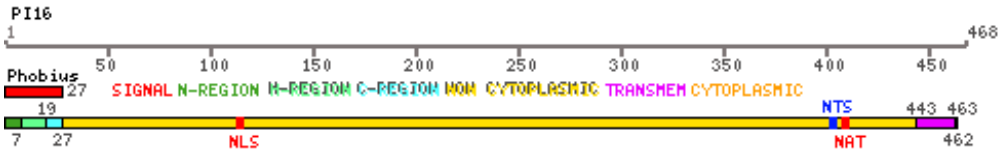


Figura 5: Regiões do peptídeo sinal encontradas pelo *software* SignalP3.0, na sequência PI16 e mapeada por programas *in-house script* (módulo PERL-GD).

Próximo ao sítio de clivagem é encontrado um padrão específico de aminoácidos e resíduos nas posições -3 e -1 (relativos ao sítio de clivagem) são pequenos e neutros (para a posição -1 e geralmente são **Ala, Gly, Ser, Cys ou Thr**) para que ocorra a clivagem corretamente (von Heijne, 1983, 1985). Em contraste, a posição -2 possui frequentemente carga aromática ou resíduo polar maior. O método matemático mais usado para predizer a localização do sítio de clivagem foi a matriz de pesos de von Heijne (1986), que também discrimina entre peptídeo sinal e não-peptídeo sinal usando *score* máximo de pontuação de sítio de clivagem. Redes neurais, na maioria das vezes do tipo *feed-forward backpropagation*, são utilizadas para resolver problemas biológicos e também aplicadas na previsão de peptídeo sinal e sítio de clivagem. Ladunga *et al.* (1991), utilizaram um algoritmo que ajusta a arquitetura da rede neural para discriminar entre peptídeo sinal e não-peptídeo sinal. No entanto, sua rede não superou o método da matriz de pesos, apesar de utilizar uma grande base de dados. Schneider e Wrede (1993), utilizaram redes neurais treinadas por um algoritmo genético para predição de sítios de clivagem. Nesse caso, o conjunto de dados era pequeno e os resultados não correspondiam ao do método da matriz de pesos. Resultados de uma rede neural não supervisionada auto-organizável mostrou boa performance para identificar sequências peptídeo sinal de um conjunto de dados de genes receptores de insulina em humano no trabalho de Arrigo *et al.* (1991). Hoje um certo número de ferramentas computacionais estão disponíveis para a detecção

de peptídeo sinal e suas habilidades para localizar peptídeo sinal e sítios de clivagem variam consideravelmente.

No trabalho de Zhang e Henzel (2004), foi caracterizado computacionalmente um conjunto de proteínas secretoras humanas utilizando o programa SignalP2.0-NN (NIELSEN *et al.*, 1997). Sua precisão foi de 78,1%, segundo os autores, este foi o melhor programa para previsão de sequências sinal. Em nosso estudo, utilizamos a ultima versão de SignalP 3.0 (BENDTSEN *et al.*, 2004), para predição de peptídeo sinal baseado em redes neurais e modelos escondidos de Markov. E também usamos Phobius, (KÄLL; KROGH; SONNHAMMER, 2004), baseado em modelos escondidos de Markov.

Dado ao fato de que o *software* SignalP 3.0-NN, utilizado neste projeto, usa Rede Neural Artificial (RNA). Para predição de peptídeo sinal, na seção seguinte será descrito a topologia de uma Rede Neural Artificial com noções básicas de funcionamento, algoritmo, modelos matemáticos e de correção, que são implementados para treinamento e aprendizado de uma RNA.

1.3 Teoria de Rede Neural Artificial (RNA)

Esta seção objetiva mostrar superficialmente o modelo matemático de uma Rede Neural Artificial. Uma Rede Neural Artificial, baseia-se em um modelo matemático que tenta imitar o mecanismo de processamento do cérebro humano. Ela usa processamento paralelo, para que os neurônios possam realizar o aprendizado simultâneo e que após treinamento possam armazenar conhecimento para executar a função desejada com eficiência (ALEKSANDER; MORTON, 1990).

Uma rede neural pode possuir uma ou múltiplas camadas. Como exemplo podemos descrever uma rede com três camadas: uma camada de entrada, onde as unidades recebem os padrões; uma camada intermediária, onde acontece o processamento e a extração das características; uma camada de saída, que conclui e apresenta o resultado final. Quanto maior o número de camadas, melhor a capacidade de aprendizado da rede. A camada de entrada deve possuir uma unidade especial conhecida como “bias”, usada para aumentar os graus de liberdade, permitindo uma melhor adaptação ao conhecimento a ela fornecido. O número de camadas define a capacidade de representação das relações entre o espaço de entrada e o de saída (NILSSON, 1965; FELDMAN; BALLARD, 1982). A inexistência da camada intermediária (modelo Perceptron) condiciona a rede a representar relações linearmente independentes. O Perceptron é o tipo de rede neural mais simples do tipo

feed-forward, uma vez que as entradas são direcionadas para a unidade de saída pelas conexões (ROSENBLATT, 1988) (Figura 6).

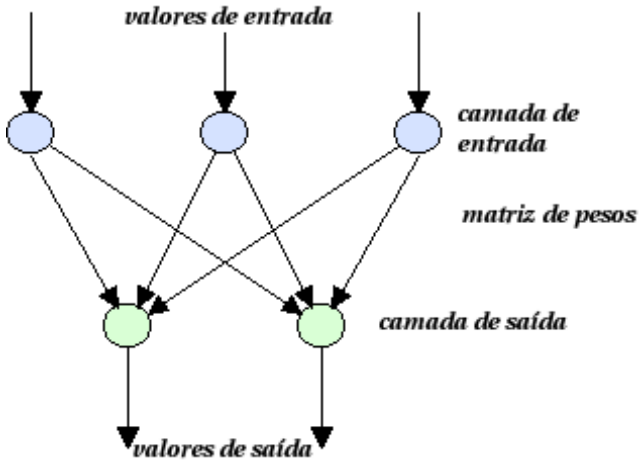


Figura 6: Estrutura simples de um Perceptron.

Ele usa matriz de valores para representar redes neurais *feed-forward* e é um classificador binário que mapeia a entrada x (vetor valores reais) para um valor de saída $f(x)$ (valor binário) em toda a matriz.

$f(x) = 1, \; if \; w \cdot x + b > 0$

$f(x) = 0, \; else$

Onde w é um vetor de valores reais e $w \cdot x$ o produto que calcula a soma ponderada. b é uma constante que não depende de qualquer valor de entrada ("bias"). O valor de $f(x)$ (0 ou 1) é usado para classificar x como um exemplo positivo ou negativo, no caso da classificação de um problema binário. O "bias" pode ser visto como a função de ativação, ou para proporcionar atividade na saída neurônio. Se b é negativo, então a combinação dos fatores de entrada deve apresentar um valor positivo superior a $-b$ para pressionar o neurônio classificador sobre o *threshold* 0. O Perceptron é o tipo de rede neural mais simples do tipo *feed-forward*, uma vez que as entradas são direcionadas para a unidade de saída pelas conexões (ROSENBLATT, 1988).

Já o Perceptron Multi-Camadas (MLP) é uma extensão do Perceptron de camada única. Possui uma camada com unidades de entrada conectada a uma ou mais camadas intermediárias ocultas e uma camada de saída. As camadas intermediárias, conhecidas como camadas ocultas, trabalham como um reconhecedor de características que armazenam nos pesos sinápticos (Figura 7).

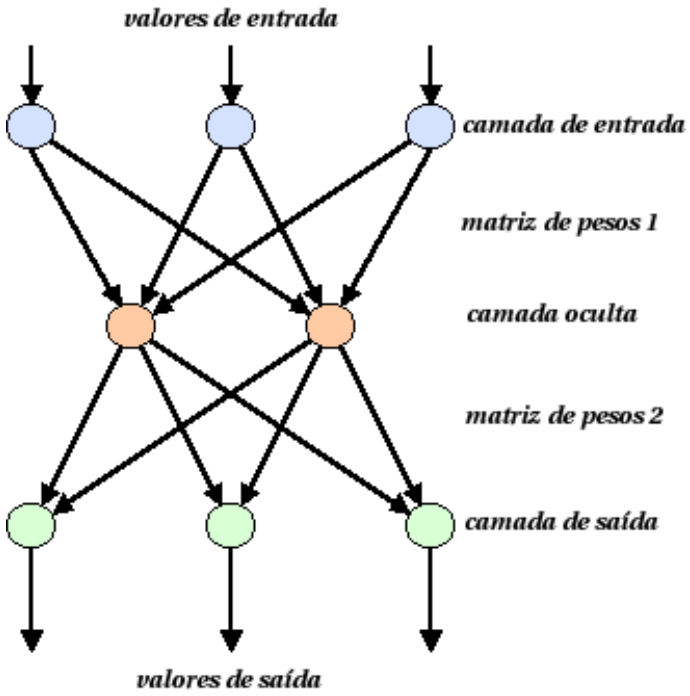


Figura 7: Estrutura do Perceptron Multi-camadas.

O algoritmo de treinamento mais utilizado em MLP é o *backpropagation*, que é um algoritmo de aprendizado supervisionado por correção de ERRO (ROSENBLATT, 1988). Algoritmo de aprendizado é um conjunto de regras definidas para a solução de um problema de aprendizado (Figura 8).

Dentre os algoritmos de aprendizado para determinados modelos de redes neurais tem-se:

Aprendizado Supervisionado (por correção de ERRO): quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;

Aprendizado Não Supervisionado (auto-organização - Aprendizado Competitivo e o Aprendizado Hebbiano): quando não existe um agente externo indicando a resposta desejada para os padrões de entrada;

Reforço, quando um crítico externo avalia a resposta fornecida pela rede;

Estes algoritmos diferem entre si pelo modo como os pesos são modificados. No *software* SignalP, a rede foi treinada usando o algoritmo para treinamento de redes Multi-Camadas *backpropagation*, com a função ERRO sugerida por Rumelhart, Hinton e Williams (1986),

O algoritmo *backpropagation* é constituído de:

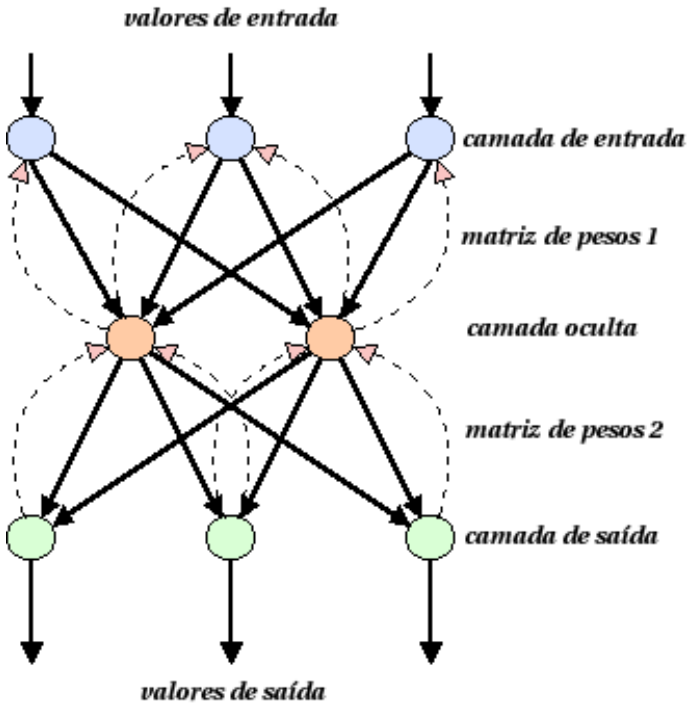


Figura 8: Estrutura do algoritmo *backpropagation*.

1. Propagação: Depois de apresentado o padrão de entrada, a resposta de uma unidade é propagada como entrada para as unidades na camada seguinte, até a camada de saída, onde é obtida a respostada rede e o ERRO é calculado;

2. Retropropagação (*backpropagation*): Desde a camada de saída até a camada de entrada, são feitas alterações nos pesos sinápticos. Durante a fase treinamento é apresentado um conjunto formado pelo par: entrada para a rede e valor desejado para resposta a entrada da rede. A saída é comparada ao valor desejado e é computado o ERRO global da rede, que influenciará na correção dos pesos, no passo de retro-propagação. Este processo é muito utilizado por apresentar uma boa solução para o treinamento de Perceptrons Multi-Camadas.

Resumo do Algoritmo *backpropagation*:

1. Inicialização: Inicialize os pesos sinápticos e os "bias" aleatoriamente, com valores no intervalo $[-1; 1]$;

2. Apresentação dos Exemplos de Treinamento: **Treinamento "on-line"**: para cada exemplo do conjunto de treinamento, efetue os passos 3 e 4. **Treinamento "em lote"**: para cada "época" do conjunto de treinamento, efetue os passos 3 e 4.

3. Computação para frente (propagação): depois de apresentar o exemplo do conjunto de treinamento

$$T = \{(x(n), d(n)\}, \tag{1.1}$$

sendo $x(n)$ a entrada apresentada à rede e $d(n)$ a saída desejada, a rede calcula o valor da ativação vj e a saída para cada unidade da seguinte forma:

$$v_j = \Sigma_{i=1}^{n} w_{ji} x_i + b, \tag{1.2}$$

para o cálculo do valor da ativação e

$$f(v) = \frac{1}{1 + e^{(-\alpha v)}}, \tag{1.3}$$

para o cálculo da saída y da unidade k, utilizando a função sigmoide, como no exemplo ou uma outra função se necessário. A rede utiliza a saída das unidades de uma camada como entradas para a seguinte, até a última camada. A saída das unidades da última camada será então a resposta da rede.

4. Cálculo do ERRO: fazendo a saída $y_j = O_j(n)$, $O_j(n)$ será a resposta da rede, o sinal de ERRO é calculado pela fórmula:

$$e_j(n) = d_j(n) - O_j(n), \tag{1.4}$$

onde $dj(n)$ é a saída desejada com resposta para cada unidade na interação (n). Este sinal de ERRO é utilizado para computar os valores dos erros das camadas anteriores e fazer as correções necessárias nos pesos sinápticos.

5. Computação para trás (retro-propagação): cálculo dos erros locais, d, para cada unidade, desde a camada de saída até a de entrada. O gradiente local é definido por:

$$\delta_j(n) = e_i(n) O_j(n)(1 - O_j(n)) \tag{1.5}$$

para a unidade da camada de saída ou

$$\delta_j(n) = O_j(n)(1 - O_j(n)) \sum \delta_k w_{jk}, \tag{1.6}$$

para as unidades das demais camadas.

Onde:

$O_j(1-O_j)$, é a função de ativação diferenciada em função do argumento, valor de ativação; d_k, é o erro das unidades da camada anterior conectadas a unidade j; w_jk, são os pesos das conexões com a camada anterior.

Após o cálculo dos erros de cada unidade, é calculado o ajuste dos pesos de cada conexão segundo a regra delta generalizada e atualizado os pesos:

$$\Delta w_{kj}(n+1) = \alpha w_{kj}(n) + \eta \delta_j y_j, \tag{1.7}$$

para o cálculo e ajustes dos pesos sinápticos é efetuado:

$$w(n+1) = w(n) + \Delta w_{kj}(n), \tag{1.8}$$

onde:

a, é a constante, quando $a = 0$, esta função funciona como a regra delta comum;

h, é a taxa de aprendizado;

d_j, é o ERRO da unidade;

y_j, é a saída produzida pela unidade j;

6. Interação: é refeito os itens 3, 4 e 5 referentes à propagação, o cálculo do ERRO, e retro-propagação, apresentando outros estímulos de entrada, até que sejam satisfeitas as condições de treinamento da rede, que podem ser:

   • o erro da rede está baixo, sendo pouco alterado durante o treinamento; • o número máximo de ciclos de treinamento foi alcançado;

Na modelagem da arquitetura de uma Rede Perceptron Multi-Camadas é importante a escolha do número de unidades de entrada, a definição da função de ativação que irá ditar o comportamento da rede, a codificação da camada de saída e a formatação da resposta da rede. Outros parâmetros são escolhidos, como a taxa de aprendizado e o conjunto de treinamento. O processamento de cada unidade é influenciado pelo processamento efetuado pelas unidades das camadas anteriores, pois cada camada desempenha um papel específico, com elevado grau de confiança (`http://www.lncc.br/~labinfo/tutorialRN/frm4_perceptronMultiCamadas.htm`).

Esta seção objetivou mostrar superficialmente o modelo matemático de uma Rede Neural Artificial, explicando como ela imita o mecanismo do processamento do cérebro humano, possuindo processamento paralelo, para que os neurônios possam realizar o aprendizado simultaneamente e pelo processo de aprendizagem armazenar conhecimento e torná-lo disponível.

Dando seguimento ao estudo de redes neurais, em Material e Métodos, é colocado de forma objetiva e simples o algoritmo da rede neural do *software* de predição SignalP-NN, como funcionam os cálculos das duas redes neurais do *software* no algoritmo, as saídas de pontuação e os valores numéricos que são calculados com as saídas das redes. A subseção seguinte apresenta a Rede Neural Artificial (RNA) do *software* SiganlP-NN (BENDTSEN *et al.*, 2004), a estrutura do algoritmo, o tipo de aprendizado da rede e o tipo de algoritmo de aprendizado supervisionado por correção de ERRO para treinamento da rede.

1.3.1 Treinamento da Rede Neural de SignalP-NN

Bendtsen *et al.* (2004), utilizaram o algoritmo do tipo *backpropagation*, para resolver o problema biológico de análise da sequência predizendo peptídeos sinal e sítios de clivagem (LADUNGA, 2000). Os dados para treinamento foram extraídos do SWISS-PROT *version* 29 (BAIROCH; APWEILER, 1997). A abordagem utilizou duas redes neurais. O *software* SignalP-NN foi implementado com o algoritmo em *POSIX-compliant shell script*, para que uma das redes fizesse o reconhecimento do peptídeo sinal e a outra para reconhecimento de sítios de clivagem (BENDTSEN *et al.*, 2004).

A rede neural foi treinada usando o algoritmo *backpropagation*, que utiliza a função de ERRO sugerida por McClelland:

$$E = -\sum_{\alpha,i} \log(1 - (O_i^\alpha - T_i^\alpha)^2), \tag{1.9}$$

ao invés da função ERRO convencional

$$E = \sum_{\alpha,i} (O_i^\alpha - T_i^\alpha), \tag{1.10}$$

onde O_i^α e T_i^α são a saída e valores alvo respectivamente para treinamento α. A função ERRO logarítmica reduziu a convergência consideravelmente e também possui propriedade de aprendizado à rede para tarefas mais complexas (comparada a medida

de ERRO padrão) (RUMELHART; MCCLELLAND, 1986), sem aumentar o tamanho da rede. O aprendizado foi mantido constante em 0.025. Os pesos foram atualizados e o treinamento para valores alvo foram 1.0 para exemplos positivos (sítio de clivagem e peptídeo sinal) e 0.0 para exemplos negativos. Para avaliar a saída da rede uma nota de corte de 0.5 foi vista como positiva. O coeficiente de correlação foi calculado com preditos corretos e incorretos pelo coeficiente de correlação (MATTHEWS, 1975) usando a fórmula seguinte:

$$C = \frac{(P^t N^t) - (N^f P^f)}{\sqrt{(N^t + N^f)(N^t + P^f)(P^t + N^f)(P^t + P^f)}}, \tag{1.11}$$

onde P^t e P^f são os números verdadeiro e falso positivos, enquanto que N^t são os números verdadeiros e falso negativos. O coeficiente de correlação de ambos os treinamentos foram monitorados e a performance foi salva em cada corrida. O teste de performance foi calculado por *cross-validation* (EFRON, 1983), onde cada conjunto foi dividido em 5 partes aproximadamente iguais e toda rede foi carregada com uma parte dos dados de teste e outras quatro partes com dados de treinamento. A performance calculou uma média sobres os 5 diferentes conjuntos de dados.

O treinamento da rede proporcionou *scores* entre 0 e 1 para cada um aminoácido. O resultado do peptídeo sinal, *S-score*, foi interpretado como uma estimativa de probabilidade de posições em torno do peptídeo sinal. Enquanto a saída de sítio de clivagem e o não sítio de clivagem da rede, o *C-score*, foi interpretado como uma estimativa da probabilidade da posição inicial na proteína madura (posição +1 relativa ao sítio de clivagem). Na figura 9, dois exemplos de valores de *C-score* e *S-score* para peptídeos sinal. Um peptídeo sinal típico com um sítio de clivagem típico, mostra como apresentam as curvas de posições da rede, onde o *C-score* tem um acentuado pico que corresponde a uma mudança em *S-score*. Em outras palavras, o exemplo tem 100% de posições previstas corretas na rede de SignalP, tanto em função *C-score* como em *S-score*. Exemplos menos típicos pode-se observar na figura 9 (b) onde o *C-score* tem vários picos.

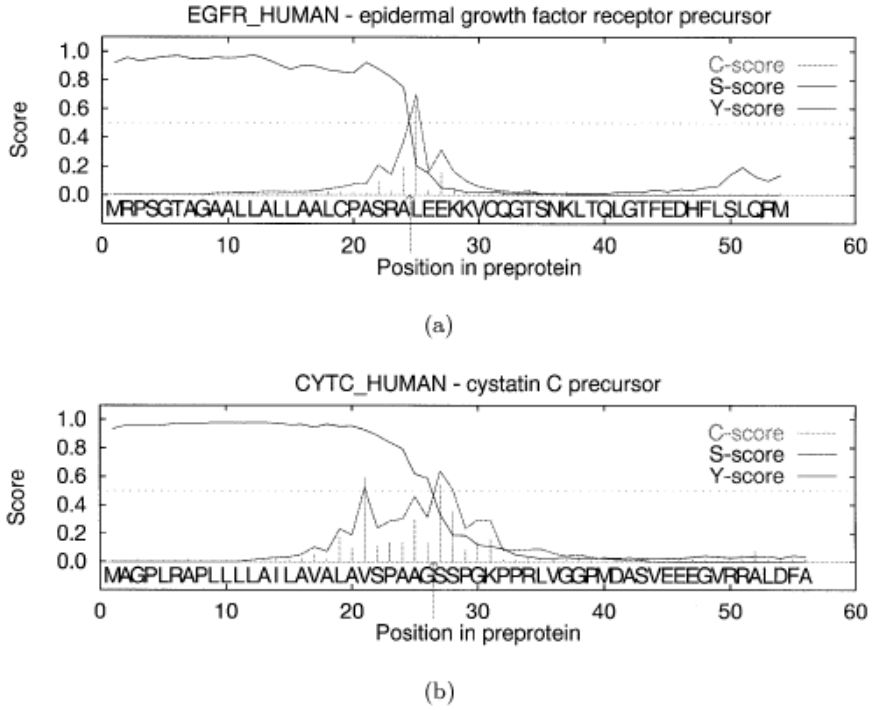


Figura 9: Exemplo de sequência predita no treinamento da rede neural SiganlP verificando sítio de clivagem e peptídeo sinal, valores *C-score*, *S-score*, e *Y-score*. O verdadeiro sítio de clivagem está marcado com seta. Em (a) tem-se um exemplo da sequência com todas as posições corretamente previstas de acordo com ambos os *C-score* e *S-score*, e em (b) tem-se duas posições com *C-score* superior a 0.5. O verdadeiro sítio é previsto incorretamente se e somente se o valor máximo de *C-score* estiver sozinho, mas combinado com o cálculo do *Y-score*, a previsão torna-se mais precisa.

O *cutoff* como positivo para *C-score* e *S-score* em SignalP é de **0.5**, e é considerado ótimo, de acordo com o coeficiente de correlação de Pearson para o caso de eucariotos:

$$r = \frac{\sum_i^n = 1(x_i - \ddot{x})(y_i - \ddot{y})}{\sqrt{\sum_i^n = 1(x_i - \ddot{x})^2}\sqrt{\sum_i^n = 1(y_i - \ddot{y})^2}} \tag{1.12}$$

onde $x_1, x_2, ..., x_n$ e $y_1, y2, ..., y_n$ são os valores medidos de ambas as variáveis, onde além disso

$$\ddot{x} = \frac{1}{n}.\sum_i^n = 1x_i \tag{1.13}$$

e

$$\ddot{y} = \frac{1}{n}.\sum_i^n = 1y_i \tag{1.14}$$

são as médias aritméticas de ambas as variáveis. A análise correlacional indica a relação entre 2 variáveis lineares e os valores sempre estarão entre $+1$ e -1. Este coeficiente, normalmente representado por r assume apenas valores entre -1 e 1.

$r = 1$, significa uma correlação positiva entre as duas variáveis. $r = -1$, significa uma correlação negativa entre as duas variáveis, isto é, se uma aumenta, a outra sempre diminui. $r = 0$, significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado $r = 0$ deve ser investigado por outros meios.

Para formalizar esta situação e melhorar a previsão na rede de SignalP, a rede foi treinada com uma série de combinações linear e não-linear dos *scores* brutos da rede e avaliou-se a porcentagem de sequências com corretos sítios de clivagem dos 5 conjuntos teste. A melhor medida e matematicamente simples foi a média geométrica do *C-score* derivado do *S-score* chamada *Y-score*:

$$Y_i = \sqrt{C_i \Delta_d S_i} \tag{1.15}$$

onde $\Delta_d S_i$ é a diferença entre a media *S-score* da posição d depois da posição i:

$$\Delta_d S_i \frac{1}{d}(\sum_j^d a = 1 S_{i-j} - \sum^d - 1_j = 0 S_{i+j}) \tag{1.16}$$

Para cada conjunto de dados em SignalP, foi escolhido no treinamento o valor d que resulta na melhor sequência no nível de performance. O *Y-score* proporciona a porcentagem de sequências corretas encontradas relativas ao *C-score* a fim de verificar se as previsões positivas de peptídeo sinal são melhores que os falsos positivos.

Para comparar as respostas dos neurônios na rede de SignalP, uma abordagem computacional com um método tradicional é feita para os resultados utilizando a matriz de pesos usada por von Heijne (1985). Após o sítio de clivagem, definido pelo algoritmo de Viterbi (RABINER *et al.*, 1989), o desempenho da matriz de pesos e os resultados do nível de performances, correspondem a matriz de pesos usada para predição de peptídeo sinal (CHOU; KENDALL, 1990). De maneira geral, o treinamento da rede neural de SignalP 3.0, teve como objetivo discriminar proteínas com peptídeo sinal e sem peptídeo sinal ou com sinal âncora e sem sinal âncora. Na seção seguinte será apresentado a parte teórica dos modelos ocultos de Markov, ou modelos escondidos de Markov, utilizados pelos *softwares* SignalP 3.0-HMM (BENDTSEN *et al.*, 2004) e Phobius (KÄLL; KROGH;

SONNHAMMER, 2004), os sítios de clivagem e o peptídeo sinal e no caso do algoritmo Phobius, inclui a previsão de peptídeo sinal, previsão de não-peptídeo sinal, topologia transmembrana, região citoplasmática.

1.4 Teoria dos Modelos Ocultos de Markov (HMMs)

O HMM pode ser descrito como um processo estocástico com pouca memória. Os algoritmos são baseados em modelos matemáticos aplicados a um processo estocástico, e utilizados em muitos problemas de classificação e processamento de padrões (RABINER *et al.*, 1989; FORSYTH; PONCE, 2002).

Um estado no HMM é um evento e há transições que descrevem a probabilidade de mover-se para um outro estado. Os modelos de Markov são chamados ocultos pois os estados não são diretamente observáveis. Para cada estado existe a uma probabilidade de observações (DUDA; HART; STORK, 2001). Nos modelos de Markov, a saída de cada estado corresponde a distribuição de probabilidade de emissão ao invés de um evento determinístico, onde as probabilidades de emissão impõem uma sequência de estados, o estado observado e a sequência de estados escondida (MITCHELL, 1997; YEH; LIM; BURGE, 2001).

O HMM geralmente é representado por um modelo matemático do tipo tripla:

$$\lambda = (\pi, A, B), \tag{1.17}$$

onde π é a distribuição inicial dos estados, A é uma matriz de probabilidade de transições e B a matriz de probabilidades dos símbolos de observação.

Outras definições são necessárias para representar um modelo oculto de Markov como N = número de estados, $S = (S_1, S_2, ..., S_N)$, representando o conjunto de estados, M = número de observações possíveis, $V = (V_1, V_2, ..., V_M)$, representando o conjunto de diferentes observações e q_t é um estado no instante t. Cada elemento π_j é calculado através da probabilidade $P(q_1 = S_j)$, onde $1 <= j <= N$. Esses elementos representam a probabilidade do modelo iniciar no estado j. Os elementos da matriz A, ou seja, cada A_{ij}, são calculados pela probabilidade $P(q_t = S_j | q_{t-1} = S_i)$, onde $1 <= i, j <= N$. Esses elementos representam a probabilidade do próximo estado ser S_j sabendo que o estado anterior era S_i. Cada elemento da matriz $B, B_{j(k)}$ é calculado pela probabilidade $P(O_t = V_k | q_t = S_j)$, onde $1 <= j <= N <= k <= M$ e O_t é a observação no tempo t. Esses elementos representam

a probabilidade da observação ser igual a V_k e ser gerada pelo estado S_j.

HMMs tem sido amplamente utilizados em diversas áreas, principalmente em sistemas para o reconhecimento de voz (RABINER *et al.*, 1989) e textos manuscritos (HU; BROWN; TURIN, 1996; HORNEGGER *et al.*, 1994). Em um trabalho Starner e Pentland (1995), descrevem um sistema para reconhecimento de voz em sentenças da língua americana utilizando HMM, onde a taxa de reconhecimento foi de cerca de 99.2% para palavras. Para o conjunto de características de textos manuscritos o sistema mostrou-se limitado, pois foi treinado para verificar as palavras sem as posições da mão durante a escrita e para corrigir esse problema Starner e Pentland (1995), sugerem a utilização de um vetor de características de posições da mão durante a escrita.

Modelos ocultos de Markov são também bastante utilizados para descrever uma sequência de padrões comportamentais. Os diversos sistemas que fazem o uso de modelos ocultos de Markov para o reconhecimento de comportamentos apresentam bons resultados. Entretanto, a taxa de reconhecimento depende muito do conjunto de características visuais utilizadas e do número de estados do modelo (SPINK *et al.*, 2001).

Os modelos ocultos de Markov (THEODORIDIS; KOUTROUMBAS, 1999; FORSYTH; PONCE, 2002; YEH; LIM; BURGE, 2001), possuem como proposta modelar essas mudanças e alterações através de probabilidades. O objetivo do projeto SignalP 3.0 (BENDTSEN *et al.*, 2004) e Phobius (KÄLL; KROGH; SONNHAMMER, 2004), utilizando modelos ocultos de Markov foi obter um sistema que automaticamente identifique proteínas classicamente secretadas, pelas características apresentadas ao modelo oculto.

Essa seção visou apresentar e exemplificar a parte teórica de fatores que limitam e são importantes no desenvolvimento de *softwares* com modelos ocultos de Markov em Bioinformática. As subseções seguintes terão como prioridade apresentar a estratégia e o algoritmo escolhido pelos *softwares* SignalP-HMM e Phobius e Material e Métodos apresentará a forma e as características do algoritmo.

1.4.1 Predição de Peptídeo Sinal por SignalP-HMM

O *software* SignalP-HMM (BENDTSEN *et al.*, 2004) foi desenvolvido em *POSIX-compliant shell script* para identificar o peptídeo sinal típico de uma proteína secretada pela via clássica. A arquitetura do SignalP-HMM foi inalterada e constituída de diagramas de estados com quatro tipos de proteínas estados: peptídeo sinal, sinal âncora, citoplasmáticas e (para eucariotos) nuclear. Todos os conjuntos foram agrupados no modelo de

diagrama de estados em subconjuntos, excluindo peptídeos sinal menores que 15 e maiores que 50 aminoácidos, pois são incomuns e difíceis de ser modelados por HMM, e também com base na sequência sinal N-terminal que é de aproximadamente 70 aminoácidos, com iniciador **Met**, N-região, H-região e C-região (NIELSEN; KROGH, 1998).

No trabalho de Nielsen e Krogh (1998), um submodelo oculto de Markov foi implementado para previsão de N-terminal, região hidrofóbica e região entorno do sítio de clivagem. Em SignalP, o modelo oculto de Markov utilizou um número de estados proteicos que são conectados por transições probabilísticas, associadas com distribuição de aproximadamente 20 aminoácidos. Por ser probabilístico, o modelo pode usar métodos e padrões de máxima verossimilhança (Gavin J. Gibson; RENSHAW, 1995), para determinar os modelos do parâmetro (DURBIN *et al.*, 1998).

Para peptídeos sinal, Nielsen e Krogh (1998), desenvolveram o modelo SignalP, que é parte correspondente de cada uma das três regiões do peptídeo sinal e o tamanho razoável é constantemente difícil de ser modelado. Para controlar a complexidade dos dados, eles estimaram do modelo simples, uma grande quantidade e variabilidade de dados para desenvolver o modelo.

Os dados foram extraídos do SWISS-PROT versão 35 (BAIROCH; APWEILER, 1997). O conjunto de dados foi utilizado para quatro tipos de proteínas: peptídeo sinal, sinal âncora, citoplasmática, e nuclear (eucariotos). Os conjuntos foram agrupados em subconjuntos para a extração das características peptídeos sinal e foram anotados para 15 e maior que 50 aminoácidos. Foram também removidas proteínas sem o iniciador **Met** e proteínas em todos os conjuntos foram truncadas após 70 resíduos, que é a região escolhida pelo *software* SignalP ao modelo, pois quase todos os peptídeos sinal são menores que 70 aminoácidos. Os conjuntos de dados de SignalP foram reduzidos por homologia, pois primeiro limita o "bias" do HMM para famílias representadas e para permitir que conjuntos devem ser utilizados para trans-validação e então finalmente cada conjunto foi dividido em cinco partes de igual tamanho aproximadamente para *cross-validation* (EFRON, 1983). Para ter uma ideia do tamanho e distribuição dos aminoácidos no modelo de Markov de SignalP, as três regiões diferentes do peptídeo sinal foi atribuído um procedimento na tentativa de encontrar N-região, H-região e C-região (DURBIN *et al.*, 1998).

Em SignalP para a H-região o conjunto de regras de HMMs, deve possuir necessariamente resíduos hidrofóbicos. Os diagramas do conjunto peptídeo sinal no conjunto de dados foram atribuídos a uma H-região, de 6-20 resíduos de comprimento, uma C-região

por definição com pelo menos 3 resíduos longos e uma N-região entre 2-7 resíduos. Estas regiões definidas no modelo foram utilizadas na modelagem e no comprimento mínimo de cada região no diagrama de estado, onde todos os estados estão ligadas entre si o que significa que possuem aminoácidos de distribuição. Os limites à duração da H-região, em conformidade com os dados experimentais foram de que se inferiores a 6 aminoácidos, não seriam capazes de promover a translocação (BIRD; GETHING; SAMBROOK, 1990). A transição de clivagem / não clivagem parece ocorrer entre clivagem da H-região, no comprimento de 17 e 20 aminoácidos. A H-região é modelada por uma matriz de 8 estados, onde os últimos 7 estados também são ligados uns aos outros (CHOU; KENDALL, 1990; NILSSON, 1965).

A C-região é modelada em um *array* com 6 estados, onde cada estado possui uma específica distribuição para capturar o padrão de distribuição do aminoácido, pouco antes do sítio de clivagem. Para permitir mais C-regiões, mais 4 C-estados são adicionados, ligados uns aos outros para captar o excesso dos aminoácidos nas regiões C com mais de 6. No último estado H existem transições a todos os C-membros, exceto aos 2 estados, pouco antes do sítio de clivagem fazendo com que o comprimento mínimo de C-regiões sejam igual a 3. Após o sítio de clivagem, quatro estados do modelo antes da transição é inserido para o final com a distribuição do aminoácido igual para o padrão de distribuição. Os seis estados anteriores ao sítio de clivagem e os 4 estados após o sítio de clivagem correspondem aproximadamente a matriz de pesos utilizada por von Heijne e Abrahmsén (1989).

Os modelos de SignalP foram estimados a partir da formação de dados pelo algoritmo de Baum-Welch (RABINER *et al.*, 1989), máxima verossimilhança (Gavin J. Gibson; RENSHAW, 1995), que é um procedimento que iterativamente aumenta a probabilidade da formação total dos dados. O treinamento foi feito por Nielsen e Krogh (1998), com os dados marcados, de tal forma que a clivagem local sempre foi corretamente posicionada durante a formação, mas o modelo foi responsabilizado a descobrir por si mesmo onde colocar as fronteiras entre as N-, H-, e C-regiões. Para predizer a clivagem local para uma nova sequência, o caminho provável que utiliza a formação modelo de SignalP é encontrado pelo algoritmo padrão Viterbi (VITERBI, 1967; RABINER *et al.*, 1989). O caminho mais provável também foi utilizado para a atribuição de uma região para cada aminoácido na sequência.

Concluindo esta subseção, o *software* SiganlP baseado em modelo oculto de Markov pode ser utilizado para rotular as três diferentes regiões de um peptídeo sinal, segundo

os autores, produzindo bons resultados. O algoritmo foi também desenvolvido para fazer uma boa discriminação entre peptídeo sinal, sinal âncora, sítio de clivagem e probabilidade de ser um peptídeo, mas a grande importância da aplicação para encontrar peptídeo sinal em SignalP-HMMs é analisar, grandes bases de dados.

1.4.2 Predição de Peptídeo Sinal e Topologia Transmembrana por Phobius

No algoritmo com modelos ocultos de Markov Phobius, Käll, Krogh e Sonnhammer (2004), combinam predição da topologia transmembrana e peptídeo sinal. O modelo oculto de Markov em Phobius foi treinado e pode ser considerado com a combinação dos modelos de SignalP-HMM e de TMHMM (*Prediction of transmembrane helices in proteins*), com a transição do último estado da proteína secretada em SignalP-HMM para o estado *loop* em TMHMM (KROGH *et al.*, 2001). Nesta transição do perfil HMM Phobius cada estado tem a sua própria probabilidade de emissão e a emissão de probabilidades dos diagramas de estados, além de semelhantes, são compartilhados e no modelo os estados são referidos como compartimentos.

A topologia convencional de preditores como TMHMM, possuem consideráveis sobreposições entre as previsões de SignalP-HMM. Para solução deste problema Phobius, propõe um modelo de Markov combinado com previsões da topologia transmembrana de TMHMM e SignalP-HMM para peptídeo sinal. Para resolver as ambiguidades, Phobius propôs submodelos sinal para ambos os peptídeos e seguimentos transmembrana, peptídeos sinal (N-região, H-região, C-região), submodelos *loops* citoplasmáticos e dois diferentes submodelos para *loops* não citoplasmáticos (Figura 10).

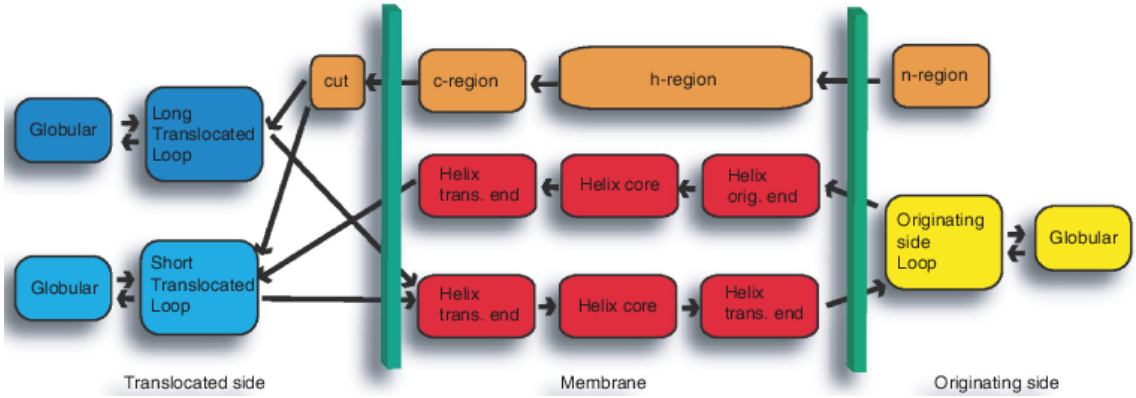


Figura 10: Figura esquemática do modelo Phobius que compreende de submodelos peptídeos sinal, hélices transmembrana, *loops* citoplasmáticos e dois diferentes submodelos para *loops* não citoplasmáticos.

O *software* Phobius foi implementado com o algoritmo em PERL *script* e as previ-

sões normais são feitas com o pacote HMM, oferecendo uma média fácil e precisa para previsão de peptídeos de sinal e topologia transmembrana a partir de uma sequência de aminoácidos. As sequências devem ser apresentadas em formato FASTA e previsões são dadas quer em "short" ou "long" - linha de texto ou de saída que são características estilo UniProt.

Todas as previsões feitas pelo algoritmo Phobius são rotuladas pela localização e probabilidade de uma localização (citoplasmática, não-citoplasmática, membrana ou peptídeo sinal) de uma dada sequência. Em Phobius maximizaram e implementaram, por uma alteração em cálculos *forward-backward* (RABINER *et al.*, 1989), a região para especificar o N-terminal, que é parte da sequência peptídeo sinal. O objetivo foi multiplicar a probabilidade de transmitir um estado nas posições da sequência.

Assim como em SignalP-HMM o modelo peptídeo sinal Phobius, usa um modelo um sinal âncora, ou seja, uma proteína com seguimento transmembrana, e um segmento N-terminal da proteína ajuda a discriminar os falsos positivos. Da mesma forma modelos N-terminal usam hélice transmembrana, peptídeos sinal e lipoproteínas peptídeos sinal em bactérias *gram*-negativas para melhorar a discriminação entre estas categorias (KÄLL; KROGH; SONNHAMMER, 2004).

O método do algoritmo HMM Phobius é baseado em modelos ocultos de Markov com o objetivo de predição tanto da topologia transmembrana de uma proteína quanto da presença do peptídeo sinal. A arquitetura do modelo pode ser considerada como uma combinação de modelos feitos em TMHMM (KROGH *et al.*, 2001) e SignalP-HMM (BENDTSEN *et al.*, 2004). Porém várias modificações foram feitas e adaptadas para ambos os modelos no *softwware*. No modelo com as modificações Phobius a arquitetura visa melhor desempenho, onde cada compartimento/estado tem o sua própria e individual probabilidade de emissão, sendo vinculados os estados a várias partes do modelo, ou seja a emissão das probabilidades são semelhantes e um estado depende do modelo dos demais estados referidos como compartimento (KROGH *et al.*, 1994).

Em Phobius a hélice transmembrana no submodelo compreende de 3 compartimentos, 4 resíduos hélice citoplásmico final e é seguido por um resíduo hélice núcleo de 7-26, e seguido por 4 resíduos hélice não citoplasmáticos finais. Os seguimentos transmembrana podem ser entre 15 e 34 aminoácidos e as probabilidades de emissão de todos os 3 compartimentos são interligadas entre as entradas e saídas no modelo. O submodelo peptídeo sinal é iniciado pela metionina **Met** dividido nas regiões N, H, C e pós-clivagem. Os diagramas de estados permitem probabilidade de comprimento e as distribuições são dis-

postas em ordem de menor e máximo comprimento da N-região. As demais regiões H e C são constituídas de compartimentos estados com transições. Em contraste com SiganlP-HMM que não permite auto-transição no modelo, o sítio de pós clivagem é constituído por 4 estados e a região em torno da clivagem contém 10 estados, fazendo parte da matriz. O ciclo de submodelos consiste de 20 compartimentos, um estado que permita qualquer comprimento *loop* que tenha entre 1 e 20 resíduos ou maior, um estado auto-*looping* globular está conectado entre membros 10 e 11, com 3 diferentes *loops* modelos: ***cytoplasmic loop, the short non-cytoplasmic loop and the long non-cytoplasmic loop***. O motivo da separação dos compartimentos *loops* é que podem conter domínios globulares. O objetivo é produzir um modelo que favoreça *loops* maiores.

O treinamento em Phobius aconteceu sem a divisão da sequências por reino, ou seja, eucarioto, *archea*, bactérias *gram*-positivas e *gram*-negativas, pois não obteve desempenho nos resultados em separá-las. O procedimento de treinamento foi TMHMM 2.0.1, mas modificando "bias" de acordo com o tamanho para regiões com hélice transmembrana e peptídeo sinal.

Esses treinamentos incluíram os passos de testar a acurácia de Phobius por *cross-validation* (EFRON, 1983), nos 9 ou 10 dados de subconjuntos de treinamento. Incluíram também verificar a probabilidade de distribuição com máxima verossimilhança (Gavin J. Gibson; RENSHAW, 1995), para determinar os modelos do parâmetro, mas mantendo o sítio de clivagem e estimando o ruído a partir da formação dos dados pelo algoritmo de Baum-Welch (RABINER *et al.*, 1989).

Em seguida os parâmetros do modelo foram atualizados para discriminação condicional de máxima verossimilhança (Gavin J. Gibson; RENSHAW, 1995), maximizando a probabilidade correta em vez da probabilidade de sequências observadas. Foram medidos o desempenho dos dez modelos, de modo que um único modelo foi obtido e o melhor algoritmo para predição em Phobius foi o modelo utilizado para realizar as previsões consideradas adequadas para modelos treinados pelo decodificador com a máxima verossimilhança (Gavin J. Gibson; RENSHAW, 1995), maximizando a probabilidade correta em vez da probabilidade de sequências observadas (KROGH *et al.*, 1994).

Finalizando esta subseção, o *software* Phobius baseado em modelo oculto de Markov possui objetivo de previsão da topologia transmembrana e peptídeo sinal, verificando a alta similaridade entre regiões hidrofóbicas de uma hélice transmembrana e um peptídeo sinal, fazendo a reação cruzada entre as previsões, e obtendo a informação da topologia das diferentes regiões de uma proteína transmembrana interligadas em uma série de estados.

1.5 Objetivos

O presente projeto tem como objetivo central usar abordagens computacionais para descrever a primeira lista do Secretoma Humano. Para tanto, as seguintes atividades serão realizadas

1 - Encontrar nas sequências as características comuns das proteínas secretoras tais como peptídeo sinal (N-região, H-região, C-região).

2 - Localizar, nas proteínas com peptídeo sinal os sítios de glicosilação *sequon*: **Asn-Xaa-Thr, Asn-Xaa-Ser, Asn-Xaa-Cys**.

3 - Validar a lista com dados de proteínas secretoras caracterizadas experimentalmente.

4 - Analisar o nível de expressão das proteínas secretoras durante a osteogêneses em dados de SAGE do tecido Osteoblasto

5 - Pela análise dos dados *in silico*, validar experimentalmente candidatas as proteínas secretadas, utilizando o processo de cultura celular, gel de poliacrilamida com análise em espectrômetro de massa Maldi TOF-TOF *Mascot*.

2 Material e Métodos

A base de dados utilizada para predição do conjunto de proteínas secretadas pela via secretora foi o banco de dados RefSeq, com o registro de 38432 (setembro, 2008, *H. sapiens*) sequências de proteína que são caracterizadas pelo NCBI (`ftp://ftp.ncbi.nih.gov/refseq/H\_sapiens/mRNAProt/human.protein.faa.gz`).

2.1 Predição de Peptídeo Sinal por SignalP-NN

No algoritmo do *software* SignalP-NN, a saída compreende das seguintes pontuações para previsão: *Max. C*, *Max. S*, *Max. Y*, e dos valore numéricos: *S-mean* e *D*. As duas diferentes redes neurais foram treinadas e são executadas, uma para predizer o peptídeo sinal (*S-score*) e a outra para prever o sítio de clivagem (*C-score*). A pontuação *S-score* é calculada para previsão do peptídeo sinal. A pontuação *C-score* consiste no cálculo do sítio de clivagem, onde para cada posição apresentada na sequência, um *C-score* é relatado, mas o verdadeiro sítio de clivagem tem que ser significantemente alto (HON *et al.*, 2009).

A média *S-mean* é uma média calculada do *S-score*, que possui o parâmetro de cálculo iniciando no primeiro aminoácido N-terminal até o aminoácido com a maior pontuação *Y-max*, calculando então o *S-mean score* com o tamanho do peptídeo sinal predito (HON *et al.*, 2009).

O valor numérico de *D* é uma média simples calculada das pontuações *S-mean* e *Y-max*, e o valor alto demonstra o desempenho em discriminar entre proteínas secretadas e não secretadas (HON *et al.*, 2009).

Para proteínas não secretadas todas as pontuações representadas na saída por SignalP-NN são muito baixas. As duas redes são necessárias para resolver o problema de predição do peptídeo sinal na sequência, pois uma foi treinada para reconhecer o sítio de clivagem e a outra foi treinada para definir se um determinado aminoácido pertence ao peptídeo sinal. A descrição de como é a topologia e treinamento de uma rede neural, foi apresentada na

introdução deste trabalho. Mas a rede de SignalP-NN tem sido atualizada e melhorada com introdução de novas funcionalidades (BENDTSEN *et al.*, 2004).

2.2 Predição de Peptídeo Sinal por SignalP-HMM

O *software* SignalP-HMM usa o modelo escondido de Markov de Nielsen e Krogh (1998) para predição de peptídeo sinal. A descrição resumida de como trabalha o modelo matemático oculto de Markov encontra-se na introdução deste trabalho e em SignalP-HMM, é construído de um diagrama de estados com quatro tipos de proteínas estados: peptídeo sinal, sinal âncora, citoplasmáticas e (para eucariotos) nuclear (NIELSEN; KROGH, 1998).

A ideia no tamanho da distribuição do modelo de Markov de SignalP, foi atribuída a tentativas de encontrar peptídeo sinal: N-região, H-região e C-região, utilizando o algoritmo de Durbin *et al.* (1998).

A estrutura do diagrama de estados de SignalP-HMM permite a estimativa direta do parâmetro do modelo a partir da observação, usando o algoritmo Baum-Welch (RABINER *et al.*, 1989) com o procedimento de máxima verossimilhança (MYUNG, 2003), que analisa o comportamento dos objetos da classe, mostra os estados que os objetos podem assumir, os eventos das transições de um estado para outro e exibe as ações dos eventos.

A C-região tem uma distribuição específica para capturar o padrão do aminoácido antes do sítio de clivagem definido pelo algoritmo de Viterbi (RABINER *et al.*, 1989), com as transições dos estados peptídeo sinal: N-região, H-região e C-região (CHOU; KENDALL, 1990).

Todas as pontuações devem ser altas, ao encontrar o peptídeo sinal, com exceção da probabilidade de sinal âncora. Além disso, o sítio de clivagem é atribuído por uma probabilidade de pontuação, juntamente com pontuações para a N-região, H-região, e C-região do peptídeo sinal, se for encontrado (NIELSEN; KROGH, 1998).

2.3 Predição de Peptídeo Sinal e Topologia Transmembrana por Phobius

O *software* Phobius consiste de um algoritmo desenvolvido na linguagem de programação PERL (*Practical Extraction And Report Language*) com modelos ocultos de Mar-

kov, que combina a predição de peptídeo sinal e predição da topologia transmembrana (SignalP-HMM e TMHMM) (KROGH *et al.*, 2001).

O sub modelo Phobius peptídeo sinal é dividido nas regiões N, H, C e pós-clivagem. Desta forma Phobius faz um relatório de saída do diagrama de estados que consiste nos seguintes dados: região com peptídeo sinal (N-região, H-região, C-região), região não citoplasmática, região transmembrana, região citoplasmática (KÄLL; KROGH; SONNHAMMER, 2004).

Nas sequência cuja região peptídeo sinal não é encontrada, o modelo de Markov Phobius gera um relatório das demais regiões encontradas.

2.4 Pipeline de Predição de Peptídeo Sinal

Para predição do peptídeo sinal pela análise da sequência N-terminal da proteína os programas de predição utilizados neste trabalho foram SignalP version 3.0 (BENDTSEN *et al.*, 2004) e Phobius (KÄLL; KROGH; SONNHAMMER, 2004) descritos nas seções anteriores.

Os *softwares* foram obtidos com os autores por endereço eletrônico, instalados e executados localmente (`http://www.cbs.dtu.dk/services/SignalP` e `http://servers.binf.ku.dk/phobius/`). As sequências foram submetidas em formato FASTA, para análise de predição de peptídeo sinal por SignalP-NN e SignalP-HMM e Phobius. Somente os primeiros 70 aminoácidos foram submetidos em SignalP3.0, pois os autores recomendam como *input* somente a região N-terminal da proteína, ou seja, não mais que 50-70 aminoácidos. As opções de parâmetro de entrada em SignalP NN e HMM, foram para dados de eucarioto, sem gráficos e para Phobius *short output format* e *default*, pois neste formato as ferramentas fornecem listas de previsão da localização peptídeo sinal, hélice tramembrana e localização do intervalo de regiões com *loop* (SONNHAMMER; von Heijne; KROGH, 1998; KROGH *et al.*, 2001; BAIROCH; APWEILER, 1997; HON *et al.*, 2009).

O *pipeline* computacional para predição do conjunto de proteínas secretadas pela via secretora admitiu como entrada a base de dados RefSeq, com o registro de 38432 sequências de proteínas. As sequências de proteína para análise *in silico* e os *softwares* utilizados, SiganlP 3.0 e Phobius, foram baixados localmente, instalados na plataforma, cuja configuração consiste de 2 x CPU Alpha, modelo EV67, 3 GB de memória RAM e 4 discos de 36.4 GB de HD (*Hard Disk*). As sequências foram executadas por programas implementados em PERL via *softwares* e analisadas por programas PERL, levando em

torno de 2 à 3 horas para cada programa com filtragem dos dados. Esse período de tempo varia, de acordo com a quantidade de sequências submetidas. As linguagens PERL-CGI (*Practical Extraction And Report Language - Common Gateway Interface*), PHP (*Hypertext Preprocessor*) e HTML (*HyperText Markup Language*) foram utilizadas para proporcionar a interface entre o banco de dados MySQL (*Structured Query Language*) e usuário.

Os *softwares* utilizados para localização do peptídeo sinal, como já descritos anteriormente, foram SignalP para dados de eucarioto: SignalP-NN, que discrimina peptídeo sinal e não peptídeo sinal, SignalP-HMM que discrimina secretora e não secretora, e Phobius que discrimina peptídeo sinal. Somente proteínas com peptídeo sinal localizados pelos 3 *softwares* foram consideradas como candidatas iniciais a proteínas secretadas.

Para mineração dos resultados de proteínas com peptídeo sinal, foi implementado programas na linguagem de programação PERL. O critério do *script* PERL para a seleção das candidatas a proteínas secretadas dos resultados do algoritmo de SignalP-NN foi selecionar as proteínas cujos valores de medida *Max. S* tivessem *position*, *value* e *cutoff* para *signal peptide (YES)*.

Na seleção das candidatas com peptídeo sinal, dos resultados do algoritmo de SignalP-HMM o critério do *script* PERL foi separar somente sequências que tivessem predição *Prediction: signal peptide*, *signal peptide probability* (valor alto), *signal anchor probability* (valor baixo) e *cleavage site*, com posições definidas.

Os dados de sequências candidatas com peptídeo sinal, nos resultados do algoritmo Phobius foram também selecionados por um *script* PERL utilizando como critério a palavra chave *FT SIGNAL* no relatório de saída, *N-REGION*, *H-REGION* e *C-REGION*, pois as demais localizações do relatório do *software* na sequência não eram parâmetro para o filtro.

Todos os resultados foram separados e organizados em uma base de dados de informação para visualização das candidatas a proteínas secretadas. O sistema de gerenciamento de banco de dados (SGBD) utilizado foi MySQL, que utiliza a linguagem SQL (*Structured Query Language*) como *interface*, é *open source* e possui consistência e confiabilidade para gerenciar os dados (Figura 11).

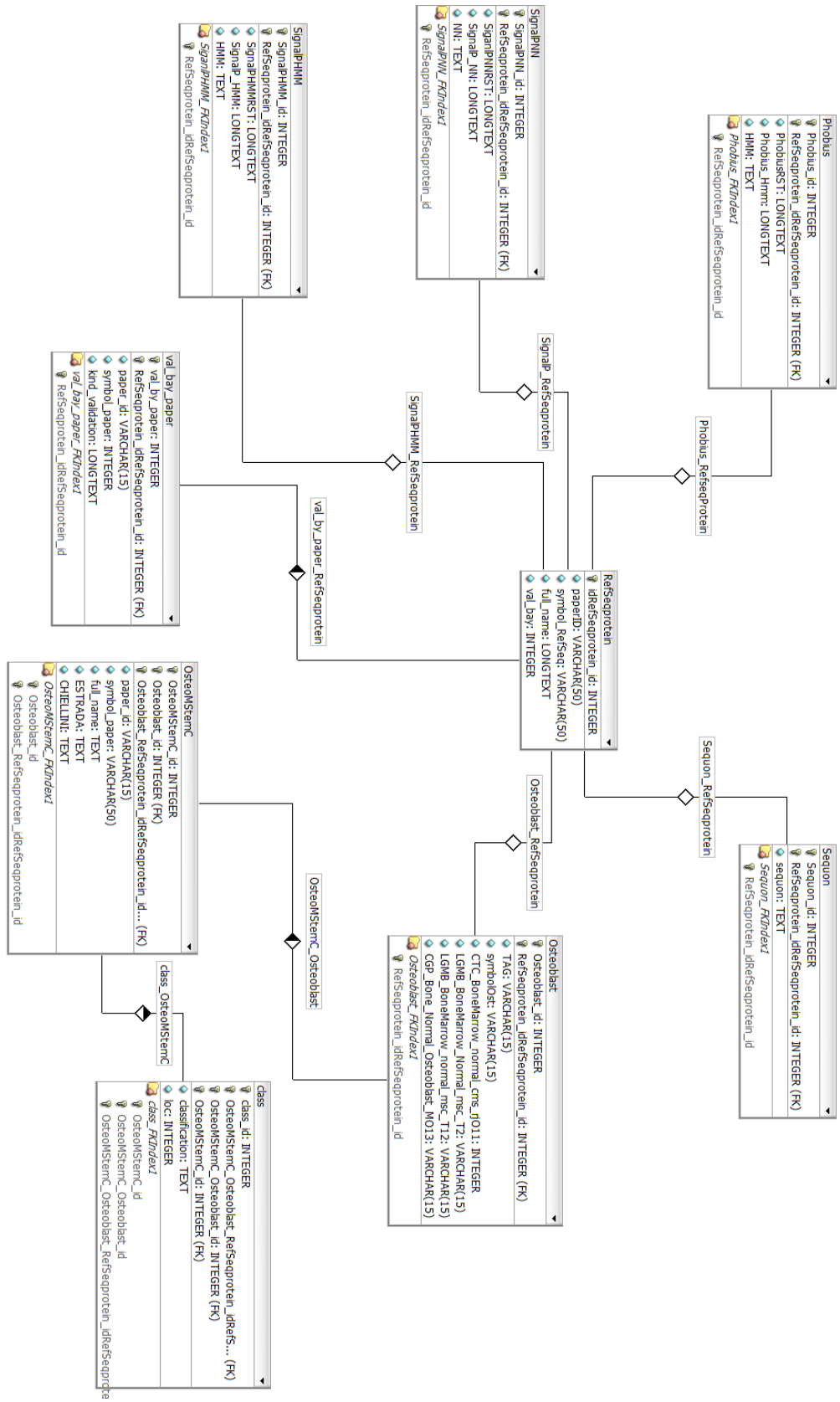


Figura 11: Banco de dados implementado para utilização neste projeto e para armazenamento dos dados do Secretoma Humano.

Para visualizar os dados durante o processamento e ao longo do projeto a linguagem de programação utilizada para gerar a visualização *web* foi PHP. Para relacionar as proteínas com peptídeo sinal encontrados nos 3 *softwares* utilizados, um programa na linguagem PERL, FindSequon.pl, foi implementado para localização de *sequons* ao longo de cada sequência candidata pelos *softwares* anteriormente descritos. O algoritmo de programação PERL, FindSequon.pl, consistiu em um Procedimento (sub-rotina), que percorria cada sequência buscando todo resíduo inicial Asparagina (*N*) e em seguida observando se o próximo resíduo seria Prolina (*P*), pois como descrito na literatura, (*P*), nesta posição inibe a glicosilação. Se encontrada Prolina o procedimento procura uma nova Asparagina (*N*) na sequência e caso não encontrando Prolina (*P*) o próximo passo era procurar os resíduos de glicosilação Serina (*S*), Treonina (*T*) ou Cisteína (*C*), para marcar o motivos *N-X-S, N-X-T, N-X-C* encontrados (Figura 12), ao longo de toda sequência de proteína (GAVEL; von Heijne, 1990).

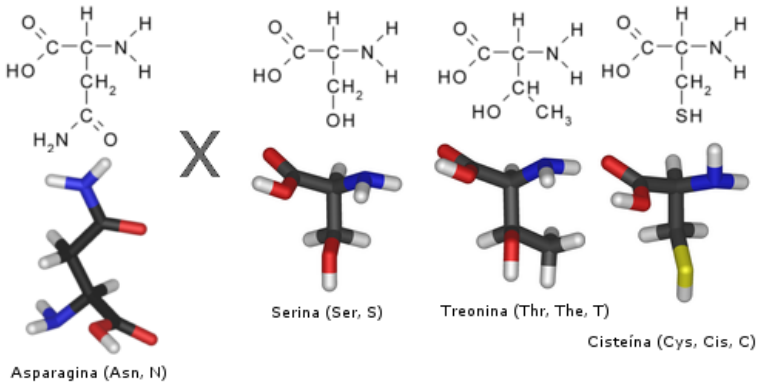


Figura 12: Estrutura linear e tridimensional dos aminoácidos, Asparagina, Serina, Treonina, Cisteína. A posição X na figura indica qualquer aminoácido (exceto Prolina), na sequência tri-petídica *sequon*.

Este esquema de classificação permitiu a identificação de proteínas com características de fortes candidatas a entrar na via secretora pelo RER. A identificação deste conjunto de proteínas em seguida foi combinada por um programa PERL a dados de proteínas secretadas descritas na literatura - validação *in silico* (Figura 13).

As proteínas por esta estratégia classificadas como candidatas a secretadas foram relacionadas a dados de expressão gênica diferencial da técnica SAGE (HU; POLYAK, 2006) (*Serial analysis of gene expression*) do tecido Osteoblasto (BRUSIN, 2008; SILLENCE; SENN; DANKS, 1979). Os dados do tecido Osteoblasto (técnica SAGE), no tecido Mesenquimal tempo 0hs, Mesenquimal tempo 2hs, Mesenquimal tempo 12hs e Osteoblasto tempo 21hs, foram relacionados aos dados das proteínas com características de secretadas deste trabalho e que estão descritas na literatura.

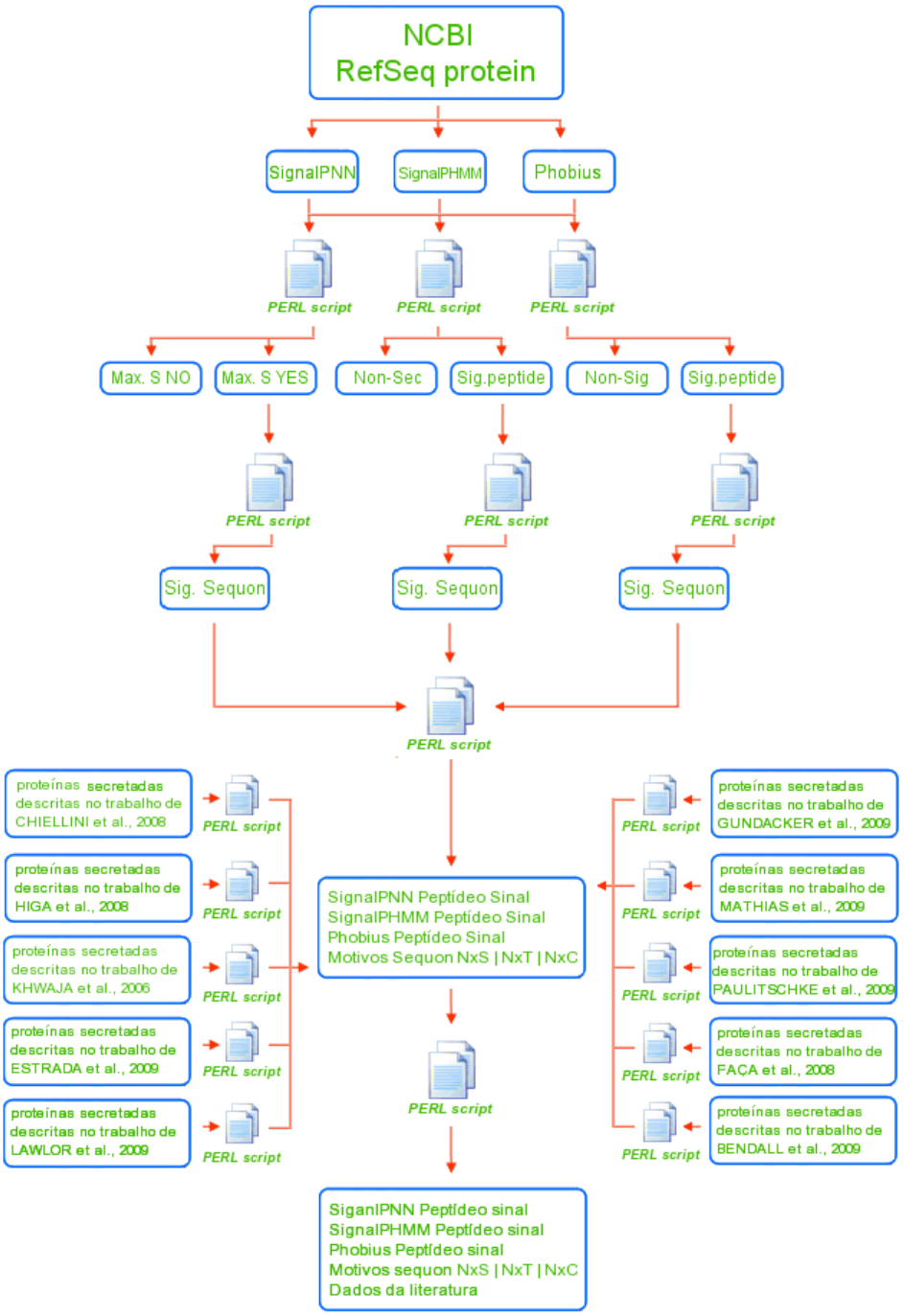


Figura 13: *Pipeline* da análise *in silico* para identificar candidatas a proteínas secretadas no Secretoma Humano.

2.5 Análise Experimental de Secretoma

2.5.1 Cultura de Células e Obtenção das Proteínas Secretadas

Células da linhagem de câncer da mama $HCC1954$ foram cultivadas em meio D-MEM suplementado com 10% de soro bovino fetal até que as células atingissem 90% em garrafas de cultura de $75mm_2$ de confluência. Após atingirem a confluência, as células foram lavadas por 3 vezes com PBS (tampão fosfato salina) $1x$ para a retirada de todo o meio de cultura suplementado com soro bovino fetal.

Foi, então, adicionado meio de cultura D-MEM sem qualquer suplementação e as células foram mantidas em cultura por 48 horas, a 37^oC em atmosfera umidificada contendo 5% de CO_2. Um volume final de $3ml$ em sistema AMICON de diálise, utilizando membrana de 10 kDa. O meio de cultura sem as células foi removido e concentrado até um determinado volume. As proteínas foram quantificadas pelo método colorimétrico de BRADFORD cuja concentração total foi de $1,6mg/ml$.

Alíquotas deste concentrado foram submetidas à eletroforese em gel de poliacrilamida contendo SDS (sulfato de sódio), como será descrito a seguir.

2.5.2 SDS-PAGE

Foi utilizado extratos da linhagem celular $HCC1954$, que foram submetidas a prévia separação por SDS-PAGE, $20\mu g$ de proteínas da amostra e gel de poliacrilamida 12,5% homogêneo.

$10cmX10cmX1mm$ e $16cmX18cmX1,5cm$

em sistemas de eletroforese vertical para géis de poliacrilamida (*GE Healthcare*). O gel foi submetido as seguintes condições de corrida: $300V$, $15mA$ constante. As proteínas foram detectadas pelo corante *Coomassie Blue Coloidal* G-250. As imagens foram adquiridas pelo *scanner ImageScanner* utilizando o *software* MagicScan (GE Healthcare). O gel foi dividido em frações que em seguida foram recortadas do próprio gel. Foi submetido a digestão e após a digestão, as frações que apresentavam duplicidade, foram passadas para um único tubo $a+b=1$, $c+d=2$, $e+f=3$, $g+h=4$, $i+j=5$, $l+m=6$, $n+o=7$.

2.5.3 Digestão Enzimática e Análise por Espectrometria de Massas

As bandas de interesse no gel foram selecionadas e recortadas para identificação. As bandas recortadas foram lavadas com solução NH_4HCO_3 (bicarbonato de amônio) $50mM$ e acetonitrila 50% para a remoção de SDS e corante, sendo em seguida adicionado acetronitrila. As bandas completamente secas em *SpeedVac*, foram reidratadas com uma solução contendo $0,5\mu g$ de tripsina em solução NH_4HCO_3 $50mM$. A reação foi mantida a 37^oC por 24 horas e interrompida pela adição de $5\mu L$ de ácido fórmico.

2.5.4 Caracterização de Proteínas por *Shotgun Peptide Sequencing*

Os peptídeos tripsínicos obtidos na seção anterior foram extraídos com $100\mu L$ de solução contendo 50% de acetonitrila e $0,1\%$ de ácido fórmico, seguidos de secagem completa em *SpeedVac*. A amostra foi resuspendida em $10\mu L$ de solução contendo 30% de acetronitrila e $0,1\%$ de ácido fórmico, seguidos de adição de $50\mu L$ de ácido fórmico $0,1\%$. A amostra foi centrifugada a $12000rpm$ por 3 minutos e foi aplicada em um nano-HPLC (*Shimadzu*), com uma coluna de fase-reversa. Em cada fração do gel foram coletados 2.5 *spots* (cada um com 0,8 μL do eluente) em sistema *Axima Shimadzu*. As amostras foram analisadas por MALDI-TOF-TOF (*Axima Performance - Kratos - Shimadzu, Manchester, UK*).

3 Resultados e discussão

3.1 Implementação e Mineração do Banco de dados

A via secretora em células de mamíferos é constituída por um conjunto de membranas, organelas transportadoras pelas quais proteínas secretadas se movem para atingir diferentes destinos. Este caminho significa que operações bioquímicas ocorrem pela via, como enovelamento, glicosilação ou biossíntese de lipídios que podem ser compartimentadas o que permite reações específicas. Dado a esta ligação entre localização subcelular e função, um mapa da distribuição quantitativa de todas as proteínas e lipídios constituintes da secreção ("secretome") é um passo importante para a compreensão da forma como ocorrem essas funções (SIMPSON; MATEOS; PEPPERKOK, 2007), por isso a entrada da pré-proteína na via secretora nos proporciona características específicas, o que nos permitiu montar a arquitetura do modelo *data mining*, considerada neste trabalho com a combinação dos algoritmos SignalP-NN, SignalP-HMM, Phobius, para busca de peptídeo sinal, como está representado na figura 14.

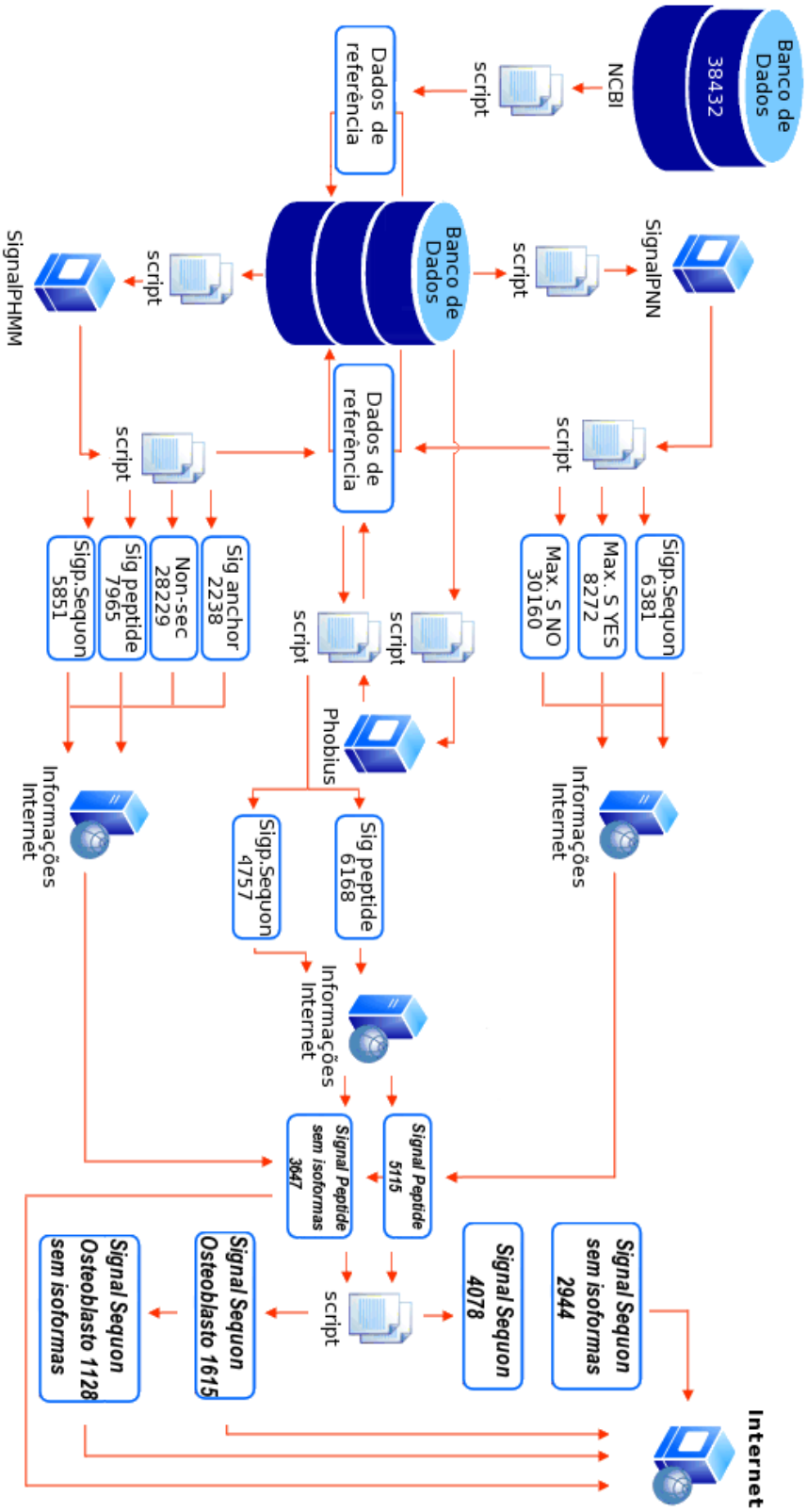


Figura 14: *Flowchart* de mineração dos dados obtidos na estratégia de identificação de proteínas com características de secretadas no Secretoma Humano.

O conjunto de dados do *pipeline* de busca de sequências proteicas secretadas em humanos iniciou com a base de dados RefSeq, com o registro de 38432 sequências com isoformas e 24244 sequências sem isoformas. Após a execução dos *softwares* em cada proteína da base de dados, várias observações visuais e computacionais, foram feitas no modelo, e a predição combinada entre os *softwares* produziu a relação entre os conjuntos de classificação das candidatas a proteínas secretadas em cada *software*, com o algoritmo PERL, FindSequon.pl que buscou os motivos de glicosilação *sequons* ao longo da sequência.

Utilizando o diagrama de Venn (GRUNBAUM, 1975) podemos observar o valor total de representação de conjunto único de proteínas com peptídeo sinal e *sequon*, com o objetivo de mostrar os valores nos conjuntos numéricos, organizando e analisando os resultados (Figura 15).

Figura 15: Este diagrama apresenta o conjunto de valor total de classificação de proteínas candidatas a secretadas (sem isoformas), por possuir *sequon* e peptídeo sinal nos *softwares* SignalP-NN, SignalP-HMM e Phobius.

Podemos observar utilizando o diagrama de Venn (GRUNBAUM, 1975), a relação entre os três conjuntos de resultados dos *softwares* SignalP-NN, SignalP-HMM e Phobius, as relações de união e intersecção entre os resultados numéricos, com o intuito de estabelecer uma melhor compreensão dos resultados das proteínas candidatas a secretadas pertencentes a cada conjunto (Figura 16).

O conjunto total de sequências de proteína submetidas as análises foram as 38432 sequências com isoformas da base de dados RefSeq, porém observamos que usando as isoformas haveria apenas redundância nos resultados por isso optamos em excluir as isoformas da análise, o que nos proporcionou o total de 24244 sequências de proteína submetidas a análise dos *softwares* (Figura 17). Do total de 24244 sequências de proteína analisadas pelos *softwares* e pelo algoritmo PERL, FindSequon.pl, que buscou os motivos de glicosilação, 4677 possuem motivo *sequon* e peptídeo sinal, de acordo com o *software* SignalP-NN, 4170 possuem motivo *sequon* e peptídeo sinal, de acordo com o *software* SignalP-HMM e 3438 possuem motivo *sequon* e peptídeo sinal, de acordo com o *software*

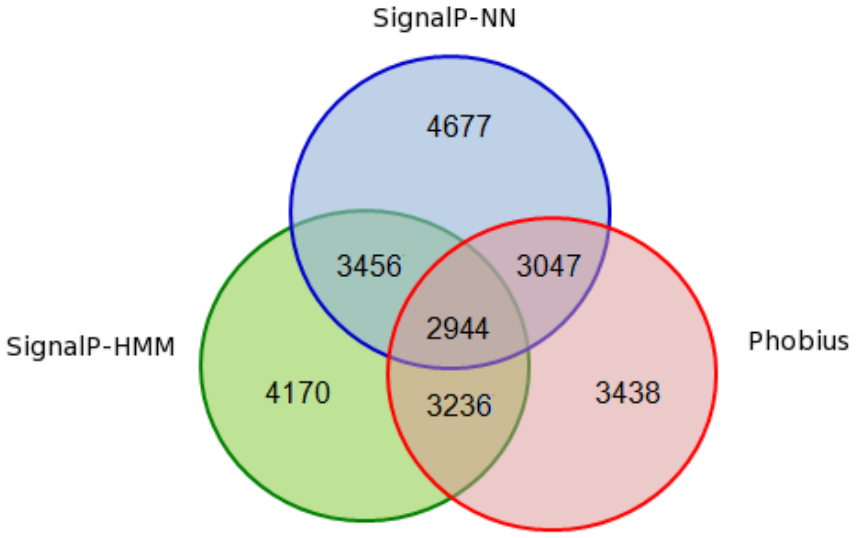


Figura 16: Diagrama de intersecção de proteínas candidatas (sem isoformas), entre os *softwares* SignalP-NN, SignalP-HMM e Phobius.

Phobius. Por esta aproximação computacional, obtivemos a intersecção combinada de 2944 sequências de proteína candidatas a secretadas da célula, por possuírem motivos de glicosilação *sequon* e peptídeo sinal.

Os dados das proteínas classificadas como candidatas a secretadas foram também relacionados aos dados de expressão gênica diferencial da técnica SAGE do tecido Osteoblasto, nos tempos Mesenquimal tempo 0hs, Mesenquimal tempo 2hs, Mesenquimal tempo 12hs e Osteoblasto tempo 21hs e esta correlação proporcionou o total de 1128 proteínas em Osteoblasto classificadas como candidatas a secretadas (Figura 17) e 49 validadas *in silico* pelos trabalhos de Chiellini *et al.* (2008) e Estrada *et al.* (2009).

	SignalPNN	SignalPHMM	Phobius	SignalP&Phobius	Osteoblasto
TOTAL	38432				1355
N-terminal Peptideo Sinal (isoformas)	8272	7965	6168	5115	1909
Peptideo Sinal (isoformas) *Sequon*	6381	5851	4756	4078	1615
TOTAL	24244				1355
Peptideo Sinal	5899	5508	4363	3647	1333
Peptideo Sinal *Sequon*	4677	4170	3438	2944	1128

Figura 17: Os dados de sequência com isoformas e sem isoformas são mostrados na figura e a quantidade de proteínas separadas por possuírem *sequon* e peptídeo sinal.

3.2 Validação *in silico* das Proteína Candidatas a Secretadas

O passo sequencial em nosso *pipeline* do secretoma humano, foi obter informações na literatura a respeito das 2944 sequências de proteína candidatas. Com esta busca encontramos 695 proteínas descritas como secretadas por experimentos feitos em trabalhos descritos na literatura (Figura 18).

	SignalPNN	SignalPHMM	Phobius	SignalP&Phobius	Osteoblasto
TOTAL	24244				1355
Peptideo Sinal	5899	5508	4363	3647	1333
Peptideo Sinal ***Sequon***	4677	4170	3438	**2944**	**1128**
Validação *in silico*				**695**	**49**
Publicações para validação *in silico*	CHIELLINI et al., 2008; HIGA et al., 2008; KHWAJA et al., 2006; ESTRADA et al., 2009; LAWLOR et al., 2009; GUNDACKER et al., 2009; MATHIAS et al., 2009; PAULITSCHKE et al., 2009; BENDALL et al., 2009; FAÇA et al., 2008;				CHIELLINI et al., 2008; ESTRADA et al., 2009

Figura 18: Os dados na figura mostram a quantidade de sequências candidatas e as validadas *in silico* do Secretoma Humano e também no tecido Osteoblasto.

Um dos trabalhos utilizados para validação dos dados deste *pipeline in silico* foi o de Chiellini *et al.* (2008). Nesse trabalho eles, estabeleceram que o tecido adiposo não só possui um papel fundamental na conservação e liberação de energia, mas também é um órgão secretor e fonte de células tronco. Entre as diferentes linhagens, as células tronco são capazes de se diferenciar em osteoblastos e adipócitos. Como proteínas secretadas podem regular o equilíbrio entre ambas as linhagens, Chiellini *et al.* (2008), tiveram por objetivo em seu trabalho, caracterizar o secretoma do tecido adiposo humano multipotente derivado de células tronco (hMADS) com a iniciativa de relacionar adipócitos e osteoblastos. A análise da diferenciação de células hMADS, em adipócitos e osteoblastos e coloração *Coomassie Blue staining of hMADS cell secretome* em conjunto com análise qRT-PCR dos níveis de mRNA específicos de adiponectina e osteogênico (marcadores de fosfatase alcalina) nos dias 3 e 14 de diferenciação em relação ao dia 0 demonstraram a Chiellini *et al.* (2008), dados dos três experimentos independentes com microfotografias de hMADS de células diferenciadas em adipócitos e osteoblastos no dia 14. Esta análise representou a escala em gel de proteínas secretadas de células hMADS no dia 0 e dia 3 para diferenciar adipócitos e osteoblastos após 6hs de incubação. O gel foi representativo nos 3 experimentos independentes. Esta abordagem proteômica, utilizando

eletroforese mono-dimensional e espectrometria de massa em tandem, permitiu a Chiellini *et al.* (2008), identificar a diferenciação de um total de 73 proteínas secretadas nos dias 0 e 3 de adipócitos e osteoblastos. A análise das proteínas identificadas mostrou que 52% corresponderam a proteínas classicamente secretadas caracterizadas por um peptídeo sinal, que 37% estão previamente descritas como proteínas do compartimento extracelular e são desprovidas de peptídeo sinal, e que 11% não possuem peptídeo sinal, nem estão descritas extracelular.

Desta abordagem proteômica, obtivemos em nosso trabalho a validação *in silico* de 36 proteínas secretadas, do total das 2944 candidatas com características *sequon* e peptídeo sinal (Figura 19).

Peptídeo sinal *Sequon*	Dados da literatura para validação *in silico*	Total	Osteoblasto	Total
	CHIELLINI et al., 2008;	36	CHIELLINI et al., 2008; ESTRADA et al., 2009;	49
	HIGA et al., 2008;	46		
	KHWAJA et al., 2006;	44		
	ESTRADA et al., 2009;	31		
	LAWLOR et al., 2009;	44		
	GUNDACKER et al., 2009;	66		
	MATHIAS et al., 2009;	16		
	PAULITSCHKE et al., 2009;	25		
	BENDALL et al., 2009;	104		
	FAÇA et al., 2008;	610		
Total 2944		695		

Figura 19: Os dados na figura mostram a intersecção a respeito das 2944 candidatas ao Secretoma Humano, as 695 validadas *in silico* e a quantidade em cada trabalho descrito na literatura.

Outro trabalho para validação dos dados do *pipeline in silico* foi o de Higa *et al.* (2008), onde eles, usaram a análise proteômica da linhagem de células hepáticas infectadas HepG2, com DV, focando o alvo de proteínas secretadas da células. A abordagem proteômica utilizou gel *1D-electrophoresis* e *liquid chromatography coupled* com *tandem mass spectrometry (LC-MS/MS)*. Os resultados das análises desta estratégia permitiu a Higa *et al.* (2008), a identificação de um total de 107 proteínas, onde em cada 35 encontradas, 24 eram encontradas no controle do secretoma e 24 somente em células do secretoma

infectadas. Para validação desses dados Higa *et al.* (2008), realizaram *2D-eletrophoresis* seguido de *MALDI-TOF/TOF*, resultando na identificação de 20 proteínas. Destas, 8 foram confirmadas por resultados de *LC-MS/MS*. Correlacionando estes dados aos nossos 2944 *in silico* e encontramos 46 proteínas secretadas, com características *sequon* e peptídeo sinal (Figura 19).

No trabalho por Khwaja *et al.* (2006), também utilizado para validação dos dados deste *pipeline in silico*, eles observaram que interações em tumor de estroma desempenham um papel importante na evolução do desenvolvimento tumoral, em sua manutenção e em sua progressão. Porém Khwaja *et al.* (2006), salientam que pouco se sabe sobre a forma como alterações genéticas de transformação da célula, provocam atípicas interações celulares que modulam estas alterações intercelular complementares. A hipótese é de que estes eventos envolvem modificações como complemento de proteínas secretadas pelas células, que agem como mediadoras de comunicação intercelular. Para testar esta hipótese Khwaja *et al.* (2006), examinaram o papel que estas proteínas desempenham na célula em um supressor tumoral importante. Esta análise foi feita por um micro-ambiente tumoral e pela regulação dos fatores de secreção. A abordagem usou uma combinação de 2-D e técnicas cICAT proteômica, onde Khwaja *et al.* (2006), encontraram 111 proteínas secretadas, 39 apresentando secreção elevada e 21 secreção inibida em resposta à expressão do supressor de tumor wt-p53. Do trabalho por Khwaja *et al.* (2006), obtivemos a validação *in silico* de 44 proteínas secretadas, das 2944 candidatas a secretadas da célula com características *sequon* e peptídeo sinal (Figura 19).

Outro trabalho utilizado para validação dos dados do *pipeline in silico*, foi de Estrada *et al.* (2009), que observaram que a medula óssea derivada de células tronco mesenquimal (MSCs), estão envolvidas em respostas a cicatrização e regeneração. Estrada *et al.* (2009), traçam o perfil global do proteoma MSCs para investigar o fator crítico que pode promover a cicatrização e regeneração de tecidos. Estrada *et al.* (2009), observaram que proteínas Cyr61 mostraram-se presentes abundantemente em MSCs. A presença de Cyr61 foi confirmada por coloração imunofluorescência *staining* e análises por *imunoblot*. Desta forma Estrada *et al.* (2009), demonstraram que Cyr61 está presente no meio de cultura (secretoma) de MSCs. O secretoma de MSCs estimula respostas angiogênicas de neovascularização *in vitro* e *in vivo*. A diminuição de Cyr61 completamente revoga a indução angiogênica na capacidade do secretoma de MSCs. Estrada *et al.* (2009), observam que coletivamente todos os dados demonstram que polipeptídeos no secretoma de MSC, contribuem para promoção da atividade de angiogênese, um evento chave necessário para a regeneração e reparação de tecidos lesados. Relacionamos nossos dados *in silico*, com

dados do trabalho de Estrada *et al.* (2009), e encontramos 31 proteínas secretadas em nosso banco de proteínas das 2944 com características *sequon* e peptídeo sinal (Figura 19).

Em outro trabalho por Lawlor *et al.* (2009), obtivemos 44 proteínas secretadas relacionadas as 2944 com características *sequon* e peptídeo sinal (Figura 19), de nossos dados *in silico*. No trabalho eles buscam um meio eficiente para a identificação de bio-marcadores prognósticos e preditivos que atualmente são essenciais em estudos com câncer. A abordagem para esta descoberta por Lawlor *et al.* (2009), foi proposta procurando por vias de proteínas individuais em vez de monitoração e alvo. Atualmente, o "secretome", um fluído biológico que geralmente é enriquecido com proteínas secretadas e ou adjacentes pode ser relevante em células cancerosas e por isso tem sido alvo de descobertas de bio-marcadores. No trabalho de Lawlor *et al.* (2009), é descrito um novo método de análise utilizando análises *stacking gels*, *label-free relative quantitation*, e análises de via metabólica. O protocolo apresentado, aumentou a capacidade de análise do secretoma em aproximadamente 1 ordem de magnitude, em comparação a metodologias anteriores, segundo Lawlor *et al.* (2009). O secretoma global foi o conjunto de dados obtidos e analisados verificando vias metabólica com *softwares*, integrando os resultados experimentais a um contexto de sinalização celular. Lawlor *et al.* (2009), sugerem que vários secretomas possam ser interligados por vias metabólicas intercelulares, e por sua vez podem eventualmente permitir a utilização de secretomas para descoberta de bio-marcadores. Quando Lawlor *et al.* (2009), aplicaram esta estratégia a duas linhagens celular de câncer de mama, verificou-se que o sinalizador IGF e os sistemas ativadores plasminogênios podem ser regulados de maneira diferente em câncer de mama invasivo, mas continua especulativo até que seja verificado clinicamente. Em resumo, a metodologia proposta otimiza a cultura celular com fracionamento e LC-MS para obter resultados mais elevados a partir de células cultivadas de secretoma, em bio-marcador por ligação putativa com câncer e segundo Lawlor *et al.* (2009), um meio eficaz para a identificação de bio-marcadores de prognósticos e predição hoje, são essenciais no tratamento do câncer.

Outro trabalho para validação *in silico*, nesta abordagem é o de Gundacker *et al.* (2009), onde obtivemos 66 proteínas secretadas relacionadas as 2944, com características *sequon* e peptídeo sinal (Figura 19). Gundacker *et al.* (2009), selecionam para análise de secretoma, células dendríticas (DCs) por estas serem células mais potentes e apresentadas em antígeno que desempenham papel fundamental na regulação da imunidade adaptativa. Gundacker *et al.* (2009), geraram imaturas células DCs por cultivo *in vitro* de monócitos do sangue periférico e funcionalmente ativou-as com o clássico padrão associado patogênico

lipossacarídeo (LPS). O objetivo desse estudo foi a identificação de perfis do proteoma relacionados à células dendríticas com fenótipos funcionalmente diferentes. Proteínas citoplasmáticas foram analisadas por *shotgun* proteômico resultando na identificação de 1690 proteínas. Embora a maturidade e alternativamente DCs ativados exibem expressão de perfis altamente distintos na proteína, VFC-tratados mostraram menor alterações no proteoma DCs. Observou na investigação por uma combinação de 2D-PAGE e *shotgun* proteômica que muitos DCs exercem funções específicas, pela da secreção. Gundacker *et al.* (2009), identificaram com sucesso uma ampla variedade de citoquinas. Em seu trabalho Gundacker *et al.* (2009), demonstram que funcionais subconjuntos distintos de DCs exibem perfis de proteoma como candidatos a bio-marcadores. Gundacker *et al.* (2009), afirmam ainda que estas proteínas podem revelar-se úteis para a interpretação da clínica complexa de dados proteicos.

Outro trabalho utilizado para validação dos dados do *pipeline in silico*, foi o trabalho de Mathias *et al.* (2009), cujo estímulo da pesquisa foi a especulação de que a transição epitélio mesenquimal (EMT) é altamente conservada por um processo pelo qual as células epiteliais perdem em sua base características morfológicas, tais como contato célula-célula e ganho de propriedades mesenquimais, tais como o aumento da motilidade e invasividade. Segundo Mathias *et al.* (2009), para obter introspecções, proteínas são liberadas a partir de células que modulam o processo de EMT. Mathias *et al.* (2009), compararam o secretoma de perfis de expressão de proteínas de células MDCK e *Ras-transformed MDCK cells (21D1)*, que se mostraram expressamente estáveis em *oncogenic Ras* usando 2D-DIGE/LC-MS/MS. Mathias *et al.* (2009), compararam os secretomas diferencialmente expressos com respectivos perfis usando a expressão gênica de perfis do sistema *Affymetrix GeneChip* e observaram que proteínas *down-regulated* são predominantemente envolvidas no contato célula-célula e célula-matriz de adesão e proteases, podem ser fatores que promovem migração. Segundo Mathias *et al.* (2009), as proteínas identificadas no secretoma talvez não tenham sido previamente identificadas no contexto EMT, mas podem esclarecer o entendimento de estudos de mecanismos associados e processos celular. Do trabalho de Mathias *et al.* (2009), relacionados a nossos resultados obtivemos a validação *in silico* de 16 proteínas secretadas, das 2944 candidatas com características *sequon* e peptídeo sinal (Figura 19).

Outro trabalho é o de Paulitschke *et al.* (2009), que traça uma nova era buscando a aproximação em detectar bio-marcadores em melanoma. Segundo Paulitschke *et al.* (2009), metástases em melanoma está associado a prognósticos e a detecção precoce e pode melhorar significativamente a sobrevida dos pacientes diagnosticados. Paulitschke

et al. (2009), apresentam em seu trabalho a descoberta de um novo bio-marcador, onde a estratégia foi baseada em perfis e análise de proteoma e secretoma. Paulitschke *et al.* (2009), mencionam que o estroma tumoral associado a proteínas secretadas das células podem atuar como promotores de tumor e esta cooperatividade celular é reversível podendo ser diretamente acessível em intervenções terapêuticas. O aparecimento e o estudo aprofundado destes eventos característicos parecem preceder a progressão tumoral e observa-se que, especificamente proteínas secretadas por essas células podem servir como bio-marcadores na doença precocemente. Dado à natureza dos vasos sanguíneos neoformados e ao aumento da pressão hidrostática (fluidostática) em tumores, as proteínas são secretadas mais facilmente para o sangue. Na análise de Paulitschke *et al.* (2009), a estratégia se resume em três diferentes sistemas modelo, incluindo as linhagens celular cultivadas estabelecidas de sistemas de modelo animal, clínico e amostras em humanos. A viabilidade é demonstrada com perfis de secretoma e proteoma gerados a partir de fibroblastos da pele humana normal comparados com fibroblastos de melanoma-associado, isolados de xenotransplantes de camundongo e fibroblastos de medula óssea de vários pacientes com mieloma. Paulitschke *et al.* (2009), fizeram ainda outras comparações incluindo perfis mútuos de proteoma de melanócitos e células de melanoma M24met. Todos dados de Paulitschke *et al.* (2009) estão disponíveis na literatura e inclusive a lista de secretoma, por material suplementar do banco de dados. O trabalho de Paulitschke *et al.* (2009), abordou uma nova estratégia para identificar marcadores de diagnóstico em proteínas secretadas auxiliando na detecção precoce do melanoma metastático para melhoraria da compreensão e de *pathomechanisms* envolvendo o micro-ambiente, visando permitir a concepção de novas estratégias terapêuticas. Utilizando o material suplementar deste trabalho encontramos 25 proteínas secretadas relacionadas as 2944 com características *sequon* e peptídeo sinal *in silico* (Figura 19).

Em mais um trabalho, por Bendall *et al.* (2009), utilizado para validação *in silico* deste *data mining*, encontramos 104 proteínas descritas como secretadas com peptídeo sinal e *sequon* (Figura 19). O trabalho de Bendall *et al.* (2009), trata-se de uma aproximação de espectrometria de massa com células tronco embrionárias, buscando a combinação de resultados de dois instrumentos para provar a melhoria alcançada por um único instrumento na identificação de proteínas. Bendall *et al.* (2009), iniciaram de cultura e derivação a longo prazo de manutenção de células embrionárias humanas (hESCs) de forma dependente e independente com apoio da célula, mas no entanto os fatores responsáveis pela preservação da viabilidade de hESCs em um estado incipiente permanecem desconhecidos. No trabalho de Bendall *et al.* (2009), o grupo descreve o método baseado em espectrome-

tria de massa para sondar o secretoma da cultura hESC micro-ambiente da proteína que regula a identificação de fatores potenciais que estão em baixa abundância. As amostras individuais foram analisadas várias vezes utilizando sucessivas massa (m/z) e retenção de tempo dirigida, sem a mesma amostragem péptido íon por duas vezes. Na abordagem Bendall *et al.* (2009), excluíram iterativa-espectrometria de massas (IE-MS) e aborda mais que o dobro de proteínas e peptídeos, em comparação ao simples método de análise em repetir o mesmo instrumento, após uma extensa amostra de pré-fracionamento. Além disso, a aplicação da abordagem IE-MS por Bendall *et al.* (2009), segundo o seu trabalho, mostrou melhor desempenho aos antigos (Q-TOF) MS. O número resultante de peptídeos identificados aproximou-se ao de uma análise em paralelo de repetir uma nova LTQ-Orbitrap MS. Na combinação dos resultados de ambos os instrumentos, Bendall *et al.* (2009), provaram que a abordagem é superior a alcançada por um único instrumento na identificação de outras proteínas. Usando a estratégia IE-MS, combinada com gel complementar e solução à base de métodos de fracionamento. O micro-ambiente hESC em cultura foi amplamente analisado e mais de 10 a 12 proteínas extracelulares foram observadas, em comparação a estudos publicados anteriormente. O objetivo do trabalho de Bendall *et al.* (2009), foi a detecção dos anteriormente indetectáveis fatores de crescimento, presentes em concentrações variando de 10-9 a 10-11 g/ml, e destacar a profundidade do perfil, com o IE-MS oferecendo uma abordagem simples e de técnica confiável aumentando consideravelmente o desempenho do instrumento pelo aumento da profundidade efetiva do *MS-based proteomic profiling*. Bendall *et al.* (2009), afirmam que esta abordagem deve ser amplamente aplicável a qualquer plataforma LC-MS/MS ou sistema biológico.

Em outro trabalho utilizado para nossa abordagem, Faca *et al.* (2008), buscaram elucidar proteínas secretadas e de superfície celular de células tumorais que são relevantes para diagnósticos moleculares, imagens de tumor e terapias. No trabalho de Faca *et al.* (2008), o grupo visou caracterizar o proteoma de superfície celular e as proteínas liberadas no meio extracelular de três linhagens celulares de câncer do ovário, CaOV3, OVCAR3 e ES2 e de células tumorais a partir de fluídos ascíticos. Para diferenciar proteínas liberadas no meio a partir de proteínas dos meios de cultura, as células foram cultivadas na presença de [13C]-marcadas com lisina, onde Faca *et al.* (2008), utilizaram a abordagem baseada em biotinilação, para captura da superfície celular das proteínas. A estratégia experimental de Faca *et al.* (2008), consistiu de fracionamento de proteínas seguida por digestão proteolítica e LC-MS/MS. As células de superfície foram fracionadas por cromatografia de fase reversa, digerida com tripsina e analisadas por espectrometria. Cada fração da fase reversa de cromatografia (as células de superfície, o que era do meio e

o que era do extrato), foram digeridas com tripsina, agrupadas em 15 a 21 *pools* para cada linhagem celular e cada compartimento celular. Proteínas e perfis das linhagens celulares tiveram substancial similaridade com os perfis a partir de células de câncer de ovário, fluído ascítico e marcadores de proteína conhecidos (CDH1 e TIMP1 e TIMP2) para serem associados a câncer de ovário. No trabalho de Faca *et al.* (2008), foram observadas das 3 linhagens versus fluídos ascítico 80% de concordância em, com um total de 5704 proteínas nas linhagens celulares (CaOV3, OVCAR3 e ES2) e 3789 nas células de tumor de fluídos ascítico de ovário, com observância de 3022 proteínas secretadas relacionadas nas linagens celulares versus tumor ascítico. A análise proteômica de Faca *et al.* (2008), indicou vários compartilhamentos de domínios extra-celular de proteínas expressas na superfície celular e elevadas taxas de secreção de algumas proteínas. Com esses resultados de proteínas de superfície celular e de proteínas secretadas, Faca *et al.* (2008), possuem a perspectiva de fornecer novos alvos para diagnóstico terapêutico, dado a profundidade do perfil proteômico das células de câncer de ovário. Desta abordagem utilizamos o material suplementar disponível na *internet* que nos proporcionou proteínas secretadas nas 3 linhagens e em fluídos ascíticos, e que correlacionadas *in silico*, nos proporcionou a validação de 610 proteínas secretadas com características peptídeo sinal e *sequon* (Figura 19).

Combinando a aproximação de nossa estratégia com os dados de proteínas secretadas em cada trabalho descrito a cima na literatura, relacionamos as 2944 candidatas que possuem características peptídeo sinal e *sequon* o que nos proporcionou a intersecção (um catálogo) de 695 proteínas secretadas validadas *in silico* pelos experimentos dos trabalhos publicados.

Em organismos multicelulares, proteínas secretadas desempenham papel crucial na regulação comunicação intercelular. Proteínas secretadas podem ser liberadas na corrente sanguínea, produzindo efeitos sistêmicos e por meio desta abordagem computacional, nós conseguimos classificar 3647 proteínas com peptídeo sinal, 2944 proteínas com peptídeo sinal e motivos de glicosilação *sequon* e 695 proteínas com peptídeo sinal e *sequon*, por meio de validação *in silico*, utilizando abordagens experimentais, descritas na literatura como secretadas.

3.3 Validação Experimental das Proteínas Candidatas a Secretoras

O gel SDS-PAGE mostra a existência de várias bandas e para a análise de *shotgun peptide*, o gel foi recortado em 7 frações como está descrito em material e métodos. Em seguida, as frações foram digeridas com tripsina e submetidas à separação dos peptídeos tripsínicos por cromatografia de fase reversa utilizando uma nono-HPLC, onde o eluente foi coletado automaticamente em uma placa de maldi como descrito em material e métodos.

Os *spots* foram submetidos ao espectrômetro de massa Maldi TOF-TOF sendo obtidos espectros ***ms*** e ***ms/ms*** para o sequenciamento de aminoácidos. Os espectros obtidos foram submetidos a anotação utilizando o banco de dados SWISS-PROT com *software Mascot search algorithm Matrix Science* e os parâmetros de busca do *software* para a identificação das proteínas foram com *score* do peptídeo igual ou maior que 20. Com este parâmetro buscando nas 7 frações recortadas do gel (Figura 20), obtivemos a lista de 116 proteínas identificadas em espectrômetro de massa Maldi TOF-TOF *Mascot*.

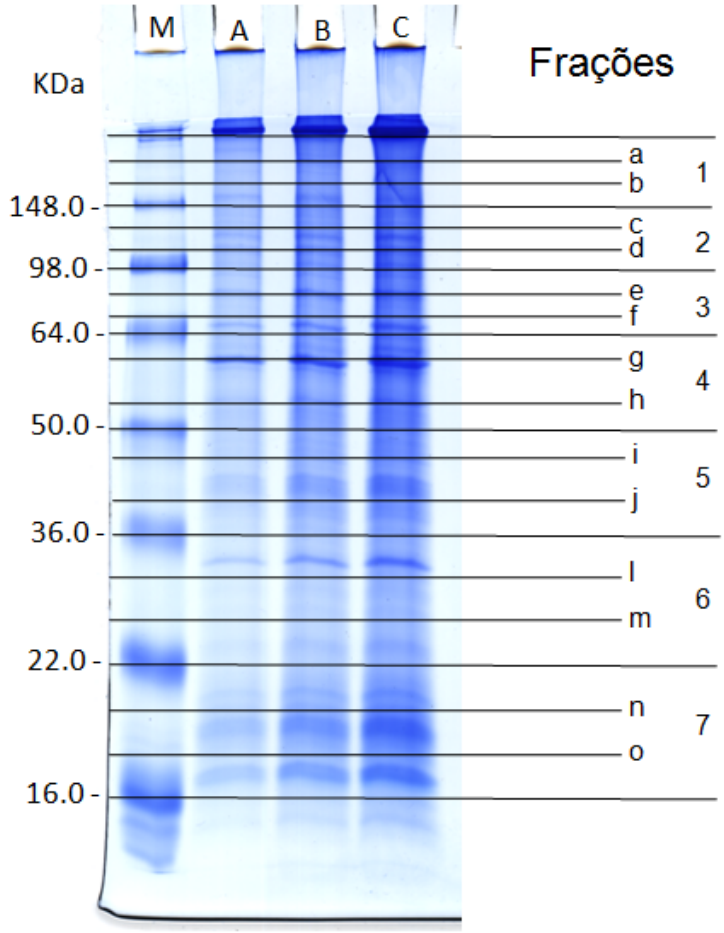


Figura 20: Separação das proteínas secretadas da linhagem *HCC*1954 por SDS-PAGE. Foi aplicado diferentes concentrações de proteínas em gel de poliacrilamida 12,5% e detectadas por Coomassie Blue Coloidal G-250. Coluna M: Marcador de peso molecular (SeeBlue - Invitrogen); Coluna A: $10\mu g$; Coluna B: $20\mu g$ e Coluna C: μg. Cada coluna foi dividida em 7 frações e cada fração foi subdividida em sub-frações (a-o). As frações e sub-frações foram cortadas e submetidas a hidrólise enzimática por tripsina in situ. Os peptídeos tripsínicos obtidos pelas frações foram separadas pelo sistema de nano-HPLC sendo coletadas na placa de MALDI *off-line* e analisados por MALDI-TOF-TOF-MS.

Das 116 proteínas do resultado do espectrômetro, pesquisamos e observamos que há dentre elas proteínas já descritas na literatura como secretadas. Correlacionando as 116 proteínas, resultado do espectrômetro, com os resultados de predição da validação *in silico* das 2944 proteínas, encontramos 11 proteínas que apresentam peptídeo sinal e *sequon*. Destas 11 proteínas identificadas (peptídeo sinal e *sequon*), 9 estão descritas nos trabalhos da literatura selecionados para validação *in silico* (Figura 21).

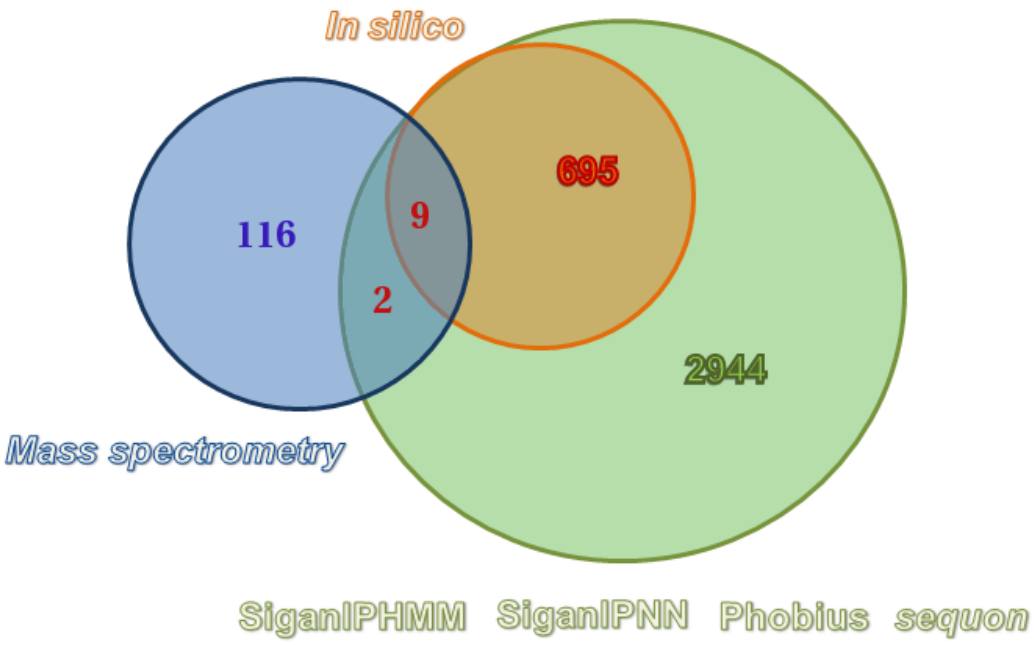


Figura 21: Separação em diagrama de Venn das proteínas secretoras relacionadas ao resultado da linhagem *HCC*1954 por SDS-PAGE.

Observamos que estas 11 proteínas que possuem peptídeo sinal e *sequon*, tiveram uma cobertura de 9,48% das 116 proteínas sequenciadas e podemos notar que destas, 9 já foram identificadas em outros trabalhos. O que conclui-se que a quantidade menor das identificadas como secretadas que continham peptídeo sinal e *sequon*, pode ser dado ao fato de que foram identificadas em células de linhagem controle e não tecido tumoral, pois em trabalhos publicados, podemos observar que a maioria das proteínas secretoras com a presença de peptídeo sinal e *sequon* são identificadas em trabalhos utilizando linhagem celular tumoral. As 105 proteínas, que não foram identificadas nesta abordagem, verificamos e observamos que são proteínas secretadas, porém nenhuma delas possuem peptídeo sinal e *sequon*.

3.4 Dados de Anotação de *GO* por GeneClass

Dados funcionais do *Gene Ontology* classificados pelo *software* GeneClass (PEREIRA *et al.*, 2006) foram associados com as 2944 sequências proteicas com características de secretadas, descrevendo a quantidade de sequências encontradas em cada termo associados as categorias funcionais GO (Processo Biológico, Componente Celular e Função Molecular). Os GeneIDs (*symbols*), solicitados para classificação e análise do *software* foram usados para associação automática no *software* Gene Class com as proteínas candidatas a secretadas. Os resultados das análises foram analisados a fim de verificar a quantidade de proteínas que participam de uma determinada categoria GO no nível 3, que foi especificado para atribuirmos a consulta como mostra os dados da figuras 22, 23 e 24.

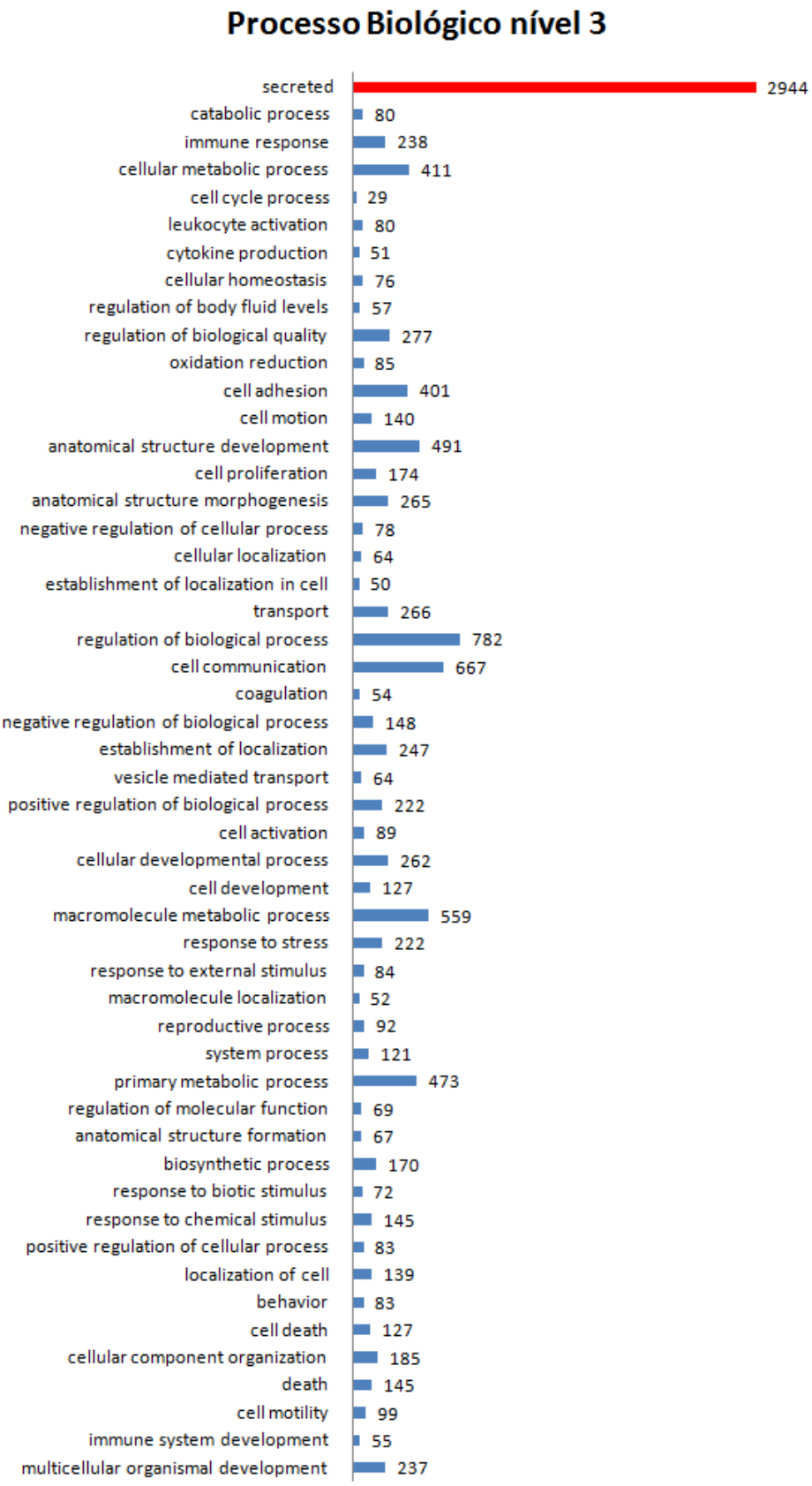


Figura 22: Os dados das 2944 proteínas classificadas como secretadas na categoria Processo Biológico, são mostrados no gráfico.

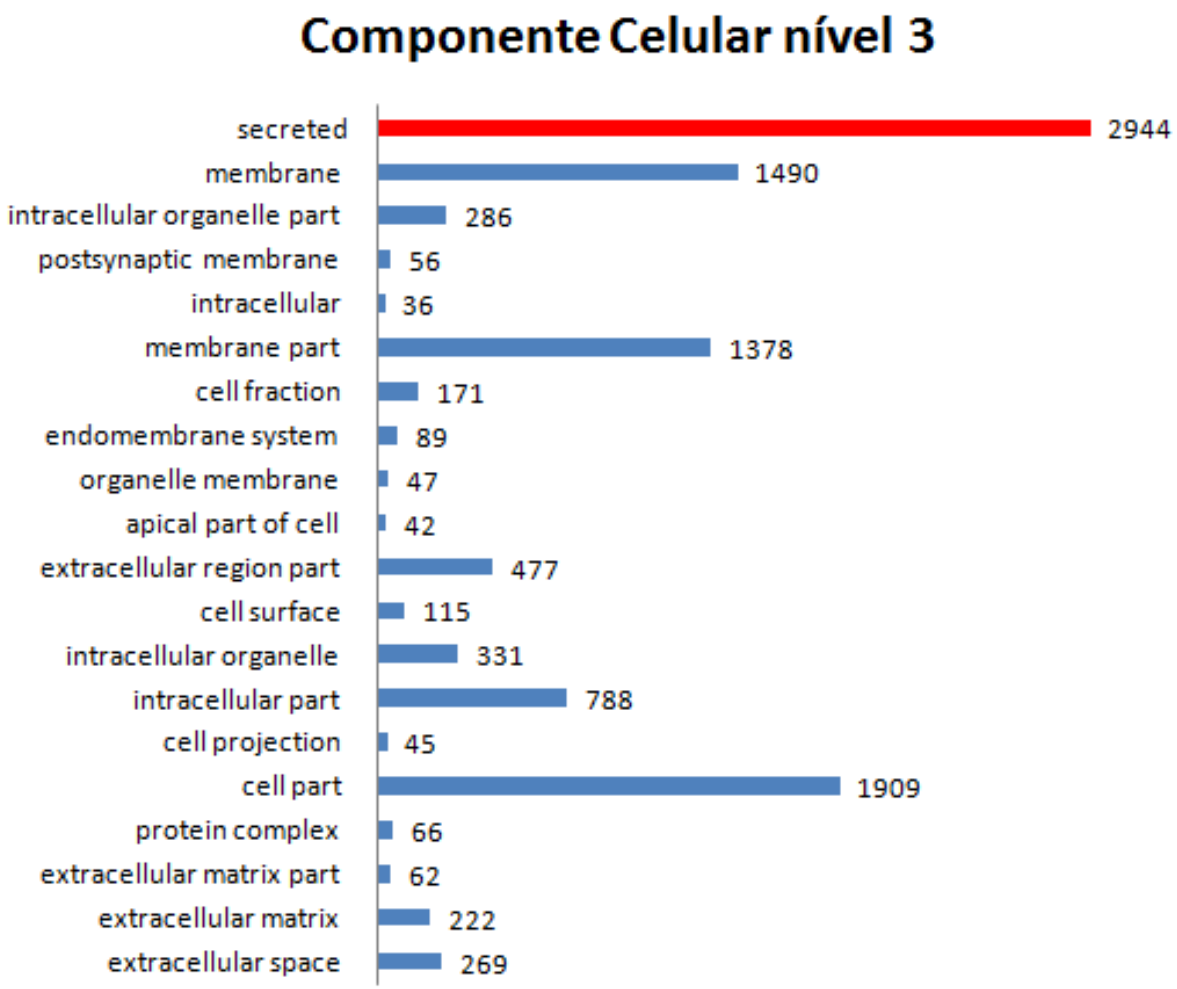


Figura 23: Os dados das 2944 proteínas classificadas como secretadas na categoria Componente Celular, são mostrados no gráfico.

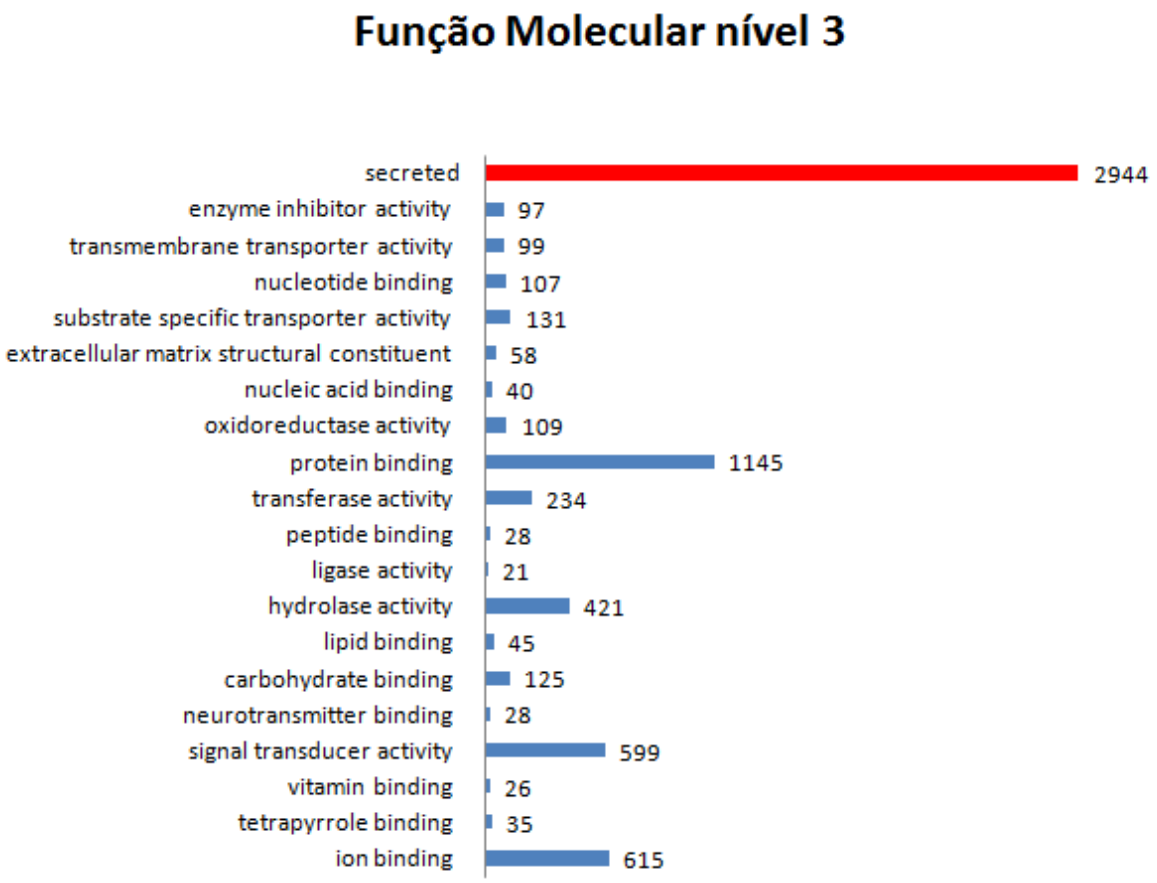


Figura 24: Os dados das 2944 proteínas classificadas como secretadas na categoria unção Molecular, são mostrados no gráfico.

Os resultados da ferramenta possuem a importância de obter novas informações sobre as proteínas secretadas, bem como complementar e ampliar a funcionalidade da mineração dos dados nas hierarquias GO, classificando funções da proteína. Do grupo de proteínas com características de secretadas da célula observamos que várias proteínas participam da mesma categoria dentro da hierarquia GO, o que podemos inferir que talvez estas participem de um mesmo determinado evento biológico nas células, podem controlar a expressão de múltiplos genes, em um determinado momento e modo adequado. Os resultados das categorias funcionais do *Gene Ontology* são importantes para monitorar o padrão de expressão de genes em determinadas condições fisiológicas e patológicas e atualmente tem sido um dos passos para entender processos biológicos que acontecem na célula.

3.5 A Importância da Bioinformática na Previsão de Peptídeo Sinal

O conhecimento de proteínas secretadas pode estimular o desenvolvimento de novos tratamentos para diversas doenças, pois a especulação é que existe um motivo pelo qual estas proteínas estão sendo liberadas (secretadas) e também que por algum motivo estão causando um efeito sistêmico no organismo por via de administração, pelo qual uma droga é colocada em contato com o organismo ou por defesa do próprio organismo proteínas solúveis secretadas por células do sistema imune, que funcionam como mensageiros para ajudar na regulação de uma resposta imune.

A previsão de proteínas secretadas por técnicas de aprendizagem de máquina tem melhorado cada vez mais nos últimos anos, graças aos métodos de desenvolvimento computacional e a abundância de informações da sequência que são utilizadas para formar estes métodos computacionais. Previsões computacionais variáveis ou exatas em proteínas secretadas indicam também o potencial da aprendizagem de máquina automática e na identificação de outros domínios, dado a quantidade de dados genômico, proteômico e etc, que se tornam disponíveis para a utilização, principalmente de Redes Neurais Artificiais e modelos ocultos de Markov. A otimização e o alto desempenho destes *softwares* podem também auxiliar na elucidação de características biológicas conservadas, físicas ou químicas de domínios que não são semelhantes em sequências.

Identificar proteínas secretadas é um importante componente hoje em alvos terapêuticos e de diagnóstico nas iniciativas em desenvolvimento. A parte experimental verifica a falta de localização para averbação de significativas proporções do proteoma humano e aumenta a importância da previsão de algoritmos de bioinformática investigando as melhores seleções de previsão do programa ou do conjunto de predição dos programas que melhor correspondem à natureza da análise. Programas de previsão também possuem recursos valiosos na investigação de localização de proteínas, mas são igualmente importantes a medida em que os algoritmos são utilizados em conjunto com dados de mineração experimental.

4 Conclusão

Em organismos multicelulares, proteínas secretadas possuem função regulatória em comunicação intercelular e vários tecidos de órgãos endócrinos mostram atividade secretora que pode ser de relevância fisiológica. O conhecimento de um conjunto de proteínas secretadas, seu "secretome", pode ajudar no esclarecimento da rede fisiológica de relacionamentos entre tecidos e proporcionar alvos terapêuticos em potencial ou fontes de novas terapias celular. O principal objetivo deste projeto foi apresentar uma abordagem computacional para a predição do Secretoma Humano *in silico* e identificamos por passos sequenciais a predição de peptídeo sinal e motivos de glicosilação, que são características importantes em proteínas secretadas pela via secretora. Por meio destas características obtivemos 2944 proteínas com peptídeo sinal e *sequon* e dentre elas, validamos *in silico* 695 proteínas, utilizando trabalhos com validação experimental descritos na literatura. Nossas análises passaram por refinamento e observação das características na validação *in silico*. Relacionamos nossas análises a dados de SAGE (*Serial Analysis of Gene Expression*) *tag*, do tecido Osteoblasto em tempos de expressão nos tecidos: Mesenquimal T0hs, Mesenquimal T2hs, Mesenquimal T12hs e Osteoblasto T21hs e encontramos 1128 com peptídeo sinal e *sequon* e 49 proteínas nas 695 validadas in *silico* pelos dados descritos na literatura. Em seguida passamos pela validação experimental, onde os passos foram SDS-PAGE e espectrometria de massa MALDI TOF-TOF *Mascot* da linhagem celular ***HCC*1954**. Obtivemos no resultado experimental de espectrometria de massa 116 proteínas e observamos que várias dentre elas fazem parte de descritas na literatura como secretadas. Das 116 proteínas do resultado experimental de espectrometria, encontramos 11 proteínas com as características peptídeo sinal e *sequon* e destas 9 faziam parte da validação *in silico*, procurando nas 2944 de nossos resultados. As 105 proteínas que não foram identificadas nesta abordagem verificamos que não possuem peptídeo sinal e *sequon*. Destes resultados e análises podemos concluir que a abordagem computacional e experimental neste trabalho, buscando proteínas secretadas, mostra a importância e implicação de ambas as áreas trabalhando em conjunto, visando auxiliar no desenvolvimento de estratégias terapêuticas de importância médica e perspectiva farmacêutica.

Referências Bibliográficas

ALEKSANDER, I.; MORTON, H. *An introduction to neural computing*. New York, NY, USA: Van Nostrand Reinhold Co., 1990. Previously published by Chapman & Hall, London. ISBN 0100-4670.

ANTELMANN, H. *et al.* A proteomic view on genome-based signal peptide predictions. *Genome Res*, v. 11, n. 9, p. 1484–502, set. 2001. ISSN 1088-9051.

ARRIGO, P. *et al.* Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *Comp Appl Biosci*, v. 7, n. 3, p. 353–7, jul. 1991. ISSN 0266-7061.

BACHER, G. *et al.* Regulation by the ribosome of the GTPase of the signal-recognition particle during protein targeting. *Nature*, v. 381, n. 6579, p. 248–51, maio 1996. ISSN 0028-0836.

BAIROCH, A.; APWEILER, R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J Mol Med*, v. 75, n. 5, p. 312–6, maio 1997. ISSN 0946-2716.

BENDALL, S. C. *et al.* An enhanced mass spectrometry approach reveals human embryonic stem cell growth factors in culture. *Mol Cell Proteomics*, v. 8, n. 3, p. 421–32, mar. 2009. ISSN 1535-9484.

BENDTSEN, J. D. *et al.* Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, v. 340, n. 4, p. 783–95, jul. 2004. ISSN 0022-2836.

BIRD, P.; GETHING, M. J.; SAMBROOK, J. The functional efficiency of a mammalian signal peptide is directly related to its hydrophobicity. *J Biol Chem*, v. 265, n. 15, p. 8420–5, maio 1990. ISSN 0021-9258.

BONIN-DEBS, A. L. *et al.* Development of secreted proteins as biotherapeutic agents. *Expert Opin Biol Ther*, v. 4, n. 4, p. 551–8, abr. 2004. ISSN 1471-2598.

BRENNER, S. E. Errors in genome annotation. *Trends Genet*, v. 15, n. 4, p. 132–3, abr. 1999. ISSN 0168-9525.

BRUSIN, J. H. Osteogenesis imperfecta. *Radiol Technol*, v. 79, n. 6, p. 535–48; quiz 549–51, 2008. ISSN 0033-8397.

BURKA, E. R. Protein synthesis by membrane-bound reticulocyte ribosomes. *Ann N Y Acad Sci*, v. 241, n. 0, p. 191–203, nov. 1974. ISSN 0077-8923.

CANEPARO, L. *et al.* Dickkopf-1 regulates gastrulation movements by coordinated modulation of Wnt/beta catenin and Wnt/PCP activities, through interaction with the Dally-like homolog Knypek. *Genes Dev*, v. 21, n. 4, p. 465–80, fev. 2007. ISSN 0890-9369.

CHIELLINI, C. *et al.* Characterization of human mesenchymal stem cell secretome at early steps of adipocyte and osteoblast differentiation. *BMC Mol Biol*, v. 9, p. 26, 2008. ISSN 1471-2199.

CHOU, M. M.; KENDALL, D. A. Polymeric sequences reveal a functional interrelationship between hydrophobicity and length of signal peptides. *J Biol Chem*, v. 265, n. 5, p. 2873–80, fev. 1990. ISSN 0021-9258.

CLAVERIE, J.-M. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, v. 6, n. 10, p. 1735–1744, May 1997. ISSN 1735-1744.

DOUDNA, J. A.; BATEY, R. T. Structural insights into the signal recognition particle. *Annu Rev Biochem*, v. 73, p. 539–57, 2004. ISSN 0066-4154.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2nd. ed. NY: John Wiley & Sons, 2001. Interscience publications. ISBN 0471056693.

DURBIN, R. *et al. Biological Sequence Analysis*. Cambridge: Cambridge University Press, 1998. Stock level updated: 15:50 GMT, 22 April 2009. ISBN 10: 0521629713.

DWEK, M. V.; BROOKS, S. A.; SCHUMACHER, U. *Functional and molecular glycobiology*. Oxford Brookes University, Oxford UK: BIOS Scientific, 2002. ISBN 1859960227.

EFRON, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. American Statistical Assoc.*, American Statistical Association, v. 78, n. 382, p. 316–331, 1983. ISSN 01621459.

ESTRADA, R. *et al.* Secretome from mesenchymal stem cells induces angiogenesis via Cyr61. *J Cell Physiol*, v. 219, n. 3, p. 563–71, jun. 2009. ISSN 1097-4652.

FACA, V. M. *et al.* Proteomic analysis of ovarian cancer cells reveals dynamic processes of protein secretion and shedding of extra-cellular domains. *PLoS ONE*, v. 3, n. 6, p. e2425, 2008. ISSN 1932-6203.

FELDMAN, J.; BALLARD, D. Connectionist models and their properties. *Cog. Science*, Elsevier, v. 6, n. 3, p. 205–254, 1982. ISSN 0364-0213.

FORSYTH, D.; PONCE, J. *Computer Vision: A modern approach*. [S.l.]: Prentice Hall Prof. Tec. Ref., 2002. ISBN 9780130851987.

GAVEL, Y.; von Heijne, G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng*, v. 3, n. 5, p. 433–42, abr. 1990. ISSN 0269-2139.

Gavin J. Gibson; RENSHAW, E. Likelihood estimation for stochastic compartmental models using Markov chain methods. *Kluwer Academic*, v. 11, n. 4, p. 347–358, Feb 1995. ISSN 0960-3174.

GERSTEIN, M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol*, v. 274, n. 4, p. 562–76, dez. 1997. ISSN 0022-2836.

GERSTEIN, M. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural consensus. *Proteins*, v. 33, n. 4, p. 518–534, Dec 1998. ISSN 1549-5469.

GERSTEIN, M.; HONIG, B. Sequences and Topology. *Curr. Op. Struc. Biol.*, v. 11, p. 327–329, 2001. ISSN 0959-440. Editorial overview.

GLICK, B. S.; MALHOTRA, V. The curious status of the Golgi apparatus. *Cell*, v. 95, n. 7, p. 883–9, dez. 1998. ISSN 0092-8674.

GREENBAUM, D. *et al.* Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res*, v. 11, n. 9, p. 1463–8, set. 2001. ISSN 1088-9051.

GRUNBAUM, B. *Venn diagrams and independent families of sets*. [S.l.]: Math. Assoc. of America, 1975. 12–23 p. ISBN 0195080300.

GUIGÓ, R. *et al.* An assessment of gene prediction accuracy in large DNA sequences. *Genome Res*, v. 10, n. 10, p. 1631–42, out. 2000. ISSN 1088-9051.

GUNDACKER, N. C. *et al.* Cytoplasmic Proteome and Secretome Profiles of Differently Stimulated Human Dendritic Cells. *J Proteome Res*, maio 2009. ISSN 1535-3893.

HARRISON, P. *et al.* A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol*, v. 316, n. 3, p. 409–19, fev. 2002. ISSN 0022-2836.

HAVEL, P. J. Update on adipocyte hormones: regulation of energy balance and carbohydrate/lipid metabolism. *Diabetes*, v. 53 Suppl 1, p. S143–51, fev. 2004. ISSN 0012-1797.

HEGYI, H.; GERSTEIN, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, v. 288, n. 1, p. 147–64, abr. 1999. ISSN 0022-2836.

HIGA, L. M. *et al.* Secretome of HepG2 cells infected with dengue virus: implications for pathogenesis. *Biochim Biophys Acta*, v. 1784, n. 11, p. 1607–16, nov. 2008. ISSN 0006-3002.

HON, L. *et al.* Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. *Human Mutation*, Wiley Subscription Services, Inc., A Wiley Company Hoboken, v. 30, n. 1, 2009. ISSN 1059-7794.

HORNEGGER, J. *et al.* Object recognition using hidden Markov models. *MIPR*, Citeseer, v. 16, p. 37–37, 1994. ISSN 0162-8828.

HU, J.; BROWN, M. K.; TURIN, W. Hmm based on-line handwriting recognition. Trans. Pattern Anal. Mach. Intell., IEEE, Computer Society, Washington, DC, USA, v. 18, n. 10, p. 1039–1045, 1996. ISSN 162-8828.

HU, M.; POLYAK, K. Serial analysis of gene expression. *Nat Protoc*, v. 1, n. 4, p. 1743–60, 2006. ISSN 1750-2799.

ITO, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNA Science USA*, v. 98, n. 8, p. 4569–74, abr. 2001. ISSN 0027-8424.

KAISER, C. A. *et al.* Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science*, v. 235, n. 4786, p. 312–7, jan. 1987. ISSN 0036-8075.

KALIES, K. U.; HARTMANN, E. Protein translocation into the endoplasmic reticulum (ER)–two similar routes with different modes. *Eur J Biochem*, v. 254, n. 1, p. 1–5, maio 1998. ISSN 0014-2956.

KÄLL, L.; KROGH, A.; SONNHAMMER, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, v. 338, n. 5, p. 1027–36, maio 2004. ISSN 0022-2836.

KHWAJA, F. W. *et al.* Proteomic identification of the wt-p53-regulated tumor cell secretome. *Oncogene*, v. 25, n. 58, p. 7650–61, dez. 2006. ISSN 0950-9232.

KLEE, E. W. The zebrafish secretome. *Zebrafish*, v. 5, n. 2, p. 131–8, 2008. ISSN 1557-8542.

KLEE, E. W.; SOSA, C. P. Computational classification of classically secreted proteins. *Drug Discov Today*, v. 12, n. 5-6, p. 234–40, mar. 2007. ISSN 1359-6446.

KORNFELD, R.; KORNFELD, S. Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem*, v. 54, p. 631–64, 1985. ISSN 0066-4154.

KROGH, A. *et al.* Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, v. 235, n. 5, p. 1501–31, fev. 1994. ISSN 0022-2836.

KROGH, A. *et al.* Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, v. 305, n. 3, p. 567–80, jan. 2001. ISSN 0022-2836.

LADUNGA, I. Large-scale predictions of secretory proteins from mammalian genomic and EST sequences. *Curr Opin Biotechnol*, v. 11, n. 1, p. 13–8, fev. 2000. ISSN 0958-1669.

LADUNGA, I. *et al.* Improving signal peptide prediction accuracy by simulated neural network. *Comp Appl Biosci*, v. 7, n. 4, p. 485–7, out. 1991. ISSN 0266-7061.

LAWLOR, K. *et al.* Pathway-Based Biomarker Search by High-Throughput Proteomics Profiling of Secretomes. *J Proteome Res*, fev. 2009. ISSN 1535-3893.

LODISH, H. *et al. Mol Cell Biology*. Fourth edition. New York, NY, USA: W. H. Freeman & Co, 1999. ISBN 0-7167-3706-X.

LOGANATHAN, D.; AICH, U. Observation of a unique pattern of bifurcated hydrogen bonds in the crystal structures of the N-glycoprotein linkage region models. *Glycobiology*, v. 16, n. 4, p. 343–8, abr. 2006. ISSN 0959-6658.

MACHAMER, C. E.; ROSE, J. K. Influence of New Glycosylation Sites on Expressionof the Vesicular Stomatitis Virus G Protein at the Plasma Membrane. *J. Biol. Chem*, v. 263, n. 12, p. 5948–5954, Apr 1988. ISSN 5948-5954.

MARCOTTE, E. M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science*, v. 285, n. 5428, p. 751–3, jul. 1999. ISSN 0036-8075.

MARSHALL, R. D. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. *Biochem Soc Symp*, n. 40, p. 17–26, 1974. ISSN 0067-8694.

MARTOGLIO, B.; DOBBERSTEIN, B. Snapshots of membrane-translocating proteins. *Trends Cell Biol*, v. 6, n. 4, p. 142–7, abr. 1996. ISSN 0962-8924.

MATHIAS, R. A. *et al.* Secretome-Based Proteomic Profiling of Ras-Transformed MDCK Cells Reveals Extracellular Modulators of Epithelial-Mesenchymal Transition. *J Proteome Res*, abr. 2009. ISSN 1535-3893.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, v. 405, n. 2, p. 442–51, out. 1975. ISSN 0006-3002.

MCGEOCH, D. J. On the predictive recognition of signal peptide sequences. *Virus Res*, v. 3, n. 3, p. 271–86, out. 1985. ISSN 0168-1702.

MITCHELL, T. *Learning Machine*. [S.l.]: Mc-Graw Hill, 1997. ISBN 0-07-042807-7.

MYUNG, I. Tutorial on maximum likelihood estimation. *J Math Psyc*, Elsevier, v. 47, n. 1, p. 90–100, 2003. ISSN 0927-7099.

NIELSEN, H. *et al.* A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst*, v. 8, n. 5-6, p. 581–99, 1997. ISSN 0129-0657.

NIELSEN, H.; KROGH, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol*, v. 6, p. 122–30, 1998. ISSN 1553-0833.

NILSSON, N. *Learning machines: Foundations of trainable pattern-classifying systems*. [S.l.]: McGraw-Hill Education, 1965. ISBN 0070465703.

PAULITSCHKE, V. *et al.* Entering a New Era of Rational Biomarker Discovery for Early Detection of Melanoma Metastases: Secretome Analysis of associated Stroma Cells. *J Proteome Res*, fev. 2009. ISSN 1535-3893.

PEREIRA, G. S. P. *et al.* Gene Class expression: analysis tool of Gene Ontology terms with gene expression data. *Genet Mol Res*, v. 5, n. 1, p. 108–14, 2006. ISSN 1676-5680.

PICKART, M. A. *et al.* Genome-wide reverse genetics framework to identify novel functions of the vertebrate secretome. *PLoS ONE*, v. 1, p. e104, 2006. ISSN 1932-6203.

POWERS, T.; WALTER, P. Protein targeting. The ribosome talks back. *Nature*, v. 381, n. 6579, p. 191–2, maio 1996. ISSN 0028-0836.

RABINER, L. *et al.* A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–286, 1989. ISSN 0018-9219.

RAPOPORT, T. A. Protein transport across the endoplasmic reticulum membrane: facts, models, mysteries. *FASEB J*, v. 5, n. 13, p. 2792–8, out. 1991. ISSN 0892-6638.

RAPOPORT, T. A.; WIEDMANN, M. Application of the signal hypothesis to the incorporation of integral membrane proteins. *In Current Topics in Membranes and Transport*, Academic Press, Inc., New York, NY, USA, v. 24, n. 1, p. 1–63, 1985. ISSN 1943-4456.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. MIT Press, Cambridge, MA, USA, p. 89–114, 1988.

RUMELHART, D.; MCCLELLAND, J. *Parallel distributed processing: Explorations in the microstructure of cognition.* [S.l.: s.n.], 1986. ISBN 0262181207.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning Internal Representations by Error Propagation.* [S.l.]: MIT Press, 1986. 318–362 p. ISBN 0-262-68053-X.

SAKAGUCHI, M. Eukaryotic protein secretion. *Current Biotechnology*, v. 8, n. 5, p. 595–601, October 1997. ISSN 09581669.

SANDERS, S. L. *et al.* Sec61p and BiP directly facilitate polypeptide translocation into the ER. *Cell*, v. 69, n. 2, p. 353–65, abr. 1992. ISSN 0092-8674.

SCHATZ, G.; DOBBERSTEIN, B. Common principles of protein translocation across membranes. *Science*, v. 271, n. 5255, p. 1519–26, mar. 1996. ISSN 0036-8075.

SCHNEIDER, G.; WREDE, P. Development of artificial neural filters for pattern recognition in protein sequences. *J Mol Evol*, v. 36, n. 6, p. 586–95, jun. 1993. ISSN 0022-2844.

SILLENCE, D. O.; SENN, A.; DANKS, D. M. Genetic heterogeneity in osteogenesis imperfecta. *J Med Genet*, v. 16, n. 2, p. 101–16, abr. 1979. ISSN 0022-2593.

SIMONS, K. T.; STRAUSS, C.; BAKER, D. Prospects for ab initio protein structural genomics. *J Mol Biol*, v. 306, n. 5, p. 1191–9, mar. 2001. ISSN 0022-2836.

SIMPSON, J. C.; MATEOS, A.; PEPPERKOK, R. Maturation of the mammalian secretome. *Genome Biol*, v. 8, n. 4, p. 211, 2007. ISSN 1465-6914.

SONNHAMMER, E. L.; von Heijne, G.; KROGH, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, v. 6, p. 175–82, 1998. ISSN 1553-0833.

SPINK, A. J. *et al.* The EthoVision video tracking system–a tool for behavioral phenotyping of transgenic mice. *Physiol Behav*, v. 73, n. 5, p. 731–44, ago. 2001. ISSN 0031-9384.

STARNER, T.; PENTLAND, A. *Visual recognition of american sign language using hidden markov models*. [S.l.]: MIT Press, 1995. ISBN 0-7695-2122-3.

TEICHMANN, S. A.; MURZIN, A. G.; CHOTHIA, C. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol*, v. 11, n. 3, p. 354–63, jun. 2001. ISSN 0959-440X.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. First edition. Cambridge, USA,: Academic Press, Inc., 1999. ISBN 0031-3203.

THORNTON, J. M. From Genome to Function. *Science*, v. 292, n. 5524, p. 2095–2097, June 2001. ISSN 0036-8075.

TJALSMA, H. *et al.* Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol Mol Biol Rev*, v. 64, n. 3, p. 515–47, set. 2000. ISSN 1092-2172.

TWEEDDALE, H.; NOTLEY-MCROBB, L.; FERENCI, T. Effect of slow growth on metabolism of Escherichia coli, as revealed by global metabolite pool ("metabolome") analysis. *J Bacteriol*, v. 180, n. 19, p. 5109–16, out. 1998. ISSN 0021-9193.

VARKI, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, v. 3, n. 2, p. 97–130, abr. 1993. ISSN 0959-6658.

VIRATYOSIN, W. *et al.* Genome-wide subcellular localization of putative outer membrane and extracellular proteins in Leptospira interrogans serovar Lai genome using bioinformatics approaches. *BMC Genomics*, v. 9, p. 181, 2008. ISSN 1471-2164.

VITERBI, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, v. 13, n. 2, p. 260–269, 1967. ISSN 2003-01-06.

von Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem*, v. 133, n. 1, p. 17–21, jun. 1983. ISSN 0014-2956.

von Heijne, G. Signal sequences. The limits of variation. *J Mol Biol*, v. 184, n. 1, p. 99–105, jul. 1985. ISSN 0022-2836.

von Heijne, G. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, v. 14, n. 11, p. 4683–90, jun. 1986. ISSN 0305-1048.

von Heijne, G.; ABRAHMSÉN, L. Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts. *FEBS Lett*, v. 244, n. 2, p. 439–46, fev. 1989. ISSN 0014-5793.

WALTER, P.; JOHNSON, A. E. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu Rev Cell Biol*, v. 10, p. 87–119, 1994. ISSN 0743-4634.

WEST, C. M. Current ideas on the significance of protein glycosylation. *Mol Cell Biochem*, v. 72, n. 1-2, p. 3–20, 1986. ISSN 0300-8177.

WILSON, C. A.; KREYCHMAN, J.; GERSTEIN, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, v. 297, n. 1, p. 233–49, mar. 2000. ISSN 0022-2836.

YEH, R. F.; LIM, L. P.; BURGE, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res*, v. 11, n. 5, p. 803–16, maio 2001. ISSN 1088-9051.

ZHANG, Z.; HENZEL, W. J. Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci*, v. 13, n. 10, p. 2819–24, out. 2004. ISSN 0961-8368.

APÊNDICE A -- 11 sequências com peptídeo sinal e sequon

Sequências com peptídeo sinal em SignalPNN, SignalPHMM, Phobius e *sequon* 2944, e 11 validadas das 116 por espectrômetro de massa MALDI TOF-TOF.

1 -HLA-B major histocompatibility complex, class I, B precursor [Homo sapiens] (Figura 25).

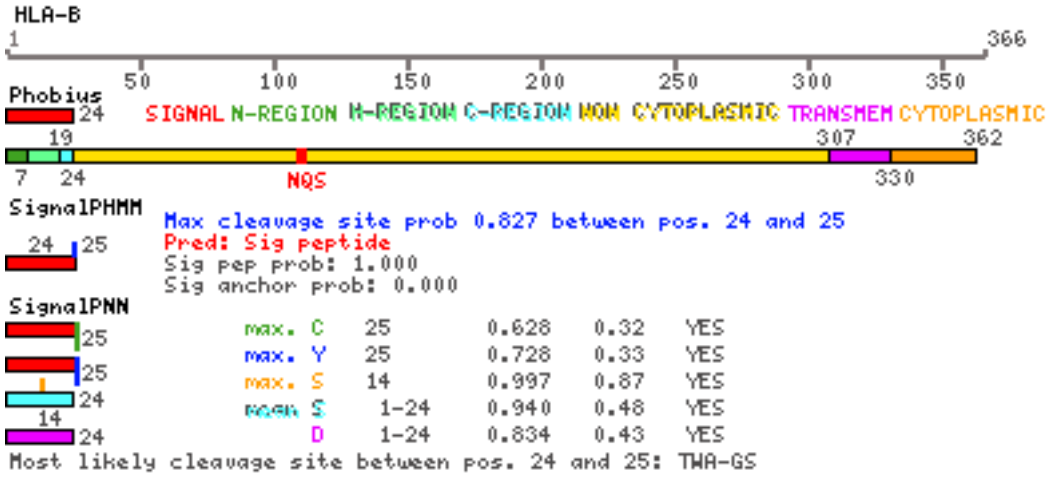


Figura 25: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

2 -LINGO3 leucine rich repeat and Ig domain containing 3 [Homo sapiens] (Figura 26).

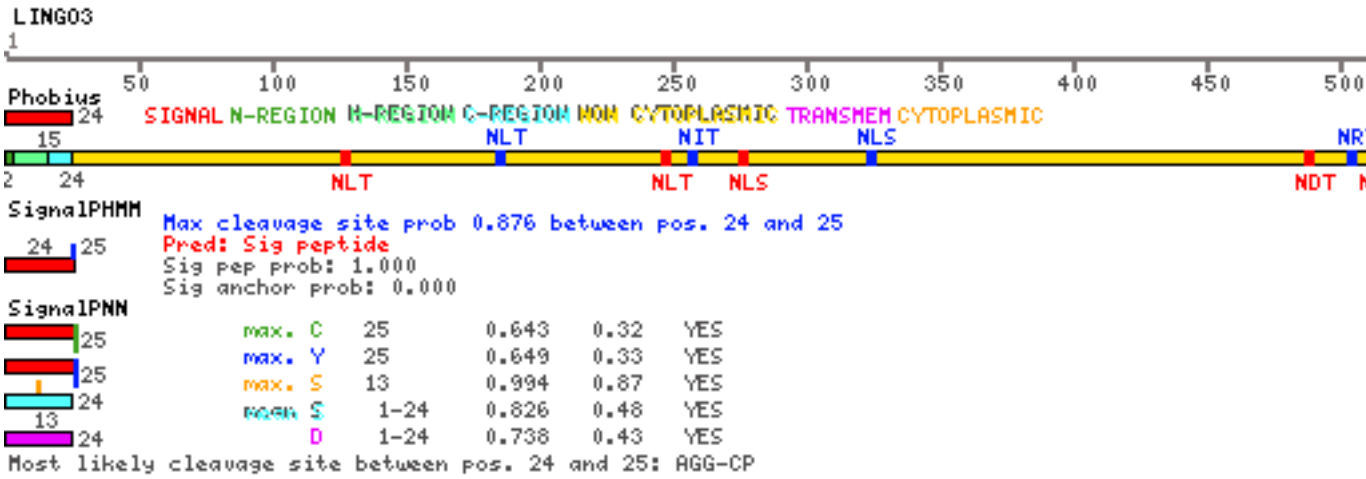


Figura 26: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

3 -HLA-A major histocompatibility complex, class I, A precursor [Homo sapiens] (Figura 27).

Previsão *in silico*: (FAÇA et al., 2008).

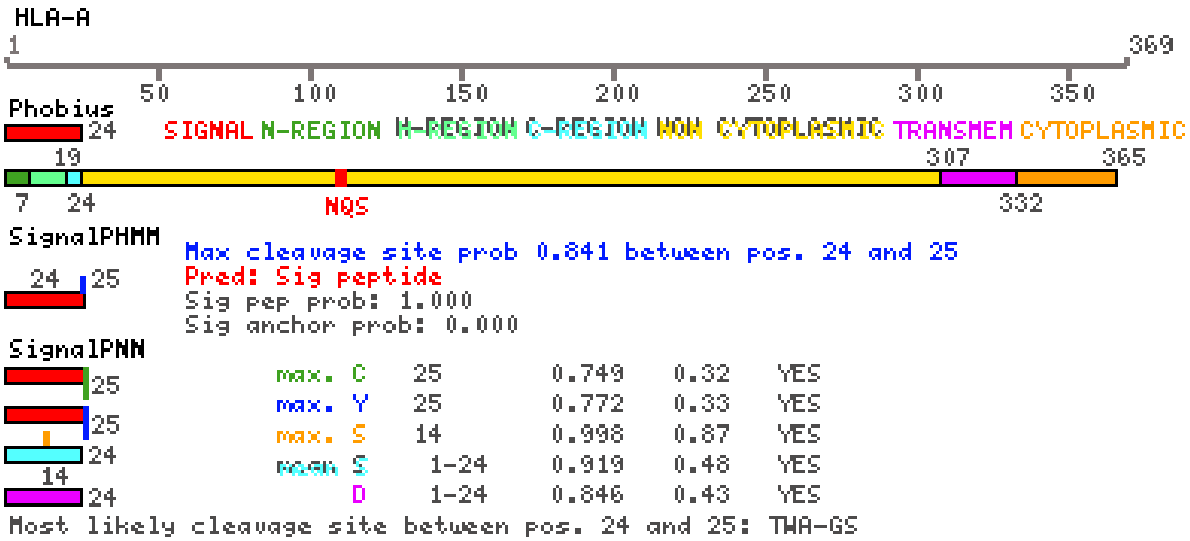


Figura 27: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

4 -CD44 antigen isoform 5 precursor [Homo sapiens] (Figura 28).

Previsão *in silico*: (KHWAJA et al., 2006; GUNDACKER et al., 2009; FAÇA et al., 2008).

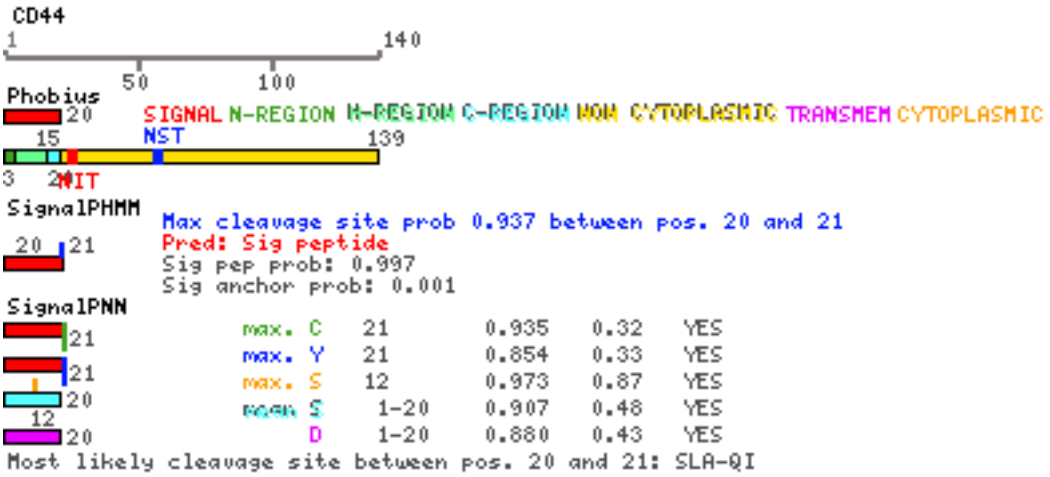


Figura 28: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

5 -HSP90B1 heat shock protein 90kDa beta, member 1 [Homo sapiens] (Figura 29).

Previsão *in silico*: (CHIELLINI et al., 2008; HIGA et al., 2008; BENDALL et al., 2009; FAÇA et al., 2008).

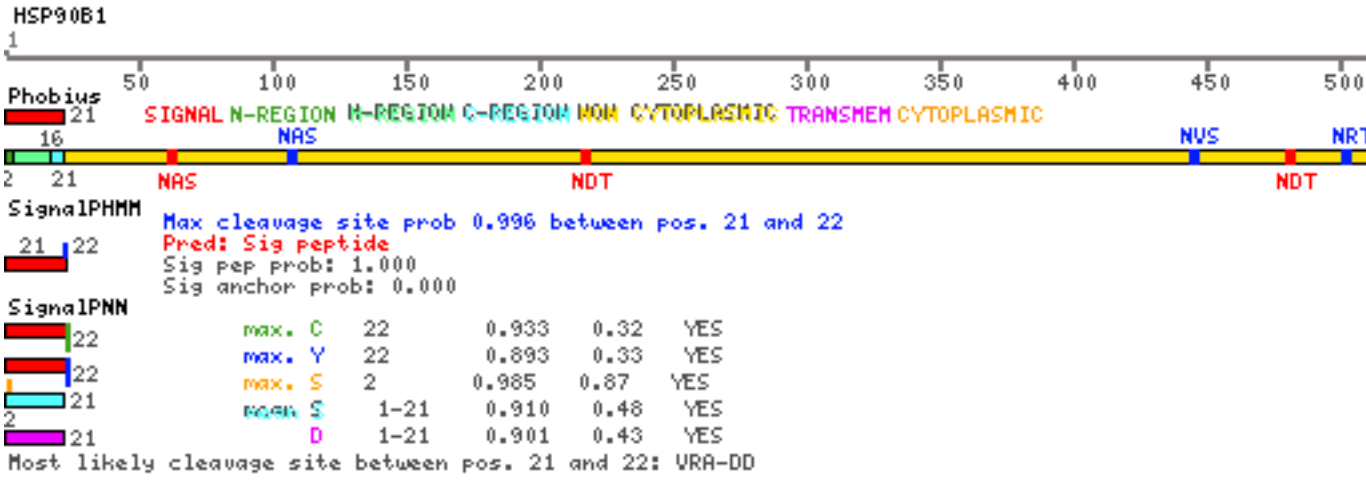


Figura 29: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

6 -HLA-DRB1 major histocompatibility complex, class II, DR beta 1 precursor [Homo sapiens] (Figura 30).

Previsão *in silico*: (GUNDACKER et al., 2009).

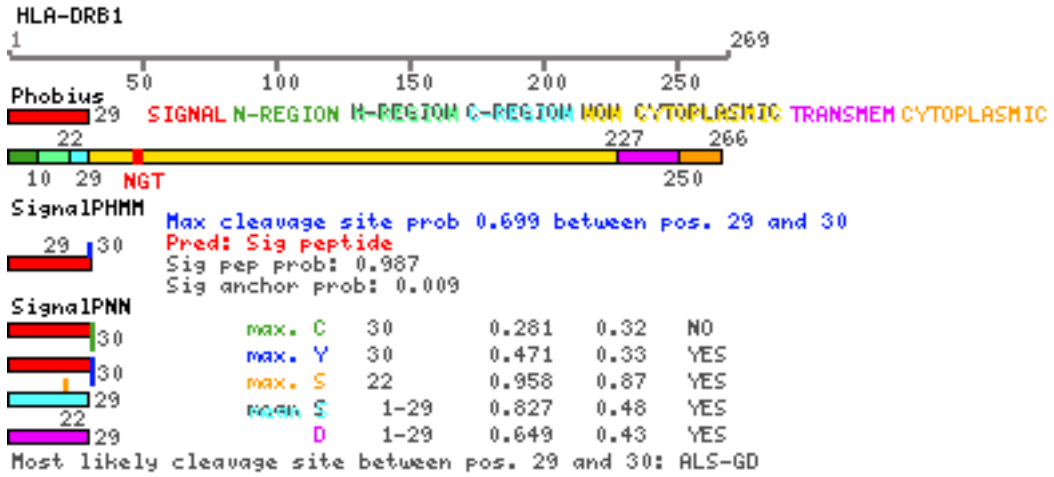


Figura 30: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

7 -CALR calreticulin precursor [Homo sapiens] (Figura 31).

Previsão *in silico*: (CHIELLINI et al., 2008; ESTRADA et al., 2009; HIGA et al., 2008; BENDALL et al., 2009; FAÇA et al., 2008).

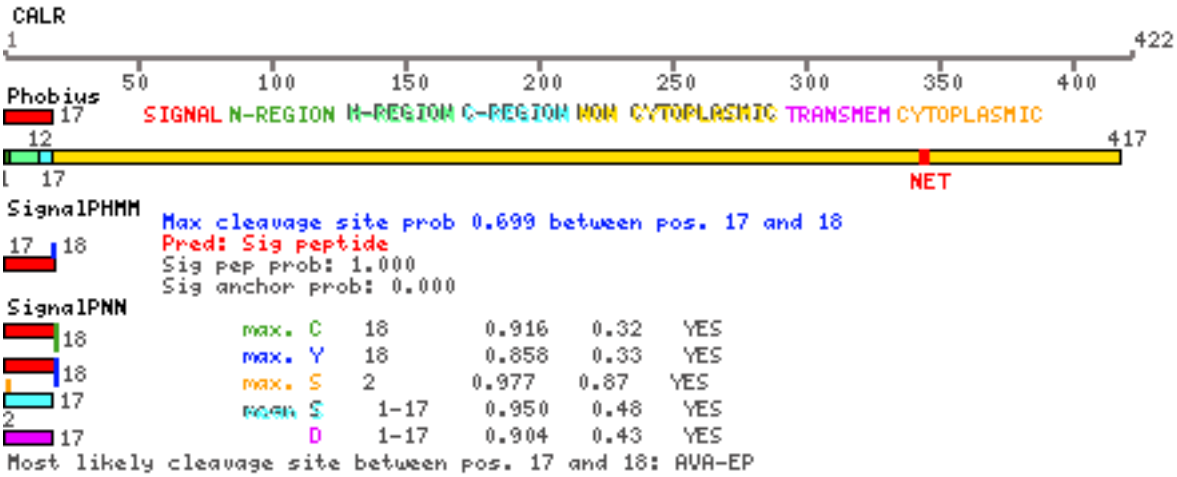


Figura 31: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

8 -TXNDC5 thioredoxin domain containing 5 isoform 1 [Homo sapiens] (Figura 32).

Previsão *in silico*: (HIGA et al., 2008; BENDALL et al., 2009).

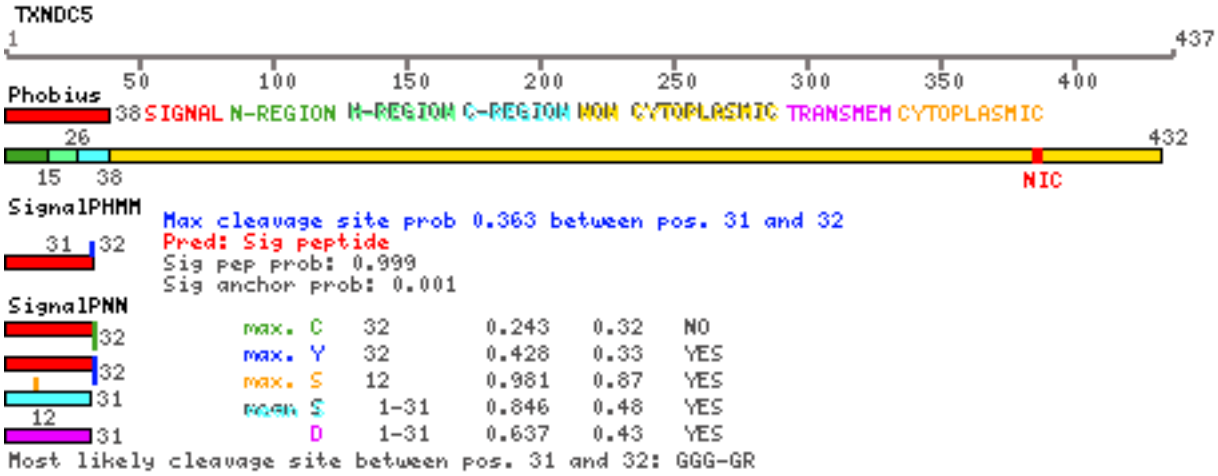


Figura 32: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

9 -ITGA3 integrin alpha 3 isoform b, precursor [Homo sapiens] (Figura 33).

Previsão *in silico*: (FAÇA et al., 2008).

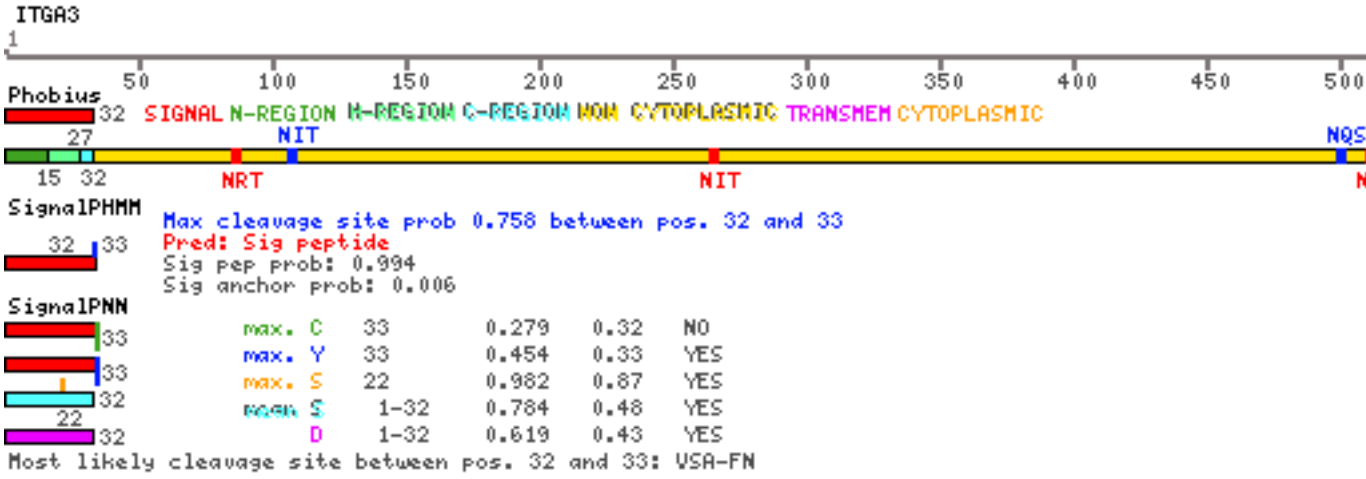


Figura 33: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

10 -ERP29 endoplasmic reticulum protein 29 isoform 1 precursor [Homo sapiens] (Figura 34).

Previsão *in silico*: (ESTRADA et al., 2009; BENDALL et al., 2009; FAÇA et al., 2008).

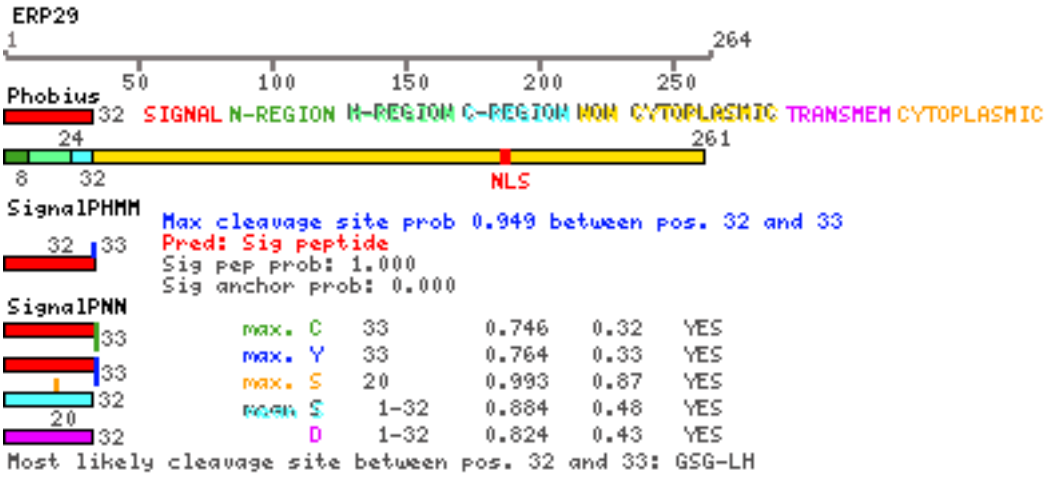


Figura 34: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

11 -PPIB peptidylprolyl isomerase B precursor [Homo sapiens] (Figura 35).

Previsão *in silico*: (CHIELLINI et al., 2008; ESTRADA et al., 2009; HIGA et al., 2008; GUNDACKER et al., 2009; BENDALL et al., 2009; FAÇA et al., 2008).

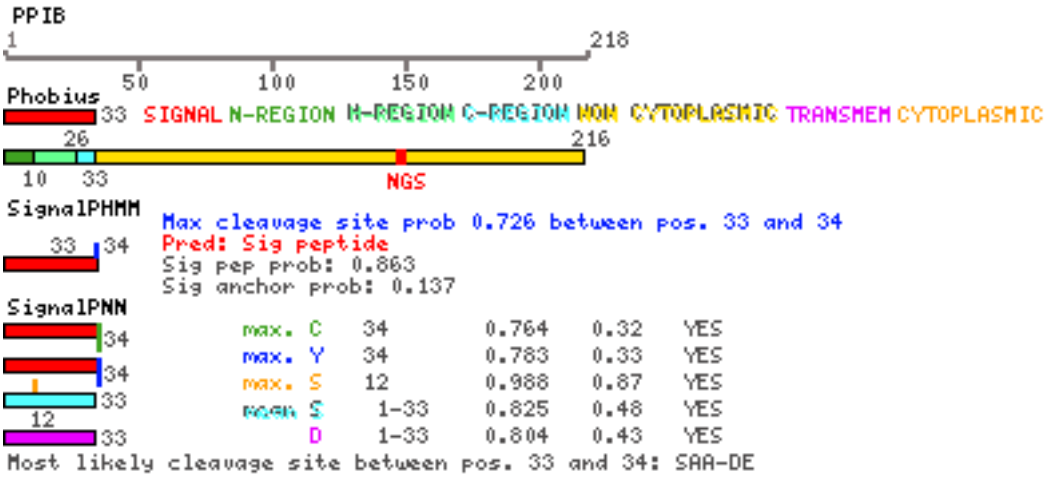


Figura 35: Os resultados dos *softwares* SignalP3.0, Phobius e as sequências com *sequon* estão representados na figura.

APÊNDICE B -- 2 sequências com peptídeo sinal

Sequências com peptídeo sinal em SignalPNN, SignalPHMM, Phobius nas 2944, e 2 validadas das 116 por espectrômetro de massa MALDI TOF-TOF.

1 -LOC652493 PREDICTED: similar to kappa immunoglobulin (subgroup V kappa I) [Homo sapiens] Ig kappa chain V-III region POM OS=Homo sapiens PE=1 SV=1 (Figura 36).

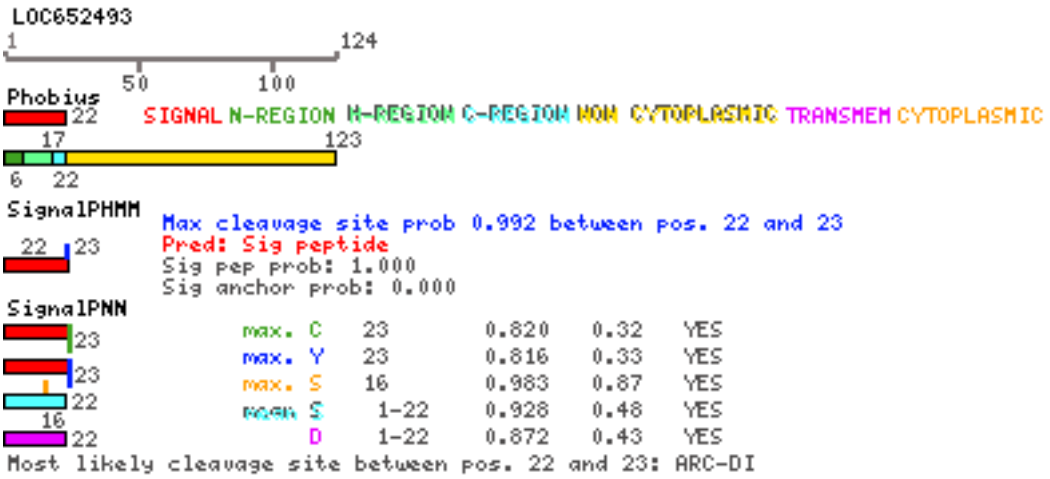


Figura 36: Os resultados dos *softwares* SignalP3.0, Phobius estão representados na figura.

2 -HSPA5 heat shock 70kDa protein 5 [Homo sapiens] 78 kDa glucose-regulated protein OS=Homo sapiens GN=HSPA5 PE=1 SV=2 (Figura 37).

Previsão *in silico*: (CHIELLINI et al., 2008; FAÇA et al., 2008).

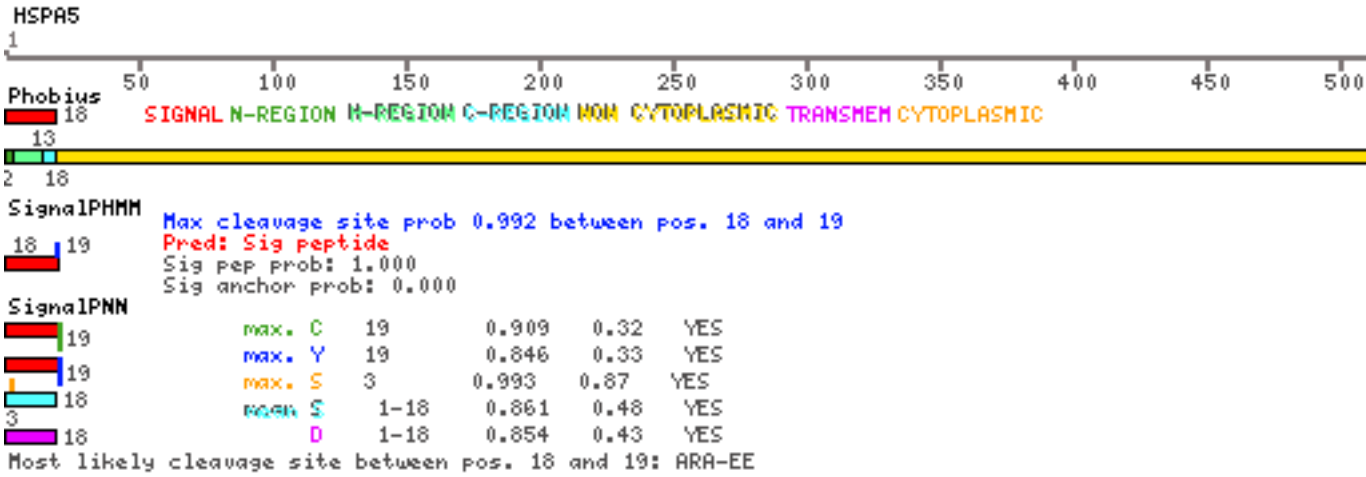


Figura 37: Os resultados dos *softwares* SignalP3.0, Phobius estão representados na figura.

APÊNDICE C -- Interface *de Visualização do Projeto*

Durante a pesquisa foi implementada a *interface web* que está disponível no endereço `http://yanomami.fmrp.usp.br/sp/`, com o objetivo da visualização dos dados no desenvolvimento do projeto para torná-lo mais intuitivo e com *interface* amigável, como podemos visualizar a seguir:

1 -Introdução com o objetivo de disponibilizar a visualização e descrição ao pesquisador da descrição a respeito de cada *software*, utilizado na pesquisa (Figura 38).

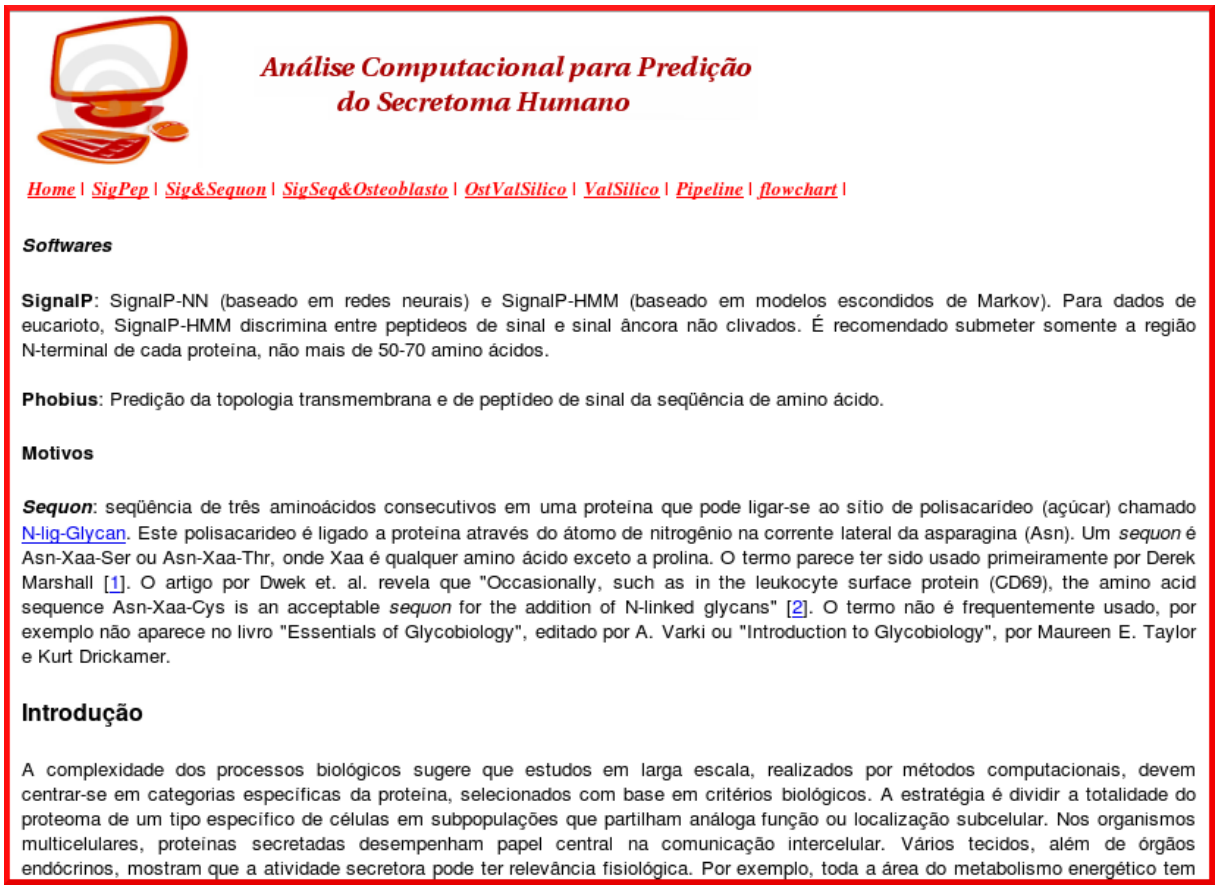


Figura 38: Figura de introdução a respeito dos *softwares* utilizados na busca de peptídeo sinal.

2 -*Interface web* disponibilizando a visualização das sequências com peptídeo sinal encontrados nos *softwares* SignalPNN, SignalPHMM e Phobius (Figura 39).

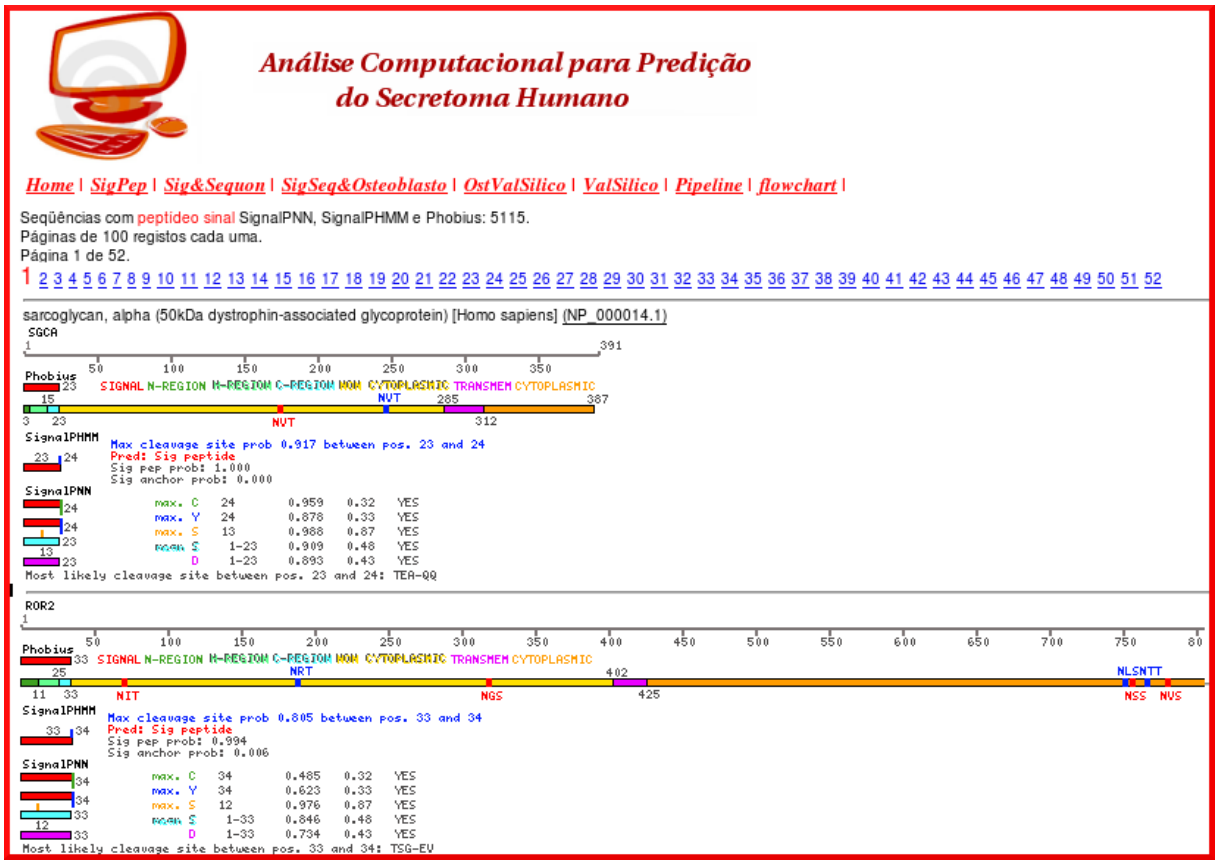


Figura 39: Visualização dos resultados do total de 5115 sequências com peptídeo sinal encontrados nos *softwares* SignalPNN, SignalPHMM e Phobius.

3 -Figura de *Interface web* com visualização das sequências com motivos *sequon* (encontradas pelo *software* FindSequon.pl) e peptídeo sinal (encontrados nos *softwares* SignalPNN, SignalPHMM e Phobius) (Figura 40).

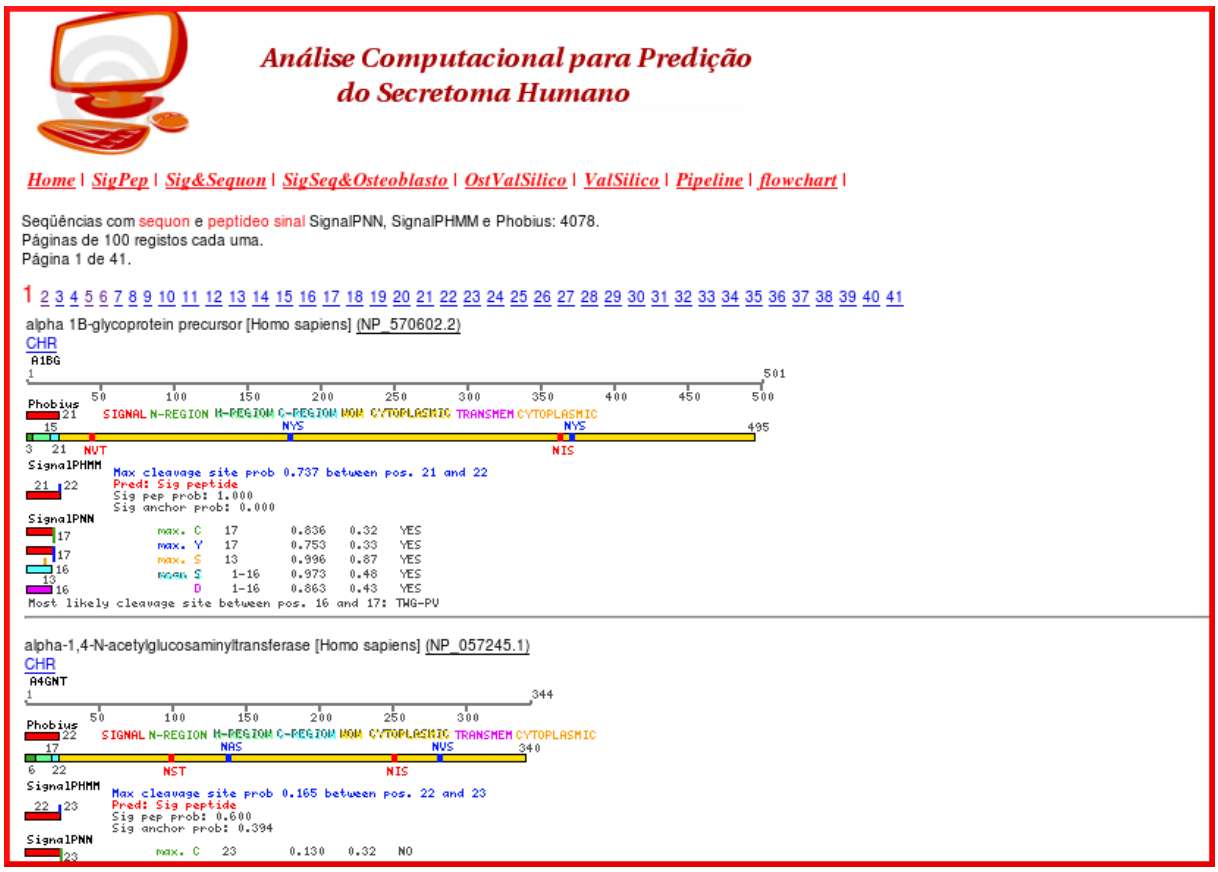
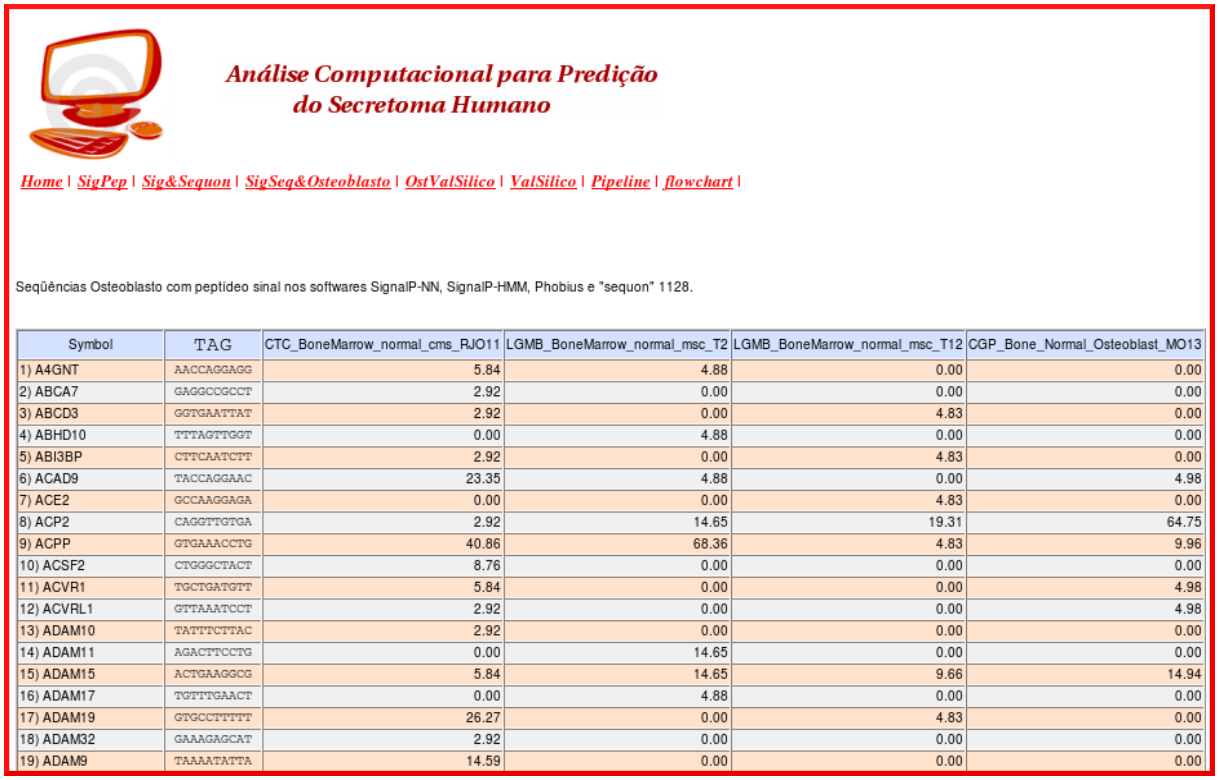


Figura 40: Visualização dos resultados do total de 4078 sequências com peptídeo sinal e motivos *sequon* (encontradas pelo *software* FindSequon.pl) e peptídeo sinal (encontrados nos *softwares* SignalPNN, SignalPHMM e Phobius).

4 -Disponibilização dos dados das sequências de Osteoblasto com peptídeo sinal nos *softwares* SignalP-NN, SignalP-HMM, Phobius e motivos *sequon* encontradas pelo *software* FindSequon.pl (Figura 41).



Análise Computacional para Predição do Secretoma Humano

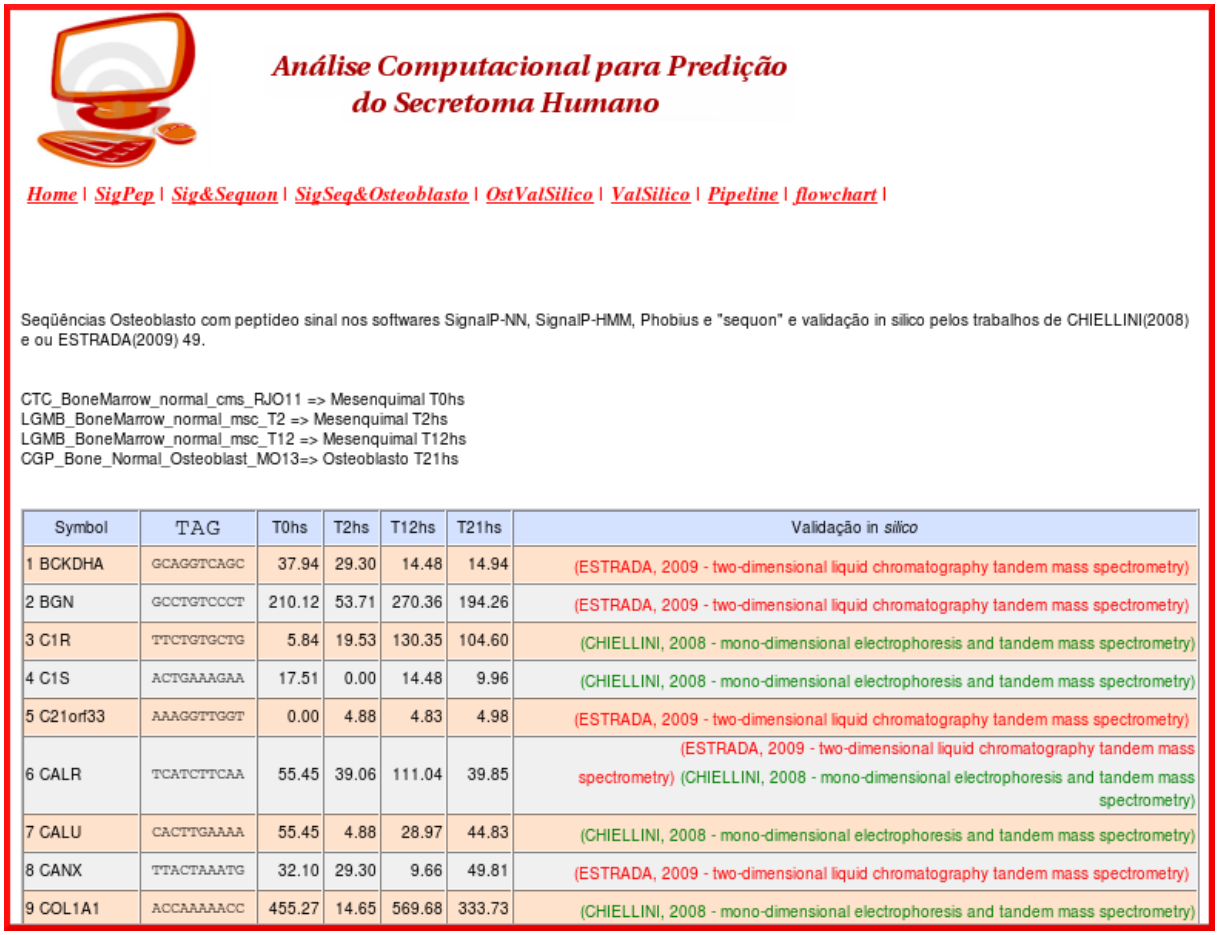
Home | SigPep | Sig&Sequon | SigSeq&Osteoblasto | OstValSilico | ValSilico | Pipeline | flowchart |

Seqüências Osteoblasto com peptideo sinal nos softwares SignalP-NN, SignalP-HMM, Phobius e "sequon" 1128.

Symbol	TAG	CTC_BoneMarrow_normal_cms_RJO11	LGMB_BoneMarrow_normal_msc_T2	LGMB_BoneMarrow_normal_msc_T12	CGP_Bone_Normal_Osteoblast_MO13
1) A4GNT	AACCAGGAGG	5.84	4.88	0.00	0.00
2) ABCA7	GAGGCCGCCT	2.92	0.00	0.00	0.00
3) ABCD3	GGTGAATTAT	2.92	0.00	4.83	0.00
4) ABHD10	TTTAGTTGGT	0.00	4.88	0.00	0.00
5) ABI3BP	CTTCAATCTT	2.92	0.00	4.83	0.00
6) ACAD9	TACCAGGAAC	23.35	4.88	0.00	4.98
7) ACE2	GCCAAGGAGA	0.00	0.00	4.83	0.00
8) ACP2	CAGGTTGTGA	2.92	14.65	19.31	64.75
9) ACPP	GTGAAACCTG	40.86	68.36	4.83	9.96
10) ACSF2	CTGGGCTACT	8.76	0.00	0.00	0.00
11) ACVR1	TGCTGATGTT	5.84	0.00	0.00	4.98
12) ACVRL1	GTTAAATCCT	2.92	0.00	0.00	4.98
13) ADAM10	TATTTCTTAC	2.92	0.00	0.00	0.00
14) ADAM11	AGACTTCCTG	0.00	14.65	0.00	0.00
15) ADAM15	ACTGAAGGCG	5.84	14.65	9.66	14.94
16) ADAM17	TGTTTGAACT	0.00	4.88	0.00	0.00
17) ADAM19	GTGCCTTTTT	26.27	0.00	4.83	0.00
18) ADAM32	GAAAGAGCAT	2.92	0.00	0.00	0.00
19) ADAM9	TAAAATATTA	14.59	0.00	0.00	0.00

Figura 41: Visualização dos resultados do total de 1128 sequências com peptídeo sinal e motivos *sequon* (encontradas pelo *software* FindSequon.pl) e peptídeo sinal (encontrados nos *softwares* SignalPNN, SignalPHMM e Phobius) nas sequências de Osteoblasto.

5 -Tela *web* de listagem das sequências de Osteoblasto com peptídeo sinal nos *softwares* SignalP-NN, SignalP-HMM, Phobius e motivos *sequon* encontradas pelo *software* Find-Sequon.pl, com validação *in silico* pelos trabalhos de Chiellini *et al.* (2008) e Estrada *et al.* (2009) (Figura 42).



Análise Computacional para Predição do Secretoma Humano

Home | SigPep | Sig&Sequon | SigSeq&Osteoblasto | OstValSilico | ValSilico | Pipeline | flowchart |

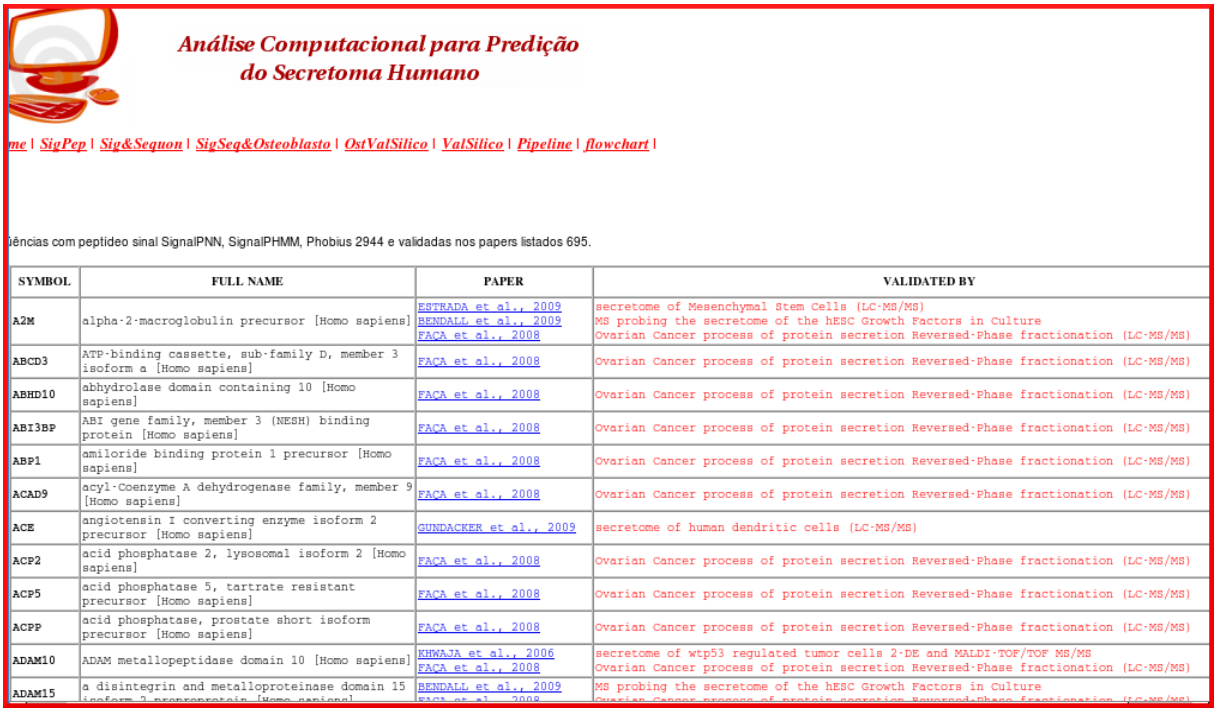
Seqüências Osteoblasto com peptideo sinal nos softwares SignalP-NN, SignalP-HMM, Phobius e "sequon" e validação in silico pelos trabalhos de CHIELLINI(2008) e ou ESTRADA(2009) 49.

CTC_BoneMarrow_normal_cms_RJO11 => Mesenquimal T0hs
LGMB_BoneMarrow_normal_msc_T2 => Mesenquimal T2hs
LGMB_BoneMarrow_normal_msc_T12 => Mesenquimal T12hs
CGP_Bone_Normal_Osteoblast_MO13=> Osteoblasto T21hs

Symbol	TAG	T0hs	T2hs	T12hs	T21hs	Validação in *silico*
1 BCKDHA	GCAGGTCAGC	37.94	29.30	14.48	14.94	(ESTRADA, 2009 - two-dimensional liquid chromatography tandem mass spectrometry)
2 BGN	GCCTGTCCCT	210.12	53.71	270.36	194.26	(ESTRADA, 2009 - two-dimensional liquid chromatography tandem mass spectrometry)
3 C1R	TTCTGTGCTG	5.84	19.53	130.35	104.60	(CHIELLINI, 2008 - mono-dimensional electrophoresis and tandem mass spectrometry)
4 C1S	ACTGAAAGAA	17.51	0.00	14.48	9.96	(CHIELLINI, 2008 - mono-dimensional electrophoresis and tandem mass spectrometry)
5 C21orf33	AAAGGTTGGT	0.00	4.88	4.83	4.98	(ESTRADA, 2009 - two-dimensional liquid chromatography tandem mass spectrometry)
6 CALR	TCATCTTCAA	55.45	39.06	111.04	39.85	(ESTRADA, 2009 - two-dimensional liquid chromatography tandem mass spectrometry) (CHIELLINI, 2008 - mono-dimensional electrophoresis and tandem mass spectrometry)
7 CALU	CACTTGAAAA	55.45	4.88	28.97	44.83	(CHIELLINI, 2008 - mono-dimensional electrophoresis and tandem mass spectrometry)
8 CANX	TTACTAAATG	32.10	29.30	9.66	49.81	(ESTRADA, 2009 - two-dimensional liquid chromatography tandem mass spectrometry)
9 COL1A1	ACCAAAAACC	455.27	14.65	569.68	333.73	(CHIELLINI, 2008 - mono-dimensional electrophoresis and tandem mass spectrometry)

Figura 42: Visualização *web* dos resultados do total de 49 sequências com peptídeo sinal e motivos *sequon* validadas *in silico* nas sequências de Osteoblasto, com validação *in silico* pelos trabalhos de Chiellini *et al.* (2008) e Estrada *et al.* (2009).

6 -Listagem *online* das 695 sequências validadas pelos dados da literatura, encontradas nas 2944 sequências com peptídeo sinal encontradas pelos *softwares* SignalPNN, SignalPHMM, Phobius e com motivos *sequon* pelo *software* FindSequon.pl (Figura 43).



Análise Computacional para Predição do Secretoma Humano

me | SigPep | Sig&Sequon | SigSeq&Osteoblasto | OstValSilico | ValSilico | Pipeline | flowchart |

ências com peptideo sinal SignalPNN, SignalPHMM, Phobius 2944 e validadas nos papers listados 695.

SYMBOL	FULL NAME	PAPER	VALIDATED BY
A2M	alpha-2-macroglobulin precursor [Homo sapiens]	ESTRADA et al., 2009 BENDALL et al., 2009 FAÇA et al., 2008	secretome of Mesenchymal Stem Cells (LC-MS/MS) MS probing the secretome of the hESC Growth Factors in Culture Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ABCD3	ATP-binding cassette, sub-family D, member 3 isoform a [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ABHD10	abhydrolase domain containing 10 [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ABI3BP	ABI gene family, member 3 (NESH) binding protein [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ABP1	amiloride binding protein 1 precursor [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ACAD9	acyl-Coenzyme A dehydrogenase family, member 9 [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ACE	angiotensin I converting enzyme isoform 2 precursor [Homo sapiens]	GUNDACKER et al., 2009	secretome of human dendritic cells (LC-MS/MS)
ACP2	acid phosphatase 2, lysosomal isoform 2 [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ACP5	acid phosphatase 5, tartrate resistant precursor [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ACPP	acid phosphatase, prostate short isoform precursor [Homo sapiens]	FAÇA et al., 2008	Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ADAM10	ADAM metallopeptidase domain 10 [Homo sapiens]	KHWAJA et al., 2006 FAÇA et al., 2008	secretome of wtp53 regulated tumor cells 2-DE and MALDI-TOF/TOF MS/MS Ovarian Cancer process of protein secretion Reversed-Phase fractionation (LC-MS/MS)
ADAM15	a disintegrin and metalloproteinase domain 15	BENDALL et al., 2009	MS probing the secretome of the hESC Growth Factors in Culture

Figura 43: Visualização *web* dos resultados do total de 695 sequências com peptídeo sinal e motivos *sequon* validadas *in silico* pelos dados da literatura.

7 -Listagem do *pipeline in silico e flowchart* de documentação (`http://yanomami.fmrp.usp.br/sp/pipe.php` e `http://yanomami.fmrp.usp.br/sp/flow.php`).

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo