

**FÁBIO ROGÉRIO DE MORAES**

**“REVELANDO AS CARACTERÍSTICAS DO NANO-  
AMBIENTE DAS INTERFACES ENTRE PROTEÍNAS”**

**CAMPINAS  
2012**



UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

**FÁBIO ROGÉRIO DE MORAES**


**“REVELANDO AS CARACTERÍSTICAS DO NANO-AMBIENTE DAS INTERFACES ENTRE PROTEÍNAS”**

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a)

*Fábio Rogério de Moraes*

e aprovada pela Comissão Julgadora.

Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Bioinformática.

  
Orientador: Prof. Dr. Goran Neshich

CAMPINAS,  
2012

FICHA CATALOGRÁFICA ELABORADA POR  
ROBERTA CRISTINA DAL' EVEDOVE TARTAROTTI – CRB8/7430  
BIBLIOTECA DO INSTITUTO DE BIOLOGIA - UNICAMP

M791p Moraes, Fábio Rogério de, 1984-  
Revelando as características do nano-ambiente das interfaces entre proteínas / Fábio Rogério de Moraes. – Campinas, SP: [s.n.], 2012.

Orientador: Goran Neshich.  
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Interações proteína-proteína. 2. Interface oligomérica. 3. Modelos classificadores. 4. Aprendizado de máquina. I. Neshich, Goran. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

**Título em Inglês:** Characteristics of protein interface nano-environment revealed

**Palavras-chave em Inglês:**

Protein-protein interactions

Oligomeric interface

Classifier models

Machine learning

**Área de concentração:** Bioinformática

**Titulação:** Doutor em Genética e Biologia Molecular

**Banca examinadora:**

Goran Neshich[Orientador]

Ricardo Aparicio

Fernando José Von Zuben

Raghuvir Krishnaswamy Arni

Marcelo Matos Santoro

**Data da defesa:** 21-08-2012

**Programa de Pós Graduação:** Genética e Biologia Molecular

Campinas, 21 de Agosto de 2012

BANCA EXAMINADORA

Prof. Dr. Goran Neshich (Orientador)



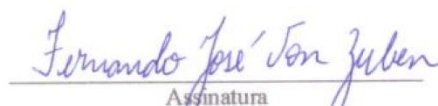
Assinatura

Prof. Dr. Ricardo Aparicio



Assinatura

Prof. Dr. Fernando José Von Zuben



Assinatura

Prof. Dr. Raghuvir Krishnaswamy Arni



Assinatura

Prof. Dr. Marcelo Matos Santoro



Assinatura

Profa. Dra. Ljubica Tasic

Assinatura

Prof. Dr. Munir Salomão Skaf

Assinatura

Prof. Dr. André Luis Berteli Ambrosio

Assinatura

"(...) Disse certa vez um poeta: 'Todo o universo está em um copo de vinho'. Provavelmente jamais saberemos o que ele quis dizer, pois os poetas não escrevem para serem entendidos. Mas é verdade que, se examinarmos um copo de vinho bem de perto, veremos todo o universo. Há as coisas da física: o líquido vivo que evapora dependendo do vento e do clima, os reflexos no copo, a nossa imaginação acrescenta os átomos. O copo é uma destilação das rochas da Terra e, em sua composição, vemos os segredos da idade do universo e da evolução das estrelas. Que estranho arranjo de substâncias químicas está no vinho? Como vieram à existência? Há os fermentos, as enzimas, os substratos e os produtos. Ali no vinho encontra-se a maior generalização: toda vida é fermentação. Ninguém descobre a química do vinho sem descobrir, como Louis Pasteur, a causa de muitas doenças. Como é vivo o clarete, impondo sua existência à consciência que o observa! Se nossas pequenas mentes, por alguma conveniência, dividem esse copo de vinho, o universo, em partes - física, biologia, química, geologia, astronomia, psicologia e assim por diante -, lembre-se de que a natureza as ignora! Assim, reunamos tudo de volta, sem esquecer para que serve afinal. Que nos conceda mais um último prazer: bebê-lo e esquecer tudo isso!"

Richard P. Feynman

## DEDICATÓRIA

Aos meus pais Sebastião e Maria e minha esposa Liane, razões deu estar aqui.

## AGRADECIMENTOS

Diversas pessoas contribuíram de várias maneiras para que eu pudesse escrever essa tese. Acredito que não conseguirei listar todas elas que me ajudaram de alguma forma nesse período de mais de 4 anos.

Em primeiro lugar agradeço a Deus, criador desse universo interessante o suficiente e que vale a pena estudá-lo, entendê-lo de forma mais aproximada o possível.

Agradeço muito ao meu orientador Goran Neshich pela caminhada e por me aceitar em seu laboratório, além, é claro, pelos grandes e valiosos ensinamentos. Por ser exemplo de profissional, comprometido com a ciência.

Agradeço aos meus pais, que lutaram por mim e aceitaram as escolhas que fiz nesse longo caminho, que acredito ter iniciado ainda no ensino médio, passando pela graduação e chegando ao doutorado. Sem o apoio incondicional deles, provavelmente não estaria aqui.

A minha esposa e eterna namorada, por caminhar comigo, entender e apoiar minhas escolhas, me animar quando estava desanimado, me criticar quando necessário e por tudo que me faz seguir em frente. Sem seu apoio, não sei se conseguiria.

Ao grande amigo Renato Milani, desde a época do ensino médio até o IB. Sempre disposto a dialogar, esclarecer e me mostrar um lado sobre todas as situações que provavelmente não conseguiria ver sem sua ajuda.

Aos outros grandes amigos Everton, Ricardinho, Rodrigo, Gustavo, Luís Fernando e Tiago pela amizade, conselhos e por tornar minha vida mais agradável.

Aos amigos e parceiros de laboratório Izabella, Ivan, José Salim, José Geraldo, José Jardine e Aduino. Cada um com sua contribuição no desenvolvimento desse trabalho, por me ajudar e tolerar meus pedidos constantes por ajuda.

Ao parceiro Wilfredo que me ensinou tudo (pessoalmente e pela sua dissertação de mestrado) o que sei sobre redes neurais e sistemas classificadores, essencial nessa tese. Agradeço também ao professor Fernando Von Zuben por apontar o Wilfredo como colaborador desse projeto.

Agradeço profundamente às agências de fomento CAPES e FAPESP pela concessão de bolsa, o que possibilitou o desenvolvimento desse trabalho, bem como minha participação em congressos da área.

Sou grato por todas elas.

## Sumário

Resumo .....	xvii
Abstract.....	xviii
1 Introdução.....	1
2 Metodologia.....	11
2.1 Seleção das estruturas dos complexos proteicos do PDB e PISA .....	11
2.2 Agrupamento do banco de dados .....	14
2.3 Conjunto teste final – Docking benchmark .....	14
2.4 BlueStar STING e STING_DB .....	17
2.5 Seleção dos descritores físico-químicos e estruturais .....	19
2.5.1 Potencial Eletrostático .....	21
2.5.2 Hidrofobicidade .....	21
2.5.3 Contatos e Densidade de Energia de Contatos .....	23
2.5.4 Rotâmeros .....	24
2.5.5 Acessibilidade .....	25
2.5.6 Contatos não Usados.....	26
2.5.7 Ordem de <i>Cross Link</i> .....	27
2.5.8 Ordem de <i>Cross Presence</i> .....	27
2.5.9 Densidade.....	27
2.5.10 Esponjicidade.....	28
2.6 Descritores ponderados pela vizinhança (WNA) .....	28
2.7 Testes estatísticos uni e multivariados .....	29
2.7.1 Teste de Welch.....	29
2.7.2 Teste de normalidade de D’Agostino .....	31
2.7.3 Teste não paramétrico de soma de ranques de Wilcoxon.....	31
2.7.4 Teste não paramétrico de duas amostras de Kolmogorov–Smirnov.....	32
2.7.5 Análise de variância multivariada.....	33
2.8 Limitação de testes estatísticos para grandes bancos de dados .....	34
2.9 Gráficos do tipo boxplot.....	35
2.10 Modelos classificadores.....	36



2.10.1	Modelos de regressão linear multivariados.....	37
2.10.2	Modelo classificador por análise discriminante linear.....	39
2.10.3	Árvore de inferência condicional não enviesada .....	40
2.10.4	Ensemble de árvores de inferência condicional via <i>bagging</i> .....	42
2.10.5	Ensemble de árvores de decisão via <i>random Forest</i> .....	43
2.10.6	Classificação por regressão logística .....	43
2.10.7	Classificador Naïve Bayes .....	44
2.10.8	Classificador por redes-neurais no R .....	46
2.10.9	Classificador por máquinas de vetores suporte.....	47
2.11	Tratamento dos dados por análise de componentes principais .....	51
2.12	Comparação entre classificadores por análise ROC .....	53
2.13	Classificadores por redes neurais artificiais com processo de seleção de variáveis e formação de ensemble de classificadores.....	55
3	Resultados e Discussão.....	63
3.1	Correlação linear entre os descritores .....	63
3.2	Testes Estatísticos de normalidade.....	66
3.3	Testes estatísticos uni e multivariados .....	67
3.4	Gráficos do tipo boxplot.....	70
3.5	Análise dos descritores por modelo de regressão linear multivariado .....	73
3.6	Desenvolvimento de novos modelos classificadores .....	76
3.7	Modelo classificador linear por LDA.....	81
3.8	Classificadores por redes neurais artificiais .....	89
4	Conclusões.....	107
5	Referências .....	111
6	Apêndice.....	117
6.1	Apêndice 1 – Resultados dos testes estatísticos univariados.....	117
6.2	Apêndice 2 – Resultado dos testes estatísticos de MANOVA.....	147
6.3	Apêndice 3 – Gráficos do tipo boxplot para todos os descritores e organizada por tipo de aminoácido. ....	157
6.4	Apêndice 4 – Comparação das curvas ROC para cada um dos oito tipos de modelos classificadores para cada aminoácido. ....	177

6.5	Apêndice 5 – Processo de seleção de variáveis para o classificador por redes neurais.....	182
6.6	Apêndice 6 – Lista com a ordem dos descritores removidos no processo de seleção de variáveis.....	191
6.7	Apêndice 7 – Lista com os códigos PDB e cadeia de cada entrada utilizada como conjunto teste na metodologia de avaliação por holdout.....	192

# Lista de Figuras

<i>Figura 1 – Ilustração da região de interface em um complexo proteína-proteína com estrutura tridimensional conhecida.</i>	4
<i>Figura 2 – Estrutura do complexo entre a proteína serino-protease com três inibidores diferentes que compartilham a mesma região de interface.</i>	5
<i>Figura 3 – Fluxograma com os procedimentos para análise, tratamento dos dados e maximização de desempenho.</i>	10
<i>Figura 4 – Filtros estabelecidos para a seleção de complexos proteicos oriundos do PDB.</i>	13
<i>Figura 5 – Ilustração da modificação conformacional sofrida por duas proteínas do grupo “difícil” do docking benchmark (versão 4).</i>	15
<i>Figura 6 – Ângulos <math>\chi</math> da cadeia lateral do aminoácido arginina (único aminoácido cuja cadeia lateral se estende até o ângulo <math>\chi</math>-5).</i>	25
<i>Figura 7 – Definição de superfície acessível de acordo com a superfície de van der Waals de cada molécula.</i>	26
<i>Figura 8 – Grandezas estatísticas medidas em um gráfico do tipo boxplot.</i>	36
<i>Figura 9 – Matriz de confusão para resultados de modelos classificadores binários.</i>	37
<i>Figura 10 – Representação de uma rede bayesiana simples (Naïve Bayes).</i>	45
<i>Figura 11 – Arquitetura da rede MLP utilizada com uma camada de entrada (dados provenientes da base de dados BlueStar STING), uma camada oculta e uma camada de saída.</i>	57
<i>Figura 12 – Criação do ensemble de redes neurais como método para evitar sobre-ajuste aos dados de treinamento.</i>	61
<i>Figura 13 – Comparação entre a distribuição dos descritores “energia total de contatos” (a) e “densidade de energia de contatos no CA-4” (b) em relação as classes IFR (branco) e FSR (cinza). Barras de mínimo e máximo foram omitidas.</i>	71
<i>Figura 14 – Comparação entre os valores médios divididos pelo desvio padrão de cada descritor físico-químico ou estrutural para as classes FSR (azul) e IFR (vermelho) para o resíduo de aminoácido valina.</i>	72

Figura 15 – Avaliação de desempenho de cada modelo classificador em relação aos 20 tipos de aminoácido.....	77
Figura 16 – Propensidade de cada tipo de aminoácido de estar localizado na superfície livre (FSR) ou na interface (IFR). .....	78
Figura 17 – Comparação do desempenho dos modelos classificadores. ....	78
Figura 18 – Comparação segundo o critério de AUC para os oito modelos classificadores (ordenados de forma decrescente, no eixo x) divididos para cada um dos 20 tipos de aminoácido. ....	80
Figura 19 – Desempenho de cada classificador específico de aminoácido durante processo de validação cruzada, avaliados pelo critério AUC (branco) e MCC (cinza). Tipos de aminoácido ordenados por maior desempenho.....	81
Figura 20 – Comparação do o desempenho do classificador por LDA em relação a escolha do limiar de similaridade sequencial na criação do banco de dados. ....	83
Figura 21 – Comparação entre os 20 classificadores específicos de aminoácidos agregados (em preto) com o classificador inespecífico de aminoácidos (em azul) por curvas ROC (a) e pelo critério de AUC utilizando boxplots (b). ....	84
Figura 22 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido. ....	85
Figura 23 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido após a incorporação dos descritores ponderados pela vizinhança (WNA). ....	86
Figura 24 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido após a incorporação dos descritores ponderados pela vizinhança (WNA) e conservação de aminoácidos (a) e comparação entre os três modelos classificadores por LDA (b). ....	87
Figura 25 – Comparação entre o classificador por LDA desenvolvido (Sting-LDA) e outros métodos com base no conjunto teste 35Enz. ....	88
Figura 26 – Processo de seleção de variável utilizando a rede neural para síntese de classificadores do tipo envoltório (wrapper). ....	91
Figura 27 – Variação do desempenho dos classificadores após a formação do ensemble de classificadores. ....	96

Figura 28 – <i>Desempenho para o conjunto de teste dos ensembles classificadores específicos para cada aminoácido, da combinação dos 20 (Sting-nn) e do inespecífico quanto ao tipo do aminoácido.</i> .....	98
Figura 29 – <i>Desempenho de cada classificador específico de aminoácido por validação cruzada.</i> .....	99
Figura 30 – <i>Desempenho de cada classificador específico de aminoácido por validação cruzada com parâmetros ponderados pela vizinhança.</i> .....	100
Figura 31 – <i>Desempenho de cada classificador específico de aminoácido por validação cruzada com parâmetros ponderados pela vizinhança e com conservação de aminoácidos.</i> .....	102
Figura 32 – <i>Comparação entre o classificador Sting-nn-WNA, desenvolvido nesta tese, e outros métodos disponíveis na literatura utilizando o mesmo conjunto de dados para teste (35Enz).</i> .....	103
Figura 33 – <i>Avaliação do classificador por rede neural em relação ao docking benchmark, utilizando as taxas de desempenho acurácia (acerto), precisão, sensibilidade e mérito, para as classes de baixa (a) e média e alta dificuldade (b).</i> .....	104
Figura 34 - <i>Predição de IFR para quatro exemplos do docking benchmark.</i> .....	106

# Lista de Tabelas

Tabela 1 – Códigos do PDB e respectivas cadeias retiradas do docking benchmark e utilizadas para comparação com outros métodos disponíveis na literatura. Os IFR são registrados através de mapeamento das estruturas dos complexos nas estruturas isoladas. ....	16
Tabela 2 - Descritores utilizados nas análises estatísticas e criação de modelos classificadores e suas abreviações em inglês. ....	19
Tabela 3 - Escala de Hidrofobicidade definida por Radzicka et al. (1988) e características físico-química dos 20 tipos de aminoácido. ....	22
Tabela 4 – 14 tipos diferentes de contatos armazenados no STING_DB e seus respectivos valores de energia de interação.....	23
Tabela 5 – Avaliação dos descritores físico-químicos e estruturais da base de dados BlueStar STING em relação a normalidade de distribuição dos valores e correlação linear entre si. ....	63
Tabela 6 – Lista de variáveis que obtiveram maiores valores de valor p (acima de 1%) no teste de normalidade de D’Agostino, separados de acordo com o tipo de aminoácido. ....	66
Tabela 7 - Resultados dos testes estatísticos univariados para cada um dos descritores do resíduo de aminoácido alanina relativo ao seu poder de distinguir IFRs dos FSRs. ....	67
Tabela 8 – Medida de desempenho dos modelos lineares para previsão dos IFR, para cada um dos 20 tipos de aminoácido em termos de coeficiente de determinação, coeficiente de determinação ajustado e valor p. ....	75
Tabela 9 – lista dos dez últimos descritores removidos no processo de seleção de variáveis no modelo envoltório usando a taxa de mérito como critério de poda. ....	91

# Lista de Quadros

Quadro 1 – comando em R para cálculo das matrizes de MANOVA.....	69
Quadro 2 – comando em R para cálculo dos testes estatísticos multivariados da MANOVA. .....	69
Quadro 3 – análise por regressão linear .....	74

# Lista de Abreviações

Acc – acessibilidade  
Ala – alanina  
ANOVA – análise de variância  
Arg – arginina  
Asn – asparagina  
Asp – ácido aspártico  
AUC – área abaixo da curva  
BFGS – Broyden–Fletcher–Goldfarb–Shanno  
CAPRI – *Critical Assesment of Predicted Interactions*  
CED – densidade de energia de contatos  
CLO – cross link order  
CPO – *cross presence order*  
Cys – cisteína  
DNA – ácido desoxirribonucleico  
DS95 – banco de dados não redundante com no máximo 95% de similaridade sequencial  
DS70 – banco de dados não redundante com no máximo 70% de similaridade sequencial  
DS30 – banco de dados não redundante com no máximo 30% de similaridade sequencial  
EP – potencial eletrostático  
FN – falso negativo  
FP – falso positivo  
FSR – *Free Surface forming Residue*  
GLM – Teoria linear generalizada  
Gln – glutamina  
Glu – ácido glutâmico  
Gly – glicina  
His – histidina  
IFR – *Interface Forming Residue*  
Ile – isoleucina  
IQR – *interquartile range*  
KKT – Karush-Kuhn-Tucker  
LDA – análise discriminante linear  
Leu – leucina  
LHA – Last Heavy Atom  
Lys – lisina  
MANOVA – análise de variância multivariada  
MCC – coeficiente de correlação de Matthew



Met – metionina  
MLP – *multilayered perceptron*  
MM – Main chain – Main chain  
MS – Main chain – Side chain  
nn – rede neural  
PCA – análise de componentes principais  
PISA – *Protein Interactions, Surfaces and Assemblies*  
Pro – prolina  
RMN – ressonância magnética nuclear  
RMSD – Desvio quadrático médio  
RNA – ácido ribonucleico  
ROC – *Receiver Operating Characteristics*  
Ser – serina  
SCOP – *Structural Classification Of Proteins*  
SS – Side chain – Side chain  
SVM – *Support Vector Machine*  
Thr – treonina  
TN – verdadeiro negativo  
TP – verdadeiro positivo  
Trp – triptofano  
Tyr – tirosina  
Val – valina  
WNA – *weighted neighbor averages* (descritores ponderados pela vizinhança)  
W – uma molécula de água  
WW – duas moléculas de água

# Resumo

Dentro do ambiente celular, há uma variedade de moléculas e a interação entre si regulam praticamente todos os processos necessários e essenciais para a manutenção da vida. Interações entre proteínas estão envolvidas no controle de vários processos intra e intercelulares, como regulação metabólica e da expressão gênica, reconhecimento antígeno-anticorpo etc. que definem as características biológicas do funcionamento da vida entre os diversos organismos. Ao conhecer a interface de interação de uma proteína chave para desenvolvimento de casos patológicos, é possível desenhar drogas com alta especificidade com o sítio de ligação. Para avançar nessa frente, o conhecimento da estrutura proteica é fundamental, porém não suficiente. É necessário conhecermos o sítio de ligação alvo para cada parceiro de interação. Este estudo visa entender as características do nano-ambiente das interfaces proteicas - área através da qual as macromoléculas se comunicam e exercem sua funcionalidade. Propomos utilizar uma abordagem de estudo das características físico-químicas e estruturais dos resíduos formadores de interfaces de complexos conhecidos e com estrutura quaternária resolvida experimentalmente, utilizando um conjunto de dados sem redundância sequencial, extraíndo os parâmetros/descriptores que descrevem de forma objetiva as diferentes classes de complexos, revelando as características principais sobre interações proteína-proteína. A finalidade deste trabalho é de conhecer os detalhes que definem uma área como interface e aplicá-lo em uma ferramenta preditiva para todas as proteínas com arranjo estrutural conhecido e/ou modelado. Propomos de forma pioneira, o uso de classificadores específicos para cada tipo de aminoácido e independente do uso de descritores sobre conservação de aminoácidos. Resultados obtidos com classificador linear e por *ensemble* de redes neurais destacam a nossa abordagem, desenhada e aplicada nesta tese, como uma com os melhores indicadores de desempenho na predição precisa dos resíduos de aminoácido na interface entre as abordagens descritas recentemente na literatura. Ainda, enquanto os outros métodos dependem de descritores sobre conservação de aminoácidos, é mostrado aqui que nenhum ganho de desempenho é obtido com a incorporação de tais descritores em nosso modelo classificador. Esse resultado indica que o uso de descritores puramente físico-químicos e estruturais é suficiente para explicar o grau de conservação dos aminoácidos.

# Abstract

Inside cells, there is a variety of molecules and their interactions regulate virtually all necessary and essential processes to the maintenance of life. Interactions among proteins are involved in the control of several processes within and out of the cell, such as, metabolic and gene expression regulation, anti-body and antigen recognition, etc. that defines biological characteristics of life among many organisms. If the protein interface amino acids of a key protein related to a given pathologic phenomenon are known, it is possible to rationally design drugs with high specificity for a specific binding site. To gain insight in this field, the knowledge of the protein three-dimensional structure is mandatory, but not sufficient. It is also necessary to know the interface between the target protein and its partners. This study focuses in understanding the characteristics of the area through which the macromolecules communicate to each other and exercise their function. Here, it is proposed an approach to study the physicochemical and structural characteristics of the interface forming residues with known quaternary structure (experimentally solved). It was selected a sequence non-redundant dataset and by extracting parameters/descriptors, that objectively describe different complex classes, it was possible to unravel the basic characteristics of protein-protein binding. The goal of this study is to unravel the details that outline a specific area as interface and apply it in a form of a predictive tool for all proteins with known atomic structure. It is proposed by the first time, the use of amino acid specific classifiers regarding amino acid type and free of amino acid conservation attributes. The results obtained here by employing linear and *ensemble* of neural network classifiers show that, based on purely physicochemical and structural descriptors, it is possible to get precise predictions about interface forming residues in protein-protein assemblies. Comparatively, the method described here retains better performance indicators than the ones recently described in the literature. In addition, we showed that, for our method, adding “conservation” attributes does not induce any performance gain, which is a major difference if compared to other described methods. This result indicates the purely physicochemical and structural descriptors are sufficient to explain how conserved amino acids are.



# 1 Introdução

No interior da célula, o material genético é organizado na forma de dupla hélice do ácido desoxirribonucléico (DNA). Como produtos do código genético, têm-se proteínas e ácido ribonucléico (RNA). Proteínas são sintetizadas no citoplasma a partir do RNA, sendo que apenas um gene pode dar origem a várias proteínas, ou ainda RNAs funcionais, aumentando assim o número e a complexidade do processo de produção de moléculas de interesse para a célula. Associados a outras pequenas moléculas (sais minerais, peptídeos, etc.) e macromoléculas (carboidratos, por exemplo), todos os componentes interagem regulando diversos processos, como transdução de sinais, regulação gênica, formação de estruturas mecânicas que conferem sustentabilidade para células, controle metabólico, respostas imunológicas etc. (Xenarios e Eisenberg, 2001; Ponstingl *et al.*, 2005; Reichmann *et al.*, 2007). A descrição e entendimento da base físico-química dos vários tipos de interações entre proteínas é um passo crítico para o desenvolvimento de uma análise da genômica funcional.

Recentemente, a comunidade científica tem voltado sua atenção para a biologia de sistemas. Essa nova frente de pesquisa visa entender o funcionamento de vias de regulação (gênica e/ou metabólica) ou mesmo organismos inteiros (Alloy e Russell, 2006). O grande avanço em técnicas de sequenciamento de DNA forneceu, para diversos organismos, uma lista de genes presentes em cada organismo. Na era pós-genômica, um grande esforço vem sendo feito para avançar no conhecimento baseado nas sequências obtidas, em especial, entender como os produtos dos genes interagem entre si. Para isso, o mapeamento das interações proteína-proteína é fundamental para o estudo do funcionamento e a função fisiológica das redes de regulação formadas. Proteínas interagem entre si com afinidades diferentes que variam de poucos milimolares até alto grau de afinidade da ordem de fento-molares, sendo a maior parte das reações altamente específicas (Reichmann *et al.*, 2007). A capacidade de proteínas interagirem com diversos parceiros simultaneamente e exclusivamente aumenta a complexidade e robustez das redes de interações (Alloy e Russell, 2006), sendo assim essenciais para o desenvolvimento, diferenciação celular e homeostase. O estudo das diversas interações em uma via, ou do organismo como um todo recebe o nome de biologia de sistemas.

Entender completamente a interação entre os diversos parceiros proteicos depende criticamente do conhecimento estrutural das moléculas de interesse (Alloy e Russell, 2006). Ao

ganharmos conhecimento estrutural sobre todos os componentes de interesse, diversos sistemas biológicos de alta complexidade podem ser estudados e entendidos, uma vez que o organismo é mais complexo do que a soma de todos os componentes. Cada componente atua de forma bem específica em uma ou poucas partes de todo o processo de regulação, com funções cíclicas, em módulos de alta complexidade e com consumo de energia (Russell *et al.*, 2004). Esses processos resultam em respostas coerentes para ações específicas (Kitano, 2002).

Alloy e Russell (2006) afirmam que atualmente é muito improvável encontrar uma proteína cuja informação estrutural não esteja disponível ou não possa ser modelada por técnicas de modelagem por homologia. Essa afirmação, que data de 2006, parece bem certa, como pode ser observado através da busca avançada no banco de dados sobre estruturas de proteínas, o *Protein Data Bank* (PDB; Berman *et al.*, 2000), que desde 2009 não há um novo (diferente) arranjo estrutural, segundo a definição do software SCOP (do inglês *Structural Classification Of Proteins*; em Julho de 2012 há 1393 arranjos diferentes segundo a mesma classificação).

Diversas técnicas experimentais podem ser utilizadas para se obter informação sobre a interação direta ou indireta de duas ou mais proteínas (Xenarios e Eisenberg, 2001). Ainda que experimentos relativamente simples, como duplo-híbrido em levedura, forneçam resultados interessantes do ponto de vista biológico, entender como tais interações ocorrem requer mais esforço. A determinação da estrutura tridimensional experimental de proteínas, e em especial de complexos entre proteínas, ainda consome grande quantidade de recursos econômicos e tempo, com várias limitações das técnicas utilizadas para obtenção de estruturas com resolução atômica: cristalografia e difração de raios-x e ressonância magnética nuclear. Outra técnica para empregada para a resolução de estruturas proteicas é a difração de nêutrons, porém menos de 0,1% das estruturas de entrada no PDB foram resolvidas com essa técnica experimental. Para cristalografia e difração de raios-x, a grande dificuldade está na obtenção de cristais com poder de difração suficiente para a geração de modelos com boa resolução (o limite sobre o que é considerado “boa resolução” é arbitrário, mas acima de 3 Å muitos detalhes sobre a posição relativa dos átomos são perdidos), enquanto que para ressonância magnética nuclear a limitação do tamanho das estruturas proteicas em torno de 30 kDa (limitação da técnica e não da qualidade de equipamentos atuais) se mostra como fator limitante, especialmente para o estudo de complexos entre proteínas.

Devido à grande importância das redes de interações proteína-proteína para a manutenção dos processos regulatórios dentro e fora das células, diversas doenças podem ser atribuídas a

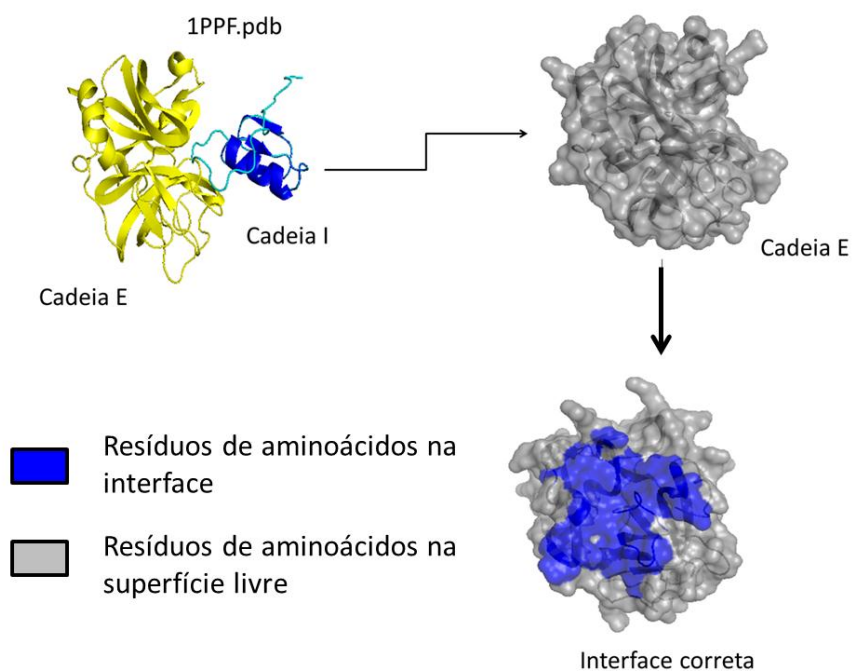
formação ou inibição de complexos proteicos em locais e/ou momentos inadequados. Por isso, nos últimos anos, grande esforço da comunidade científica é voltado para estudos com finalidade de desenho racional de drogas, cujo conhecimento dos contatos estabelecidos nas interfaces de complexos proteicos importantes em problemas envolvidos com a saúde humana, interações planta-patógeno etc. são utilizados para a determinação de pequenas moléculas que consigam interagir de forma específica com tais alvos, fundamentais para o aparecimento dessas doenças e no desenvolvimento de novas drogas (Lybrand, 1995; Beeley e Duckworthy, 1996; Parrill, 1996; Wade, 1997; Zsoldosa *et. al.*, 2003; Acharya *et. al.* 2011).

Entender como as proteínas se inter-relacionam, de como os arranjos atômicos se organizam, pode ser utilizado em experimentos de *docking* (recurso computacional que diz respeito à técnica de acoplamento de estruturas tridimensionais de duas ou mais proteínas, macromoléculas e ligantes), dinâmica molecular etc. ou até mesmo servir de guia para experimentos laboratoriais, por exemplo, indicando alvos para experimentos de mutagêneses, resíduos de aminoácido importantes em relação a contatos e características de área e volume que devem ser preservadas ou mudadas.

A capacidade de prever como duas proteínas interagem é um importante problema abordado pela bioinformática e biologia computacional, e testada na competição internacional CAPRI (na sigla em inglês para *Critical Assesment of Predicted Interactions*; Janin e Wodak, 2007). Nessa competição, estruturas de monômeros (ou homo-oligômeros) são disponibilizadas e algoritmos de diversos grupos de pesquisa são testados sobre a qualidade do modelo do complexo gerado e como esse desvia em relação ao complexo resolvido de forma experimental e não disponibilizado. Apesar de várias metodologias serem bem sucedidas em alguns alvos (Janin e Wodak, 2007), nenhuma se mostrou eficaz para todos os alvos. O elevado número de modelos não aceitáveis de complexos preditos (2114 de um total de 2390) mostrados nas últimas rodadas do CAPRI (Janin e Wodak, 2007) indica que os métodos empregados ainda não conseguem prever de forma acurada complexos proteicos, dadas as estruturas tridimensionais de cada um dos constituintes. Essa informação evidencia que nenhuma lei geral de interação entre proteínas foi obtida até o momento, e, portanto, nenhuma metodologia pode ser proposta de forma a incorporar as informações necessárias para a geração de modelos confiáveis.

Vários estudos vêm sendo reportados na literatura sobre métodos que tentam entender a natureza de tal associação entre macromoléculas (Chotia e Janin, 1975; Jones e Thornton, 1996, 1997; Sheinerman *et. al.* 2000; Xenarios, I. e Eisenberg, D. 2001; Ofran e Rost, 2003; Bahadur *et. al.*

2004; Ponstingl *et al.*, 2005; Gruber *et. al.* 2006; Tsuchiya *et al.*, 2006; Chen e Skolnick, 2008), e, apesar do grande avanço de ferramentas de biologia computacional, ainda não é possível entender totalmente como tal interação ocorre, de forma que a predição de estruturas atômicas de complexos proteicos ainda é um problema em aberto. Para fins de classificação, diversas metodologias de aprendizado de máquina e estatístico foram propostas na literatura. Essas técnicas combinam diversos descritores de entradas a mapeiam regras e funções que levam cada conjunto de dados de cada resíduo de aminoácido a uma classe ou, ainda, associam probabilidades de classificação. A figura 1 ilustra de forma o problema em questão para a proteína elastase de leucócitos humanos e o terceiro domínio do inibidor de ovomucóide de peru (código PDB 1PPF). Dado proteínas que interagem entre si com estrutura tridimensional conhecida, buscamos diferenciar os resíduos de aminoácido da interface em relação aqueles que estão na superfície livre.

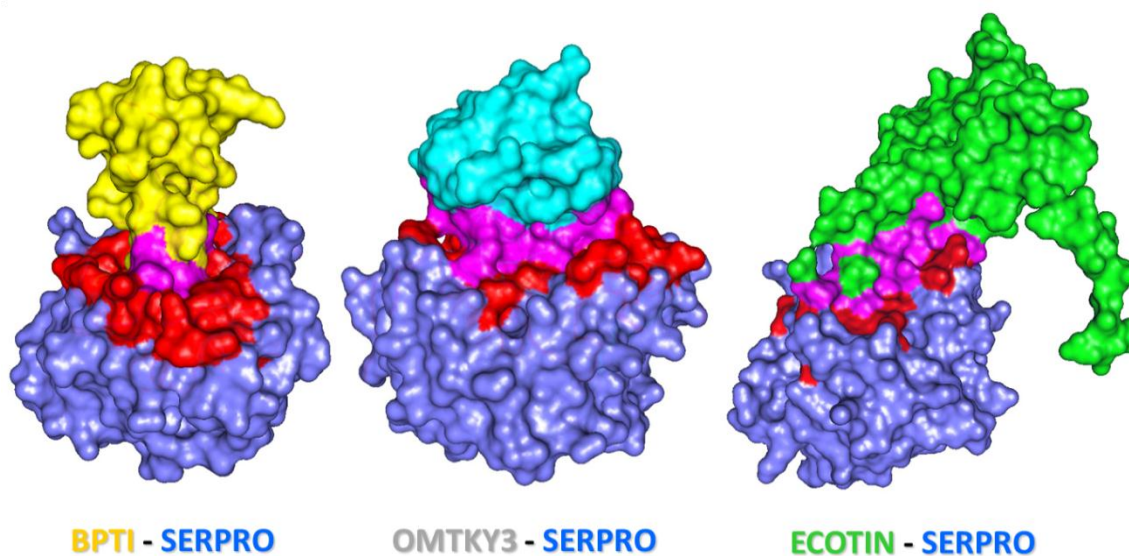


**Figura 1 – Ilustração da região de interface em um complexo proteína-proteína com estrutura tridimensional conhecida.**

*Estrutura da proteína elastase de leucócitos humanos (cadeia E) e o inibidor de ovomucóide de peru (cadeia I). Ao extrair a estrutura da cadeia E, os resíduos de aminoácidos que estavam em contato com resíduos de aminoácidos da cadeia I são pintados em azul, formando uma região de interface entre as duas proteínas. O objetivo de metodologias de classificação de aminoácidos formadores de interface buscam prever corretamente os resíduos de aminoácido da interface e diferenciá-los em relação aos resíduos de aminoácido da superfície livre (pintados em cinza).*



Uma mesma proteína pode possuir diversos parceiros de interação. Ainda, é possível que a área de interface seja compartilhada entre uma mesma proteína e seus diversos parceiros. Isso é ilustrado na figura 2 que mostra a proteína serino-protease (SERPRO) com três distintos inibidores. Os três inibidores compartilham a mesma área de interface, no entanto, os detalhes dos contatos entre os resíduos de aminoácido inter-cadeias são específicos em cada caso e conferem a especificidade da proteína com cada inibidor (Ribeiro *et al.*, 2010). A presente tese não busca estudar a questão da especificidade, apenas a predição da área de interface, uma vez que os detalhes de interação são específicos entre os parceiros de cada complexo.



**Figura 2 – Estrutura do complexo entre a proteína serino-protease com três inibidores diferentes que compartilham a mesma região de interface.**

*A mesma área de interface da proteína serino-protease (SERPRO) é compartilhada pelos inibidores BPTI, OMTKY3 e ECOTIN. Os detalhes de interação, como pares de contatos entre os resíduos de aminoácido inter-cadeias, conferem a especificidade das interações.*

Recentemente, Zhou e Qin (2007-a) realizaram uma comparação entre seis métodos, disponíveis na literatura como servidores on-line, utilizando um mesmo conjunto de dados: ProMate (Neuvirth *et al.*, 2004), PPI-Pred (Bradford e Westhead, 2005), PINUP (Liang *et al.*, 2006), SPPIDER (Porollo e Meller, 2007), cons-PPISP (Chen e Zhou, 2005) e Meta-PPISP (Qin e Zhou, 2007).

O software ProMate (Neuvirth *et al.*, 2004) está implementado como um servidor on-line, que usa uma função linear baseado em descritores estruturais, como distribuição de pares de resíduos de aminoácido e estrutura secundária, propensão de tipos de aminoácidos em estar

na interface de complexos proteicos e conservação da sequência de resíduos de aminoácido. O único parâmetro físico químico utilizado no software ProMate, diz respeito ao descritor de hidrofobicidade. A forma de predição é a formação de grupos de resíduos de aminoácido (ou *patches*) com maior pontuação do modelo linear. Uma predição é considerada bem sucedida pelos autores, quando pelo menos metade do grupo predito como interface está realmente localizado na interface. Esse método foi treinado em um conjunto de dados de hetero-complexos transientes e, usando o critério de sucesso definido pelos autores, atingiu 70% de sucesso na predição de interface em complexos proteína-proteína.

O software PPI-Pred (Bradford e Westhead, 2005) usa o modelo de máquinas de vetores suporte (SVM na sigla em inglês para *Support Vector Machine*) com descritores estruturais, físico-químico e de conservação de sequência de aminoácidos. Combinando os descritores de potencial eletrostático, hidrofobicidade, conservação de aminoácidos, propensidade de localização na interface, forma de superfície e acessibilidade ao solvente, PPI-Pred prevê regiões de interface entre proteínas com o mesmo mecanismo de *patch* descrito para o software ProMate. Uma predição bem sucedida, como descrito no artigo pelos autores, é tida quando pelo menos um dos três melhores *patches* possui 20% de cobertura da região de interface com 50% de especificidade (definida pelos autores como o número de aminoácidos de interface no *patch* dividido pelo número total de resíduos de aminoácido no *patch*). Com essa métrica, os autores mostram predições bem sucedidas em 76% dos 180 complexos proteicos no banco de dados utilizado.

O software PINUP (Liang *et. al.*, 2006) também utiliza uma função linear com os descritores de conservação de aminoácidos, propensidade de localização na interface e um fator de energia da cadeia lateral. Esse último termo é composto de descritores para acessibilidade ao solvente, ligações de hidrogênio, sobreposição de volume atômico, energia de interação eletrostática, área de átomos não polarizáveis e polarizáveis, que não são acessíveis ao solvente, propensidade de ângulos da cadeia lateral e energia livre de solvatação. Utilizando uma metodologia de validação cruzada, PINUP atinge 44,5 % de taxa de acerto com 42,2% de cobertura.

O software SPPIDER (Porollo e Meller, 2007) utiliza uma rede neural com os descritores de número de contatos, hidrofobicidade, conservação de carga, conservação de hidrofobicidade, conservação de tamanho, conservação de aminoácidos e área acessível ao solvente prevista, ou seja, o valor real de área acessível ao solvente presente na estrutura não é utilizado, mas sim uma ferramenta preditiva de área acessível baseado na sequência de aminoácidos. Esse trabalho também introduziu o conceito de descritores ponderados pela vizinhança, utilizado nesse estudo e

descrito na seção 2.6. Utilizando um conjunto de teste com 149 entradas não redundantes com o conjunto de treinamento, os autores conseguiram prever os resíduos de aminoácido formadores de interface com uma taxa de acerto de 72%, cobertura de 60% e confiabilidade 64%.

O software cons-PPISP (Chen e Zhou, 2005) também utiliza uma rede neural para prever resíduos de aminoácido formadores de interface, sendo que os dados de entrada são o perfil de sequência do PSI-Blast (dados de conservação de aminoácidos) e a área acessível ao solvente. Utilizando o conjunto de teste com 68 cadeias proteicas de 40 complexos proteína-proteína com estrutura monomérica e em complexo conhecidas (porção do *docking benchmark*, seção 2.3), cons-PPISP atingiu desempenho de 42% de taxa de acerto, com uma cobertura de 38%.

O software Meta-PPISP (Qin e Zhou, 2007) utiliza o resultado de saída de três servidores: cons-PPISP, ProMate e PINUP. Baseado nos valores de saída sobre a probabilidade de classificação de cada resíduo de aminoácido como formador de interface, os autores construíram um modelo de regressão linear que supera os resultados dos três métodos isolados. Essa abordagem de comitê de classificadores, ou *ensemble*, permite que aminoácidos que falham em sua predição por um método, possam ser corretamente predito pelos outros métodos, diminuindo assim a probabilidade de errar cada uma das classificações. Meta-PPISP aumentou a confiabilidade de cada um dos métodos em 21,7 pontos percentuais (cons-PPISP), 52,5 pontos percentuais (ProMate) e 9,6 pontos percentuais (PINUP).

No Laboratório de Biologia Computacional da Embrapa Informática Agropecuária foi desenvolvido o maior banco de dados de descritores físico-químicos, estruturais e biológicos sobre estruturas proteicas, STING\_DB, e disponibilizado pelo servidor on-line que se encontra em sua versão BlueStar STING ([www.cbi.cnptia.embrapa.br/SMS](http://www.cbi.cnptia.embrapa.br/SMS), Neshich *et al.*, 2005 e 2006, Oliveira *et al.*, 2007). BlueStar STING é uma suíte de programas com ferramentas para a visualização e análise estrutural de proteínas. Estes programas (módulos) estão concentrados em um único pacote que visa oferecer um instrumento para estudos das macromoléculas, suas estruturas e as relações estrutura-função. Informações como posição dos resíduos de aminoácido na sequência e na estrutura, busca de padrões, identificação de vizinhança, ligações de hidrogênio, ângulos e distâncias entre átomos, são facilmente obtidas, além de dados sobre natureza e volume dos contatos atômicos inter e intra-cadeias, a conservação e relação entre os contatos intra-cadeia e parâmetros funcionais.

De todas as 97 classes, com mais de 900 descritores, utilizaremos um conjunto menor com os parâmetros tidos como mais importantes para o estudo de reconhecimento de interfaces em

complexos proteína-proteína. Com esses descritores aplicados a um banco de dados com complexos proteicos, podemos analisar a capacidade dos diversos parâmetros físico-químicos e estruturais em distinguir os resíduos de aminoácido importantes para a formação dos complexos observados, em relação aos resíduos de aminoácido que não oferecem nenhuma participação direta na estipulação de suas interfaces.

A relevância do assunto introduzido às ciências biológicas e a falta de entendimento sobre as características importantes que levam macromoléculas a interagirem serviram como motivação para a proposta desse estudo. Em especial, a falta de integração entre estudos que visam explicar as características das interfaces em complexos proteicos, ou seja, como os resíduos de aminoácido formadores de interface se distinguem dos resíduos de aminoácido de superfície livre, atrelado à criação de uma ferramenta preditiva para uso prático, indicam uma forma metodológica ainda a ser explorada. Ainda, ao se buscar um conjunto de descritores físico-químico e estruturais que descrevem o nano-ambiente dos resíduos de aminoácido de interface, o uso dos descritores de conservação de aminoácido deve ser evitado, uma vez que esse atributo não pode ser medido na estrutura proteica. O fato de todos os softwares descritos previamente, além de outros disponíveis na literatura e de nosso conhecimento, fazerem uso de descritores sobre conservação, revelam outro espaço para investigação que ainda não foi explorado: a criação de modelos classificadores de resíduos de aminoácido na interface que independem do uso de descritores de conservação.

Todos os seis métodos, descritos previamente, utilizam descritores de conservação de aminoácidos, assim como os descritores sobre propensão de localização na interface ou superfície livre. Apesar de resíduos de aminoácido importantes para o estabelecimento de interfaces entre proteínas serem mais conservados do que os resíduos de aminoácido que não são relevantes, é possível encontrar estruturas de proteínas sem função conhecida e com baixa homologia com qualquer outra proteína estudada. Para esses casos, métodos de predição de região de interface com o uso de conservação de aminoácidos não devem retornar resultados confiáveis. Ainda, ao tentar entender as características básicas que definem regiões de interações entre proteínas, o uso de descritores sobre conservação de aminoácidos apaga a informação sobre o nano-ambiente de cada aminoácido, uma vez que esse descritor não é medido na estrutura proteica, mas sim de um conjunto de proteínas.

A diferença de perspectiva referente ao uso do descritor de conservação de aminoácidos consiste no fato de que os métodos descritos acima buscam apenas a criação de um modelo

preditivo. Outros trabalhos disponíveis na literatura descrevem estudos estatísticos sobre as diferenças entre resíduos de aminoácido da interface de complexos proteína-proteína, e como esses se diferenciam dos resíduos de aminoácido da superfície livre, porém não aplicam essa informação em casos práticos de predição. Dessa forma, nosso estudo tende a combinar a descrição das características fundamentais que diferenciam os resíduos de aminoácido de interface dos de superfície livre ao mesmo tempo em que classificadores são treinados no estabelecimento de uma ferramenta preditiva. Para isso, um processo de seleção de variáveis foi incorporado no melhor modelo classificador encontrado.

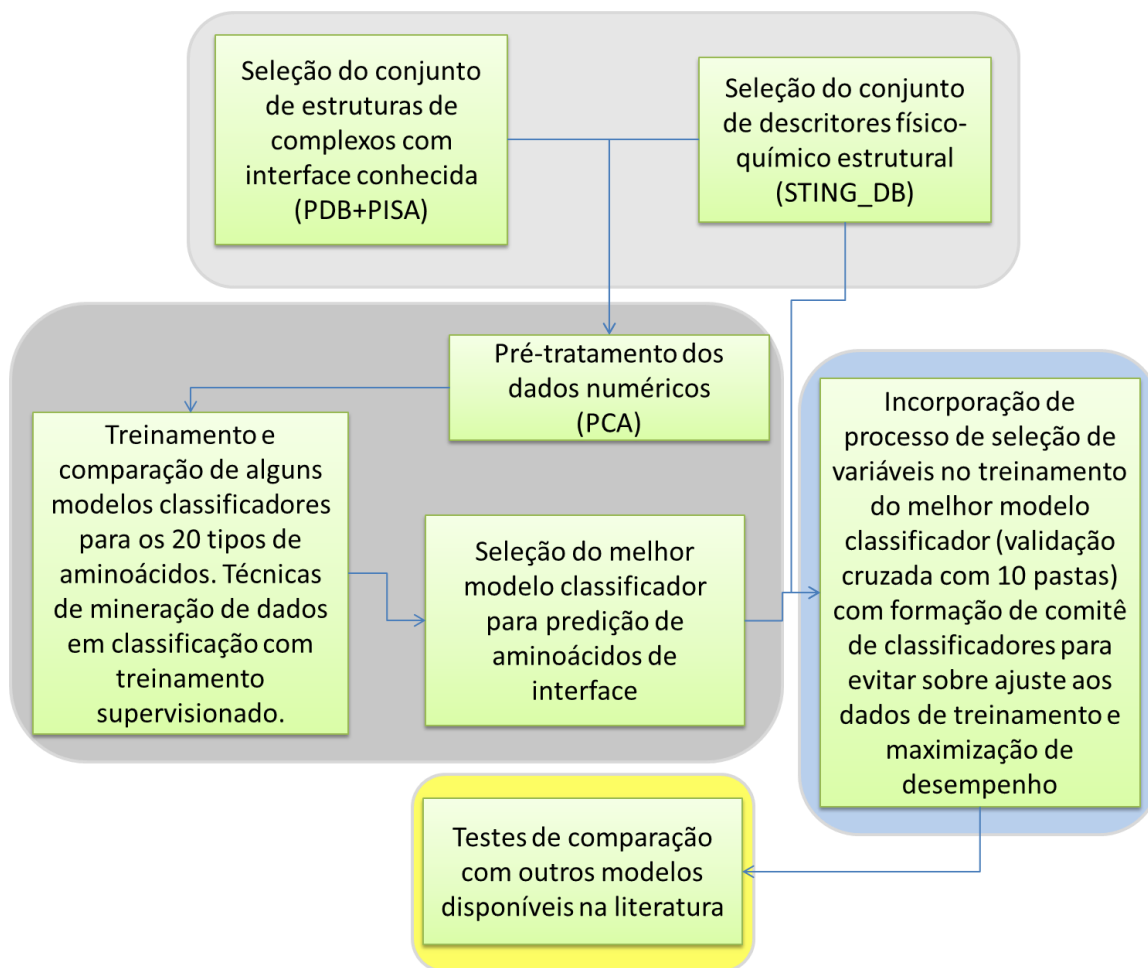
Propomos também o uso de classificadores específicos para tipo de aminoácido, sendo que a interface total é obtida com o agrupamento de todos os vinte classificadores. Esse processo pode substituir os descritores de propensão de localização dos resíduos de aminoácido na interface, e, como mostrado na seção de resultados, ao reduzir a faixa de valores para alguns descritores, é observado ganho de desempenho.

O problema estudado consiste em classificar corretamente resíduos de aminoácido, baseados em um conjunto de descritores físico-químico e estruturais, em duas classes: resíduos de aminoácido formadores de interface (IFR, na sigla em inglês para *Interface Forming Residue*) ou de superfície livre (FSR, na sigla em inglês para *Free Surface forming Residue*). Utilizando métodos de mineração de dados, buscamos responder as perguntas:

- 1) As características necessárias para estabelecimento de interações proteína-proteína estão presentes na estrutura proteica?
- 2) É possível determinar quais são os resíduos de aminoácido formadores de interface e distingui-los dos resíduos de aminoácido de superfície livre?
- 3) Quais características físico-químicas e estruturais diferenciam os resíduos de aminoácido formadores de interface dos resíduos de aminoácido da superfície livre da estrutura proteica?

A figura 3 ilustra resumidamente o fluxo de passos desenvolvidos neste estudo, desde a coleta de complexos proteicos com interface conhecida, seleção de descritores físico-químicos e estruturais, exploração de modelos de mineração de dados para a classificação com aprendizado supervisionado dos complexos selecionados, comparação e seleção do melhor modelo avaliado para cada um dos 20 tipos de aminoácidos, estabelecimento de um método para seleção de

variáveis para descrição do nano-ambiente dos resíduos de aminoácido formadores de interface e formação de comitê de classificadores para redução de sobre ajuste aos dados de treinamento. Por fim, o classificador final de resíduos de aminoácido formadores de interface é comparado com os modelos descritos previamente, que são inespecíficos quanto ao tipo de aminoácido e utilizam descritores de conservação de aminoácidos.



**Figura 3 – Fluxograma com os procedimentos para análise, tratamento dos dados e maximização de desempenho.**

*Fluxo desde a coleta dos dados até a comparação do melhor modelo obtido com as metodologias disponíveis na literatura.*

## 2 Metodologia

### 2.1 *Seleção das estruturas dos complexos proteicos do PDB e PISA*

Todas as informações em relação à disposição dos átomos dos complexos estudados foram retiradas de bancos de dados públicos. O PDB concentra todas as estruturas tridimensionais de proteínas (e complexos entre proteínas e outras moléculas) que foram resolvidas experimentalmente, e, portanto, tornou-se uma das fontes mais importantes para a realização dessa pesquisa.

No entanto, estudos reportam que algumas entradas no PDB não passam por filtros rigorosos de confiabilidade das informações presentes (Xu *et al.* 2008). Em várias entradas, os conceitos de unidade assimétrica e unidade biológica são confundidos. Unidade assimétrica é definida como o arranjo mínimo da estrutura proteica capaz de gerar o cristal, enquanto unidade biológica é o complexo mais provável de acontecer *in vivo* por ser funcional e com desempenho otimizado. Um exemplo de um caso extremo da diferença entre unidade assimétrica e biológica pode ser representado por estruturas de capsídeos de vírus que formam complexos podendo chegar a 900 moléculas interagindo entre si (900-mero), enquanto que a unidade assimétrica no PDB para a entrada 2btv.pdb (estrutura do complexo proteico do núcleo do vírus *bluetongue*) é de apenas 15 moléculas (15-mero).

Como ilustrado na figura 4 e descrito nesta seção, o banco de dados PISA (da sigla em inglês para *Protein Interfaces, Surfaces and Assemblies*; Krissinel e Henrick, 2007), foi utilizado em conjunto com o PDB, uma vez que é considerado mais confiável em relação à qualidade dos dados sobre complexos proteicos (Xu *et al.* 2008). A partir dos complexos disponíveis no PDB, o banco de dados PISA estabelece cálculos teóricos sobre a estabilidade dos complexos definindo interfaces mais acuradas que as disponíveis no PDB.

No entanto, a unidade assimétrica é igual à unidade biológica em muitos casos. Para eliminar possíveis problemas relacionados a este filtro de escolha do banco de dados, resolvemos selecionar apenas as entradas em que a unidade assimétrica do PDB está de acordo com o estado oligomérico segundo o PISA. Dessa forma, deixamos de selecionar estruturas de complexos que estão erroneamente classificados no PDB e ganhamos confiabilidade sobre o verdadeiro estado oligomérico das entradas do conjunto de dados.

Mesmo sendo uma etapa preliminar, sobressaltamos que a escolha do conjunto de estruturas proteicas é uma das etapas mais importantes para a realização desse projeto. As etapas de análise e posterior discussão dos resultados dependem criticamente e pode ser facilitada devido a uma boa escolha do espaço amostral.

A figura 4 ilustra todas as etapas e filtros utilizados para a escolha dos complexos representativos da interação entre proteínas. Os filtros estabelecidos seguem em parte os filtros de outros trabalhos publicados, e discutidos posteriormente, sobre previsões de interfaces em complexos proteicos.

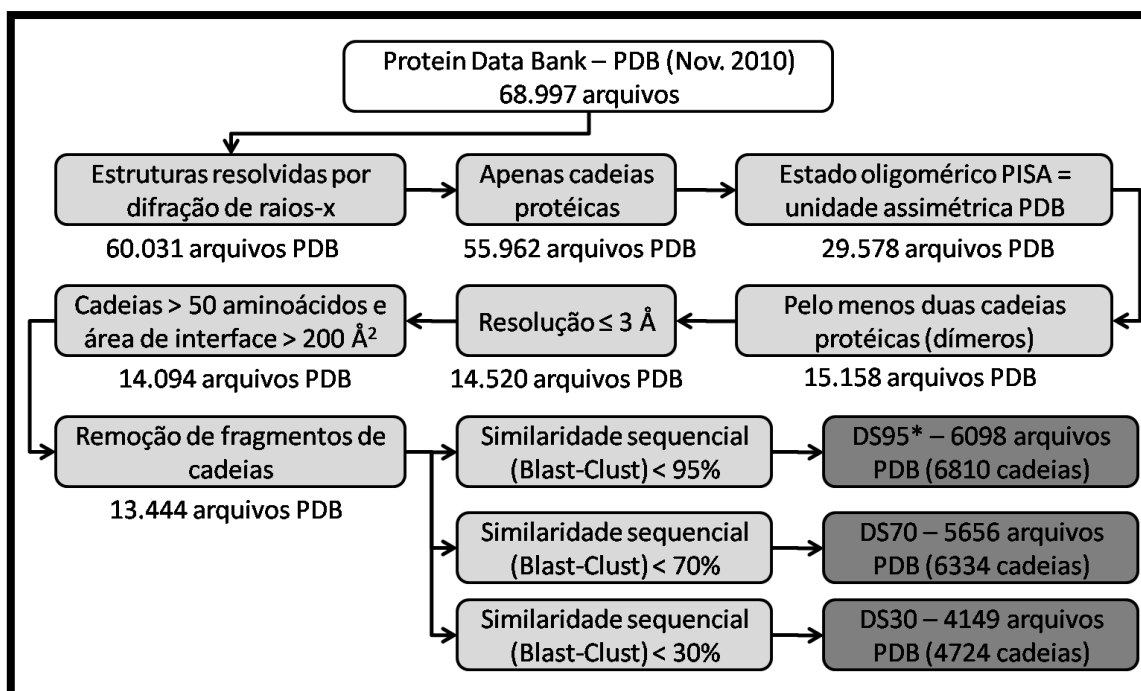
Primeiramente, havia 68.997 entradas PDB, em novembro de 2010. Os filtros utilizados consistem em escolher apenas estruturas resolvidas por cristalografia e difração de raios-x, uma vez que essa restrição refere-se à maioria dos dados presentes no PDB (apenas 8.966 não seguem esse critério). Esse filtro foi aplicado devido à observação que há diferenças entre estruturas proteicas resolvidas por cristalografia e difração de raios-x e resolvidas por ressonância magnética nuclear (RMN), principalmente na diferença de contatos estabelecidos entre resíduos de aminoácido (Schneider *et al.*, 2009). Além do mais, estruturas resolvidas por RMN possuem em média 20 estruturas equiprováveis e listadas no arquivo PDB, sendo que não é possível escolher a melhor estrutura para os cálculos dos parâmetros físico-químicos e estruturais.

Nesse estudo, não é de interesse avaliar as interações entre proteínas e DNA e/ou RNA, por isso as cadeias desse tipo de macromolécula foram excluídas. Como descrito acima, foram utilizadas apenas as entradas que estão em acordo entre os bancos de dados PDB e PISA.

Apenas dados provenientes de complexos proteicos foram selecionados (ao menos duas cadeias na unidade assimétrica), atingindo 15.158 PDBs. Esse filtro elimina entradas monoméricas e reduz em aproximadamente 50% o número de entradas (de 29.578 para 15.158 PDBs). A obtenção de estruturas na forma isolada (monomérica) se torna mais simplificada em relação à baixa energia de ligação entre complexos transientes, o que não favorece a formação de cristais (Liang, *et al.* 2006). Quanto menor o valor da resolução da estrutura, melhor consegue-se definir as posições dos átomos e melhor é o modelo estrutural. Seguindo outros trabalhos disponíveis na literatura, escolhemos arbitrariamente o máximo como sendo de 3 Å de resolução. Da mesma forma, foi definido um limite inferior para o número de resíduos de aminoácido presentes em cada cadeia como sendo de 50 resíduos. Para eliminar complexos cristalinos (que não possuem nenhuma relevância biológica) retiramos entradas com área de interface menor que  $200\text{\AA}^2$  (Ponstingl *et al.* 2005). Para eliminar possíveis dados introdutórios de ruído em nossa análise, as



entradas classificadas como fragmentos de proteínas pelo banco de dados UnitProt (Martin 2005; UnitProt consortium, 2009) foram removidas.



**Figura 4 – Filtros estabelecidos para a seleção de complexos proteicos oriundos do PDB.**

Na data de início da seleção havia um total de 68.997 estruturas depositadas no PDB. Escolhemos apenas estruturas resolvidas por cristalografia e difração de raios-X (60.031 entradas), removendo cadeias de DNA ou RNA (55.962 entradas). O estado oligomérico definido pelo PISA deve ser igual ao número de cadeias na unidade assimétrica (29.578 entradas), sendo que deve haver pelo menos duas cadeias na unidade assimétrica (15.158 entradas). Escolhemos estruturas com resolução melhor do que 3 Å (15.520 entradas), com no mínimo 50 resíduos de aminoácido e com área de interface maior do que 200 Å<sup>2</sup> (14.094 entradas). Fragmentos de proteínas foram removidos segundo a integração do banco de dados Unitprot (13.444). A similaridade sequencial foi removida utilizando Blast-Clust da busca avançada do PDB, resultando em conjunto de dados não redundantes ao nível de 95% (6098 entradas), 70% (5656 entradas) e 30% (4149 entradas).

Por último, as cadeias de aminoácidos que compartilham mais do que 95%, 70% e 30% de similaridade sequencial foram removidas. Para definir qual entrada seria mantida, utilizamos o mesmo critério que o sistema de busca avançada do PDB, escolhendo os arquivos com melhor resolução, com entrada mais recente e por ordem alfabética.

## **2.2 Agrupamento do banco de dados**

A geração de modelos classificadores descrita nessa tese foi dividida em duas etapas neste estudo. A primeira etapa visa explorar qual a importância do modelo classificador para o processo de predição e classificação de resíduos de aminoácido como formadores de interface ou de superfície livre. Na segunda etapa, foi explorado, com maior rigor estatístico, possível viés referente à escolha do conjunto de treinamento do banco de dados, a melhora em desempenho obtida pela divisão do problema de classificação por tipo de aminoácido, a comparação com outros modelos disponíveis na literatura e, por fim, discussão da lista dos descritores mais importante relacionados com a distinção entre IFR e FSR.

Ao dividir o conjunto de dados por tipo de resíduo de aminoácido, nenhuma distinção entre cisteínas que estabelecem ou não estabelecem ligação de sulfeto é realizada. No entanto, utilizamos o descritor sobre energia de ligação de sulfeto para esse tipo de resíduo de aminoácido.

Para a primeira etapa, utilizando um gerador de números aleatórios, as entradas retidas no conjunto de dados DS95, descrito acima, foram separadas em dois grupos de dados que serão utilizados nas etapas posteriores: conjunto de *treinamento* (60% das entradas - 4114 cadeias não redundantes) e conjunto *teste* (40% - 2733 cadeias).

Para a segunda etapa, as entradas do banco de dados foram aleatoriamente divididas em 10 grupos. A técnica estatística de validação cruzada com 10 pastas foi utilizada. Nessa técnica, 9 grupos são utilizados para treinar o modelo classificador enquanto o último grupo é utilizado para teste. Na segunda rodada, outro grupo é selecionado para teste enquanto os 9 grupos restantes são utilizados para treinamento. Ao repetir o processo 10 vezes, variando o grupo de teste, têm-se uma medida estatística do desempenho do classificador, sendo que cada entrada é utilizada como teste uma única vez.

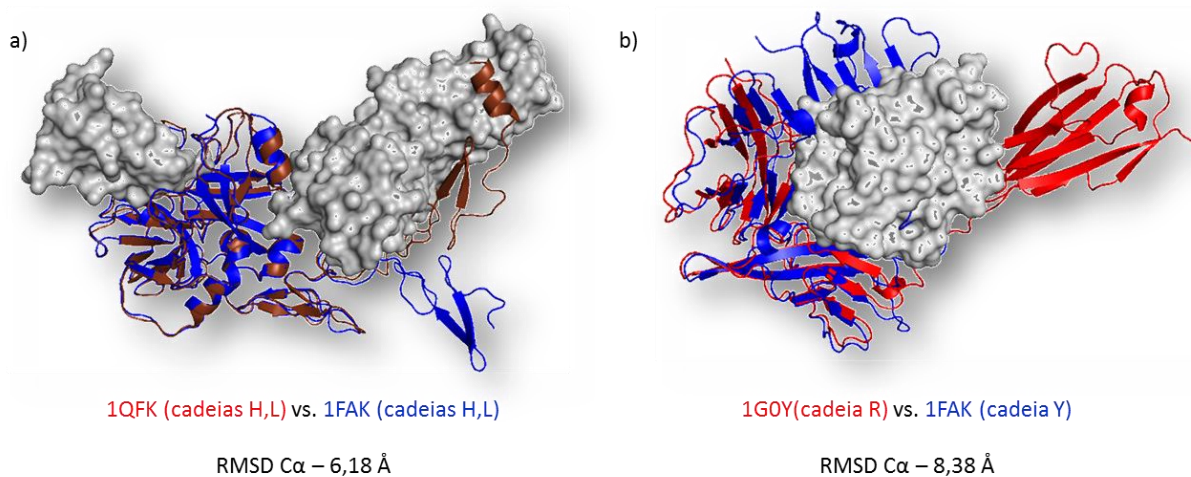
## **2.3 Conjunto teste final – Docking benchmark**

Para a área de predição de interface em complexos proteína-proteína, o banco de dados sobre *docking* proteína-proteína (Hwang *et al.*, 2010) é considerado como o *benchmark*, e encontra-se atualmente na quarta versão. Por definição, *benchmark* é um conjunto de dados que deve ser descartado durante treinamento e usados em predições para comparações entre técnicas diferentes. Nesse caso, as estruturas dos pares de proteínas que interagem foram resolvidas tanto para cada proteína isolada como também para as proteínas complexadas. Dessa forma, a predição

baseada na cadeia proteica isolada condiz mais com a realidade de predições futuras, onde a estrutura utilizada não se encontra ligada a um determinado parceiro.

Um total de 176 pares de proteínas compõe o *benchmark* (Hwang *et al.*, 2010), sendo divididos em 121 casos fáceis (corpo-rígido), 30 médios e 25 difíceis. Os graus de dificuldade estão relacionados com a reorganização das cadeias proteicas após se ligarem entre si. Para os indivíduos presentes na classe fácil, utilizamos a informação contida na estrutura do complexo, enquanto que para as demais classes a predição foi feita na estrutura isolada (não complexada). Essa atitude foi motivada pelo fato de que estruturas presentes na classe fácil sofrem pouca mudança conformacional e, portanto, as estruturas isoladas e complexadas são muito parecidas. Para os casos fáceis, o RMSD de carbono  $\alpha$  médio é de 0,82-Å (em média 212 resíduos de aminoácido por cadeia proteica, nesta classe). De forma comparativa, as classes de média e alta dificuldade possuem um RMSD de carbono  $\alpha$  de 1,8 e 3,6-Å, respectivamente (em média 239 e 235 resíduos de aminoácido por cadeia proteica, respectivamente).

A figura 5 ilustra a mudança conformacional sofrida pelas cadeias proteicas após a formação do complexo para dois casos do grupo “difícil”. Para a figura 5-a, as cadeias H e L da proteína do fator de coagulação VII-a (1QFK\_HL isolada e 1FAK\_HL em complexo) possui um RMSD (de carbono  $\alpha$ ) superior a 6 Å. Para a figura 5-b, as cadeias R e Y do receptor de interleucina-1 (1GOY\_R isolada e 1IRA\_Y em complexo) possuem um RMSD superior a 8 Å.



**Figura 5 – Ilustração da modificação conformacional sofrida por duas proteínas do grupo “difícil” do docking benchmark (versão 4).**

Para a proteína do fator de coagulação VII-a (a) o RMSD de carbono  $\alpha$  atinge 6,18-Å e para o receptor de interleucina-1 (b) atinge 8,38-Å. A forma em isolamento é ilustrada em azul enquanto a forma complexada em vermelho. A superfície proteica em cinza ilustra o parceiro de interação de cada proteína na forma em complexo.

Uma fração do *docking benchmark* foi utilizada recentemente para comparar 8 modelos classificadores de IFR descritos na literatura (Zhou e Qin, 2007-a). Para 35 casos do *benchmark* (conjunto de dados referido como 35Enz), os autores utilizaram a estrutura monomérica e mapearam a interface das estruturas dos complexos, respeitando a ordem da tabela 1.

**Tabela 1 – Códigos do PDB e respectivas cadeias retiradas do docking benchmark e utilizadas para comparação com outros métodos disponíveis na literatura. Os IFR são registrados através de mapeamento das estruturas dos complexos nas estruturas isoladas.**

PDB isolado	Cadeia isolada	PDB complexo	Cadeia complexo	PDB isolado	Cadeia isolada	PDB complexo	Cadeia complexo
1BA7	B	1AVX	B	1M08	B	7CEI	B
1A19	B	1AY7	B	1RGH	B	1AY7	A
1HOE	A	1BVN	T	1PIG	A	1BVN	P
1HPT	A	1CGI	I	9RSA	B	1DFJ	E
1K9B	A	1D6R	I	1E1N	A	1E6E	A
2BNH	A	1DFJ	I	1GJR	A	1EWY	A
1CJE	D	1E6E	B	4PEP	A	1F34	A
9PTI	A	1EAW	B	2PKA	X	1HIA	A
1CZP	A	1EWY	C	2PKA	Y	1HIA	B
1ECZ	A	1EZU	A	1J06	B	1MAH	A
1ECZ	B	1EZU	B	1UDH	A	1UDI	E
1F32	A	1F34	B	2BBK	J	2MTA	H
1BX8	A	1HIA	I	2BBK	M	2MTA	L
1FSC	A	1MAH	F	1CCP	A	2PCC	A
1LUO	A	1PPE	I	1SUP	A	2SIC	E
1B1U	A	1TMQ	B	1UNK	D	7CEI	A
2UGI	B	1UDI	I	1JB1	A	1KKL	A
2RAC	A	2MTA	A	1EGL	A	1ACB	I
1YCC	A	2PCC	B	2HPR	A	1KKL	H
3SSI	A	2SIC	I				

O uso de estruturas resolvidas pela técnica de RMN, traria informações sobre a flexibilidade das estruturas proteicas. No entanto, a presença de 20 estruturas equiprováveis por arquivo, inviabiliza o cálculo dos parâmetros físico-químicos e estruturais. Além disso, apenas 7% das entradas por RMN são de complexos entre duas ou mais cadeias proteicas. Somado a diferença entre os contatos estabelecidos por estruturas resolvidas por RMN e cristalografia e difração de raios-x (Schneider *et al.*, 2009), as entradas por RMN foram excluídas das etapas posteriores

deste estudo. Da mesma forma, o uso de outras técnicas computacionais para o estudo da flexibilidade da estrutura proteica, como dinâmica molecular, por exemplo, ainda não podem ser feitas em larga escala e de forma automatizada, sendo de grande importância apenas para estudos de casos que podem ser realizados posteriormente ao uso da ferramenta de predição de aminoácidos formadores de interface.

## **2.4 BlueStar STING e STING\_DB**

STING é uma suíte de programas com ferramentas para a visualização e análise estrutural de proteínas. Estes programas (módulos) estão concentrados em um único pacote que visa oferecer um instrumento completo para estudos das macromoléculas, suas estruturas e as relações estrutura-função. Informações como posição dos resíduos de aminoácido na sequência e na estrutura, busca de padrões, identificação de vizinhança, ligações de hidrogênio, ângulos e distâncias entre átomos, são facilmente obtidas, além de dados sobre natureza e volume dos contatos atômicos inter e intracadeias, a conservação e relação entre os contatos intracadeia, e parâmetros funcionais. Todo esse pacote de análise foi implementado em uma interface amigável ao usuário que pode ser acessada via web (<http://www.cbi.cnptia.embrapa.br/SMS>).

Associada aos módulos, uma extensa base de dados (STING\_DB) contém todos os dados que foram usados, enquanto outras fontes serão utilizadas como auxílio para a obtenção de dados não redundantes. STING\_RDB permite a realização de buscas complexas dentro de todo o conjunto de dados, reduzindo os requerimentos de armazenamento e conferindo a capacidade de análise de várias estruturas de proteínas ao mesmo tempo (o que não era possível com STING\_DB; Oliveira *et al.*, 2007).

Entre os módulos do STING, há o *Java Protein Dossier* (<sup>1</sup>PD; Neshich *et al.*, 2004) que é uma ferramenta de visualização que comunica muita informação através de um único gráfico, mostrado em formas de cores diferentes de acordo com o valor adotado para cada parâmetro. O <sup>1</sup>PD fornece aos usuários uma vasta coleção de parâmetros físico-químicos descrevendo a estrutura da proteína, estabilidade e interações com outras macromoléculas. Ao mesmo tempo, o <sup>1</sup>PD é um passo na direção de compilar uma base de dados diversificada de descritores de estrutura-função, que podem ser usados como uma plataforma para a aquisição de novos conhecimentos. <sup>1</sup>PD pode mostrar e analisar, simultaneamente, todos os parâmetros físico-

químicos de duas estruturas que tenham sido previamente superpostas, permitindo uma comparação direta de parâmetros entre estruturas similares.

Para vários descritores, obtemos informação não só do resíduo de aminoácido em questão mas também de seus vizinhos estruturais. Apesar de várias entradas do PDB possuírem moléculas de água cristalizadas junto com os resíduos de aminoácido da estrutura, essas moléculas não são usadas para o cálculo dos parâmetros descritos nas próximas sub-seções. Em parte, porque apenas estruturas de melhor resolução, trazem a informação sobre a real posição das moléculas de água.

O Grupo de Pesquisa em Biologia Computacional da Embrapa Informática Agropecuária possui diversas linhas de pesquisas visando o desenvolvimento de ferramentas computacionais que utilizam os descritores físico-químico, estruturais e de conservação para aumentar o conhecimento no campo da biologia estrutural, além da geração de produtos disponíveis publicamente.

Entre os trabalhos, destacamos o estudo aprofundado do efeito hidrofóbico em complexos proteína-proteína, que criou um índice de hidrofobicidade capaz de explicar, apenas do ponto de vista sobre hidrofobicidade, mais do que 98% das interfaces completas de complexos proteicos presentes no PDB.

Em outro estudo, foi possível aprofundar no conhecimento de *hot spots*, brevemente definidos como os resíduos de aminoácido com maior contribuição energética de contatos através da interface em complexos proteína-proteína. Utilizando os descritores físico-químicos e estruturais da base de dados BlueStar STING, foi possível prever resíduos de aminoácido como *hot spots* com maior precisão (74%) e sensibilidade (91%) do que outros métodos disponíveis na literatura. Esse trabalho culminou no desenvolvimento de uma dissertação de mestrado (Pereira, 2012)

De forma complementar, um estudo ainda em desenvolvimento está se mostrando promissor ao empregar os descritores físico-químicos e estruturais para prever os resíduos de aminoácido que compõe o sítio catalítico de enzimas, além de diferenciar entre as classes de enzimas existentes.

O conjunto desses três estudos somados à presente tese visa aprofundar o conhecimento sobre complexos proteína-proteína, no sentido de descrever o nano-ambiente em que os resíduos de aminoácido, ou conjunto de resíduos de aminoácido, estão inserido. Dessa forma, temos ferramentas com alto poder de aplicação para o campo de desenho racional de drogas e

agroquímicos, na definição de sítios específicos, em relação a interações com proteínas, para guiar experimentos ou identificar os fármacos promíscuos para serem evitados no desenvolvimento de tais drogas e agroquímicos.

Por fim, mas não menos importante, foi possível observar padrões e marcadores para a previsão de elementos de estrutura secundária em estruturas proteicas. Nessa frente de pesquisa, busca-se encontrar os elementos básicos que regulam a formação de estruturas secundárias no processo de enovelamento das proteínas. Essa abordagem visa aprofundar o conhecimento teórico sobre como o arranjo atômico das proteínas opera para formar a estrutura proteica final.

## 2.5 Seleção dos descritores físico-químicos e estruturais

Dentre todos os parâmetros disponíveis no STING\_DB, escolhemos um número reduzido de dados que, de acordo com nosso ponto de vista, apresentam maiores probabilidades de estarem associados com os processos de reconhecimento entre as proteínas. Os descritores escolhidos para a primeira etapa de desenvolvimento do projeto estão listados na tabela 2 e detalhados nas subseções 2.5.1 – 2.5.10.

Diferentemente dos trabalhos disponíveis na literatura, não foi utilizado descritores referentes à conservação de aminoácidos. Apesar de estes parâmetros estarem presentes no STING\_DB (em duas formas diferentes – HSSP e SH<sub>2</sub>Q<sup>S</sup>), o uso desse descritor inviabiliza a definição do nano-ambiente em que cada resíduo de aminoácido formador de interface está inserido, uma vez que conservação sequencial de aminoácidos é uma medida de um conjunto de proteínas homólogas, e não reflete nenhuma característica presente na estrutura proteica. A incorporação dos descritores de conservação está prevista para fim comparativo entre o desempenho de classificação dos resíduos de aminoácido formadores de interface.

**Tabela 2 - Descritores utilizados nas análises estatísticas e criação de modelos classificadores e suas abreviações em inglês.**

Acessibilidade (Accessibility - Acc)			
Acc em isolamento		Acc relativa	
Potencial Eletrostático (Electrostatic Potential - EP)			
EP no C $\alpha$	EP no LHA	EP média	EP na superfície
Hidrofobicidade			
Hidrofobicidade em isolamento			

Densidade de Energia de Contatos (Contact Energy Density - CED) Internos (INT)				
CED no C $\alpha$ INT (3)	CED no C $\alpha$ INT (6)	CED no LHA INT (3)	CED no LHA INT (6)	
CED no C $\alpha$ INT (4)	CED no C $\alpha$ INT (7)	CED no LHA INT (4)	CED no LHA INT (7)	
CED no C $\alpha$ INT (5)		CED no LHA INT (5)		
Cross Link Order (CLO)				
CLO no C $\alpha$		CLO no C $\beta$		CLO no LHA
Cross Presence Order (CPO)				
CPO no C $\alpha$		CPO no C $\beta$		CPO no LHA
Ângulos Diedrais				
	PHI		PSI	
Rotâmeros				
CHI 1	CHI 2	CHI 3	CHI 4	
Densidade				
Densidade no C $\alpha$ (3)	Densidade no C $\alpha$ (6)	Densidade no LHA (3)	Densidade no LHA (6)	
Densidade no C $\alpha$ (4)	Densidade no C $\alpha$ (7)	Densidade no LHA (4)	Densidade no LHA (7)	
Densidade no C $\alpha$ (5)		Densidade no LHA (5)		
Esponjicidade (Sponge)				
Esponjicidade no C $\alpha$ (3)	Esponjicidade no C $\alpha$ (6)	Esponjicidade no LHA (3)	Esponjicidade no LHA (6)	
Esponjicidade no C $\alpha$ (4)	Esponjicidade no C $\alpha$ (7)	Esponjicidade no LHA (4)	Esponjicidade no LHA (7)	
Esponjicidade no C $\alpha$ (5)		Esponjicidade no LHA (5)		
Número de Contatos				
Hidrofóbicos	Ligação de Hidrogênio - MM	Ligação de Hidrogênio - MWM	Ligação de Hidrogênio - MWWM	Aromáticos
Carregados (Atrativos)	Ligação de Hidrogênio - MS	Ligação de Hidrogênio - MWS	Ligação de Hidrogênio - MWWS	Ligação SS
Carregados (Repulsivos)	Ligação de Hidrogênio - SS	Ligação de Hidrogênio - SWS	Ligação de Hidrogênio - SWWS	
Energia de Contatos				
Hidrofóbicos	Ligação de Hidrogênio - MM	Ligação de Hidrogênio - MWM	Ligação de Hidrogênio - MWWM	Aromáticos
Carregados (Atrativos)	Ligação de Hidrogênio - MS	Ligação de Hidrogênio - MWS	Ligação de Hidrogênio - MWWS	Ligação SS
Carregados (Repulsivos)	Ligação de Hidrogênio - SS	Ligação de Hidrogênio - SWS	Ligação de Hidrogênio - SWWS	Energia total
Número de Contatos não Usados				
Hidrofóbicos	Ligação de Hidrogênio - MM	Ligação de Hidrogênio - MWM	Ligação de Hidrogênio - MWWM	Aromáticos
Carregados (Atrativos)	Ligação de Hidrogênio - MS	Ligação de Hidrogênio - MWS	Ligação de Hidrogênio - MWWS	Ligação SS
Carregados (Repulsivos)	Ligação de Hidrogênio - SS	Ligação de Hidrogênio - SWS	Ligação de Hidrogênio - SWWS	
Energia de Contatos não Usados				
Hidrofóbicos	Ligação de Hidrogênio - MM	Ligação de Hidrogênio - MWM	Ligação de Hidrogênio - MWWM	Aromáticos
Carregados (Atrativos)	Ligação de Hidrogênio - MS	Ligação de Hidrogênio - MWS	Ligação de Hidrogênio - MWWS	Ligação SS
Carregados (Repulsivos)	Ligação de Hidrogênio - SS	Ligação de Hidrogênio - SWS	Ligação de Hidrogênio - SWWS	Energia total não usada

**Rótulo:** C $\alpha$  = Carbon alpha; LHA = Last Heavy Atom; MM = Main chain – Main chain; MS = Main chain – Side chain; SS = Side chain – Side chain; W = Water molecule; WW = 2x Water molecules



### 2.5.1 Potencial Eletrostático

Os átomos que compõem os resíduos de aminoácido das proteínas podem, em determinadas condições, apresentar carga elétrica, que interagem com outras regiões carregadas da própria proteína ou ainda com outras moléculas e/ou íons de seu ambiente. Portanto, em um determinado ponto do espaço é possível calcular o potencial eletrostático devido as cargas presentes nas macromoléculas ao redor do ponto.

Blue Star STING utiliza o software *DelPhi* (Rocchia e Neshich, 2007) para cálculos de potencial eletrostático para todas as proteínas presentes no PDB através da resolução da equação de Poisson-Boltzmann da eletrostática:

$$\nabla \cdot [\epsilon_r(\mathbf{r})\nabla\varphi(\mathbf{r})] = \frac{1}{\epsilon_0} \left[ \rho^{cargas\_fixas}(\mathbf{r}) - \frac{\epsilon_{solv}\kappa^2(\mathbf{r})}{4\pi} \varphi(\mathbf{r}) \right] \quad (01)$$

onde  $\varphi(\mathbf{r})$  é o potencial eletrostático,  $\epsilon_r(\mathbf{r})$  é a constante dielétrica local,  $\epsilon_{solv}$  é a constante dielétrica para o solvente,  $\rho^{cargas\_fixas}(\mathbf{r})$  é a distribuição de cargas,  $\epsilon_0$  é a constante de permissividade do vácuo e  $\kappa^2(\mathbf{r})$  é o parâmetro de Debye, que depende da temperatura absoluta, concentração de sal, da carga líquida e da constante dielétrica.

Valores de potencial no Carbono- $\alpha$ , no último átomo pesado da cadeia lateral (LHA da sigla em inglês *Last Heavy Atom - LHA*), média do resíduo de aminoácido e média na superfície são armazenados no STING\_DB e foram utilizados nesse estudo. O arquivo de saída do *DelPhi* ainda apresenta valores para outras regiões específicas de cada aminoácido, por exemplo, do átomo de enxofre de cisteínas, no entanto, esses valores específicos não foram utilizados.

### 2.5.2 Hidrofobicidade

A maior parte das proteínas cuja estrutura foi resolvida e depositada no PDB são hidrofílicas. Porém, nem todos os tipos de aminoácidos que constituem as proteínas possuem em sua cadeia lateral átomos de nitrogênio e oxigênio capazes de estabelecerem ligações de hidrogênio com as moléculas de água. Para esses aminoácidos é associado o termo *hidrofílico* enquanto que para o restante dos aminoácidos é associado o termo *hidrofóbico*. A tabela 3 refere aos 20 tipos de aminoácido e suas características quanto à hidrofobicidade.

A característica *hidrofóbica* é associada ao favorecimento energético de átomos apolares (em especial o átomo de carbono) a estarem juntos espacialmente, reduzindo assim a área de contato com solvente polar. Dessa forma, os átomos de nitrogênio e oxigênio ficam mais expostos ao solvente.

Atualmente várias escalas de hidrofobicidade são aceitas na literatura. O software BlueStar STING utiliza a escala definida por Radzicka *et. al* (1998) (tabela 3). Para cada valor dessa escala multiplica-se pelo valor da área acessível do presente resíduo de aminoácido, resultando no valor de hidrofobicidade para cada resíduo da estrutura proteica:

$$\text{Hidrofobicidade} = \text{Radzicka}_i \frac{\text{area}_{acc,i}}{\text{area}_{total,i}} \quad (02)$$

onde  $\text{Radzicka}_i$  é o valor da escala de hidrofobicidade (última coluna da tabela 3);  $\text{area}_{acc,i}$  é o valor da área acessível do *i*-ésimo resíduo de aminoácido (descrição sobre o cálculo desse valor é mostrado na seção 2.5.5) e  $\text{area}_{total,i}$  é a área total do *i*-ésimo resíduo de aminoácido como calculada no tri-peptídeo Gli-X-Gli, onde cada um dos 20 tipos de aminoácido é intercalado por dois resíduos de glicina.

**Tabela 3 - Escala de Hidrofobicidade definida por Radzicka et al. (1988) e características físico-química dos 20 tipos de aminoácido.**

Aminoácido	Polaridade da Cadeia Lateral	Carga da Cadeia Lateral - pH 7,4	Constante de hidrofobicidade
(Arg)	polar	positiva	-14.92
(Asp)	polar	negativa	-8.72
(Glu)	polar	negativa	-6.81
(Asn)	polar	neutro	-6.64
(Lys)	polar	positiva	-5.55
(Gln)	polar	neutro	-5.54
(His)	polar	neutro/positiva	-4.66
(Ser)	polar	neutro	-3.4
(Thr)	polar	neutro	-2.57
(Tyr)	polar	neutro	-0.14
(Gly)	apolar	neutro	0.94
(Cys)	apolar	neutro	1.28
(Ala)	apolar	neutro	1.81
(Trp)	apolar	neutro	2.33

Aminoácido	Polaridade da Cadeia Lateral	Carga da Cadeia Lateral - pH 7,4	Constante de hidrofobicidade
(Met)	apolar	neutro	2.35
(Phe)	apolar	neutro	2.98
(Pro)	apolar	neutro	3.5
(Val)	apolar	neutro	4.04
(Ile)	apolar	neutro	4.92
(Leu)	apolar	neutro	4.92

Fonte: Radzicka et al., (1988).

### 2.5.3 Contatos e Densidade de Energia de Contatos

A maior parte dos resíduos de aminoácido estabelecem diferentes contatos com outros resíduos de aminoácido além da ligação peptídica que une a sequência primária de organização das proteínas. O software BlueStar STING diferencia 14 tipos de contatos, como mostrado na tabela 4, que é armazenada em sua base de dados ao nível atômico, ou seja, possui informações sobre quais pares de átomos dos resíduos de aminoácido estão estabelecendo os contatos, sendo que para cada tipo de contato é estabelecido um valor de energia associado. BlueStar STING define contatos baseados em distâncias relativas do par de átomos ou resíduos de aminoácido, este último no caso de contatos do tipo aromático, que estabelecem os diferentes tipos de contato (da Silveira *et al.*, 2009). A energia de cada tipo de contato utilizada e armazenada no STING\_DB é mostrada na tabela 4.

**Tabela 4 – 14 tipos diferentes de contatos armazenados no STING\_DB e seus respectivos valores de energia de interação**

	Tipo de Interação	Energia Associada (kcal/mol)
1	Hidrofóbica	0.06
2	Carregada Atrativa	10.00
3	Carregada Repulsiva	10.00
4	Ligação de Hidrogênio Cadeia Principal - Cadeia Principal	2.60
5	Ligação de Hidrogênio Cadeia principal - Cadeia Lateral	2.60
6	Ligação de Hidrogênio Cadeia Lateral - Cadeia Lateral	2.60

	Tipo de Interação	Energia Associada (kcal/mol)
7	Ligação de Hidrogênio Cadeia Principal - Cadeia Principal (1 - H <sub>2</sub> O)	2.60
8	Ligação de Hidrogênio Cadeia Principal - Cadeia Lateral (1 - H <sub>2</sub> O)	2.60
9	Ligação de Hidrogênio Cadeia Lateral - Cadeia Lateral (1 - H <sub>2</sub> O)	2.60
10	Ligação de Hidrogênio Cadeia Principal - Cadeia Principal (2 - H <sub>2</sub> O)	2.60
11	Ligação de Hidrogênio Cadeia Principal - Cadeia Lateral (2 - H <sub>2</sub> O)	2.60
12	Ligação de Hidrogênio Cadeia Lateral - Cadeia Lateral (2 - H <sub>2</sub> O)	2.60
13	Aromática	1.50
14	Ponte Dissulfeto	85.00

Fonte: [http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/energy\\_contacts\\_table.html](http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/energy_contacts_table.html)

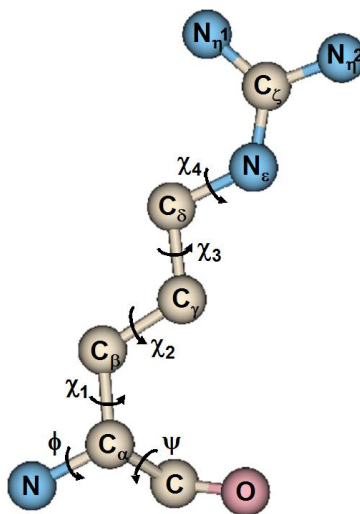
O parâmetro definido como Densidade de Energia de Contatos (ou CED – *Contact Energy Density*, na sigla em inglês) utiliza uma sonda esférica (de raio variável entre 3 e 7 Å) centrada no carbono- $\alpha$  ou *LHA* (totalizando 5 opções de cálculo) de cada aminoácido. Todos os contatos entre resíduos de aminoácido dentro da sonda esférica são somados e divididos pelo volume da esfera. Dessa forma, para cada resíduo de aminoácido há 20 descritores a partir deste parâmetro, pois se consideram de forma separada os contatos internos (entre resíduos de aminoácido de uma mesma cadeia) e externos (entre resíduos de aminoácido de cadeias diferentes). Para este trabalho utilizamos apenas os descritores de contatos internos.

#### 2.5.4 Rotâmeros

Os átomos dos resíduos de aminoácido se organizam formando ângulos entre si. Para os átomos da cadeia principal (nitrogênio, carbono- $\alpha$ , carbono-carbonila e oxigênio) há dois ângulos, PSI e PHI. Para os átomos da cadeia lateral pode haver até cinco ângulos,  $\chi$ -1 até  $\chi$ -5, porém como a cadeia lateral dos resíduos de aminoácido possuem comprimentos diferentes, nem todos possuem o mesmo grupo de ângulos, por exemplo, o ângulo  $\chi$ -5 é calculado apenas para resíduos de arginina, enquanto que para resíduos de glicina não há nenhum dos ângulos  $\chi$ , já que sua

cadeia lateral é um átomo de hidrogênio apenas. A figura 6 mostra os ângulos para um resíduo de arginina.

Assim como os ângulos PHI e PSI que apresentam preferências de configuração devido a possíveis choques estéricos entre os resíduos de aminoácido, os ângulos da cadeia lateral também apresentam algumas preferências (Schrauber *et al.* 1993), podendo apresentar um padrão de orientações.



**Figura 6 – Ângulos  $\chi$  da cadeia lateral do aminoácido arginina (único aminoácido cuja cadeia lateral se estende até o ângulo  $\chi$ -5).**

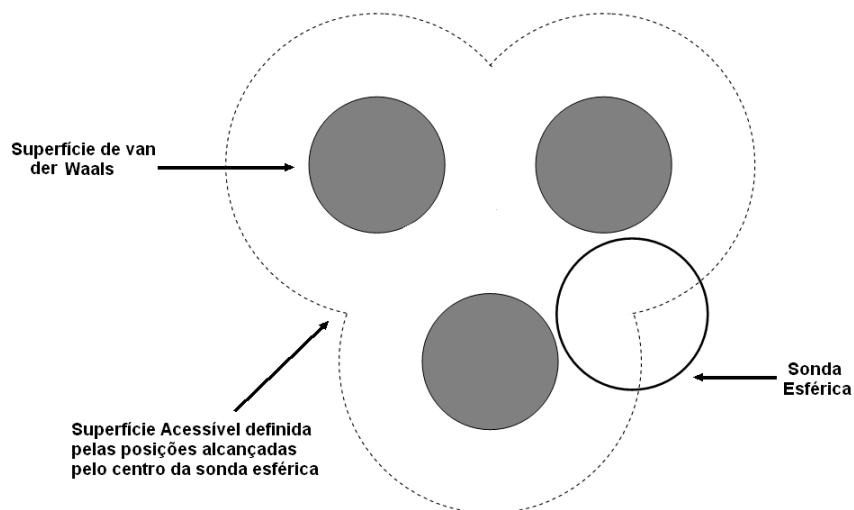
Cada resíduo de aminoácido possui determinados ângulos preferenciais como reportado por Schrauber *et al.* (1993). Os valores de cada ângulo para todos os resíduos de aminoácido de todas as estruturas proteicas presentes no PDB estão armazenados no STING\_DB, podendo ser acessadas pelo servidor do BlueStar STING.

### 2.5.5 Acessibilidade

Nem todos os resíduos de aminoácido que constituem as proteínas estão acessíveis ao solvente. Alguns estão enterrados no núcleo das estruturas proteicas e normalmente estão associados à estabilidade de tais estruturas, promovendo contatos que não são estabelecidos com o solvente, como por exemplo, os contatos hidrofóbicos entre átomos de carbono dos resíduos de aminoácido.

Os limites pré-definidos para o volume dos átomos é tido como o volume de van der Waals, que é o volume de uma esfera com raios distintos para cada tipo de átomo. A sobreposição de todos os volumes do arranjo tridimensional de átomos da estrutura de cada proteína gera sua

superfície de van der Waals. Para calcular a área acessível que cada resíduo de aminoácido fornece para a superfície proteica utiliza-se uma sonda esférica com raio igual ao raio da molécula de água (1,4 Å) e fazendo com que role pela superfície de van der Waals. A superfície gerada pelo centro geométrico da esfera sonda é a superfície acessível ao solvente (figura 7) (Lee e Richards, 1971).



**Figura 7 – Definição de superfície acessível de acordo com a superfície de van der Waals de cada molécula.**

Para calcular a área acessível que cada resíduo de aminoácido contribui para a superfície proteica, o software Blue Star STING utiliza o algoritmo *SurfV* (Sridharan *et al.*, 1992).

### 2.5.6 Contatos não Usados

Na seção 2.5.3 definimos como o software BlueStar STING define e armazena em sua base de dados os contatos estabelecidos pelos resíduos de aminoácido na estrutura tridimensional de proteínas e complexos proteicos. Através de STING\_DB pode-se facilmente obter valores de máximo para cada um dos 14 tipos de contatos estabelecidos por cada um dos 20 tipos de aminoácido em toda a base de dados. Comparando os valores do número de contatos presentes em cada resíduo de aminoácido em todas as proteínas do PDB, é possível calcular quantos contatos não foram estabelecidos em relação ao valor máximo pré-calculado.

Esse parâmetro estabelece um potencial de contatos não usados, que não leva em conta o meio ambiente em que cada resíduo de aminoácido está inserido, mas para interfaces apresenta-se como um bom parâmetro. O *número e energia de contatos não usados* define quais tipos de

contatos os átomos que a compõe ainda não realizaram quando comparados com outros dados presentes no STING\_DB.

### **2.5.7 Ordem de *Cross Link***

Devido ao dobramento da sequência de resíduos de aminoácido na estrutura tridimensional, resíduos distantes são colocados próximo no espaço, e podem portanto interagir entre si. O parâmetro *Cross Link* presente no software Blue Star STING leva essa característica em conta. A ordem no nome do descritor é definido em relação ao número dos contatos estabelecidos entre seguimentos de resíduos de aminoácido de, no mínimo, 15 resíduos de aminoácido na estrutura primária da proteína, e que estejam dentro da sonda esférica com raio 3,5 Å (devido ao dobramento da proteína) que pode ser centrada no carbono- $\alpha$ , carbono- $\beta$  ou no LHA.

### **2.5.8 Ordem de *Cross Presence***

Seguindo a mesma definição do parâmetro *Cross Link*, descrito na seção 2.5.7, o descritor *Cross Presence* conta todos os resíduos de aminoácido que estão dentro da sonda esférica de raio 3,5 Å centrada no carbono- $\alpha$ , carbono- $\beta$  ou no LHA, mesmo que os resíduos de aminoácido não estejam estabelecendo nenhum contato entre si.

### **2.5.9 Densidade**

A densidade local de cada resíduo de aminoácido é calculada utilizando uma abordagem de sonda esférica assim como estabelecido para o descritor de *densidade de energia de contatos* (seção 2.5.3). Para cada aminoácido, uma sonda esférica de raio variável (entre 3 e 7 Å) é centrada no carbono- $\alpha$  e no LHA. As massas dos átomos internos à sonda esférica são somadas e divididas pelo volume da sonda esférica. Utilizando os diferentes raios para sonda esférica, centradas em dois lugares diferentes, e distinguindo entre cadeias isoladas e em complexo proteicos, há um total de 20 variáveis deste descritor. Para este estudo utilizamos apenas as 10 variáveis relacionadas com as cadeias isoladas.

### 2.5.10 Esponjicidade

Seguindo a mesma abordagem descrita para o descritor *densidade* (seção 2.5.9), utilizamos uma sonda esférica de raio variável (entre 3 e 7 Å) centrada no carbono- $\alpha$  e no *LHA*. Porém, para o cálculo deste descritor não somamos as massas de todos os átomos, mas sim o volume ocupado por cada átomo (utilizando o raio de van der Waals e descontando o volume de sobreposição). Esse valor é subtraído e normalizado pelo volume da sonda esférica.

O descritor de *esponjicidade* é uma medida do espaço vazio do nano-ambiente de cada aminoácido.

## 2.6 Descritores ponderados pela vizinhança (WNA)

Para a versão final do modelo classificador de IFR por LDA e *ensemble* de redes neurais (apresentados na seção 2.10.2 e 2.13), testamos se a vizinhança dos resíduos de aminoácido formadores de interface e dos resíduos de aminoácido de superfície livre pode influenciar no desempenho e classificação. Para isso, cada dos descritores descritos acima recebe dois novos descritores ponderados pelos outros resíduos de aminoácido próximos espacialmente.

Assim como definido por Porollo e Meller (2007), uma sonda esférica com raio 15 Å centrada no carbono- $\alpha$  do resíduo de aminoácido de interesse, selecionamos os vizinhos espaciais de cada aminoácido. Os autores comentam que o raio da sonda esférica de 15 Å aumenta o poder de discriminação de cada um dos descritores, porém nenhuma busca por um valor ótimo desse raio é feita. Segundo da Silveira *et. al* (2009), após 15 Å de distância do resíduo de aminoácido de interesse, o número de vizinhos é próximo de zero. Dessa forma, 15 Å como limiar de distância mostra-se de forma adequada para mapear os vizinhos próximos estruturalmente. Para cada descritor  $D$  ponderamos os valores dos vizinhos e associamos o valor resultante ao resíduo de interesse segundo as equações:

$$D_{WNA}^{surf} = \sum_{i=0}^N D_i Acc_{relativa} \quad (03)$$

$$D_{WNA}^{dist} = D_0 + \sum_{i=1}^N \frac{D_i}{d_i} \quad (04)$$

Onde  $D_0$  é o valor do descritor  $D$  para o resíduo de aminoácido de interesse enquanto os outros  $D_i$  são os valores do mesmo descritor para os  $N$  vizinhos espaciais,  $Acc_{relativa}$  é a área



relativa acessível ao solvente e  $d_i$  é a distância do  $i$ -ésimo vizinho ao resíduo de aminoácido de interesse (medidos pelo carbono- $\alpha$ ). O subscrito WNA é a abreviação do nome em inglês para *Weighted Neighbor Averages*. A equação 03 indica que os resíduos de aminoácido com maior porcentagem de área acessível têm influencia maior sobre o aminoácido de interesse, enquanto a equação 04 indica que resíduos de aminoácido mais distante espacialmente influencia menos o resíduo de aminoácido de interesse independente da acessibilidade de cada aminoácido.

## 2.7 Testes estatísticos uni e multivariados

Com a coleta de todos os descritores físico-químicos e estruturais, podemos utilizar testes de inferência estatística para averiguar se os dados indicam diferenças nos padrões de valores para todos os descritores entre as duas populações: IFR e FSR. Com a utilização de técnicas estatísticas, podemos confirmar que os dados encontrados não são devido a processos aleatórios, mas que diferenças ou igualdades realmente representam diferenças de populações observadas.

### 2.7.1 Teste de Welch

O teste indicado para essa situação é o teste  $t$  de Welch para a comparação entre duas médias (Zar, 1999). Se considerarmos que não há diferença entre as duas populações, temos duas hipóteses para serem testada com os dados, em relação à média populacional  $\mu$ :

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

Ou seja, se não houver diferença entre as populações suas médias devem ser iguais assim como sua variância. Como estamos trabalhando apenas com uma amostra da população, não conhecemos os valores  $\mu_1$  e  $\mu_2$ , mas podemos estimar com a definição de média estatística amostral:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (05)$$

Da mesma forma não conhecemos a variância da população  $\sigma^2$ , apenas a estimativa dada por  $s^2$ .

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (06)$$

Se tivermos como hipótese nula que as populações são iguais, então ambas as medidas  $s^2_1$  e  $s^2_2$  são boas estimativas da variância  $\sigma^2$  da população. Dessa forma utilizamos a variância do conjunto das duas amostras:

$$s_p^2 = \frac{s_1^2 + s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad (07)$$

Como queremos testar a diferença entre as médias, precisamos da variância da diferença entre as duas amostras, ou seja, precisamos do valor dado por  $\sigma^2_{\bar{x}_1 - \bar{x}_2}$ . A variância da diferença de duas amostras independentes é obtida como a soma das variâncias:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (08)$$

Mas como estamos testando a hipótese que  $\sigma_1^2 = \sigma_2^2$ , temos:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \quad (09)$$

E como discutido acima,  $s_p^2$  é uma boa estimativa de  $\sigma^2$ . Dessa forma, temos todas as grandezas necessárias para o cálculo do teste estatístico utilizando a distribuição  $t$  de Student:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad (10)$$

Podemos, dessa forma, comparar o valor obtido dessa equação com tabelas disponíveis para a distribuição  $t$  no teste bicaudal com no máximo 1% de probabilidade de cometer o erro do tipo  $\alpha$  (rejeitar uma hipótese nula quando ela for verdadeira) e com  $n_1 + n_2 - 2$  graus de liberdade. A probabilidade de encontrarmos um teste estatístico com o resultado obtido mesmo estando errados ao descartar a hipótese nula é associada ao *valor p*, saída da maioria dos testes estatísticos.

### 2.7.2 Teste de normalidade de D'Agostino

O teste de Welch tem como premissa que os indivíduos das amostras (os resíduos de aminoácido) são independentes entre si e que provém de uma população normal (gaussiana). Por isso, antes de realizar os testes de Welch é necessário utilizar um teste estatístico preliminar para assegurar que os dados de entrada para o teste obedecem as condições para a amostra ser considerada normalmente distribuídas. Dentre algumas opções, escolhemos o teste de normalidade de D'Agostino (Zar, 1999). Assim com o teste de Welch, o teste de normalidade de D'Agostino retorna um *valor p* associado, permitindo que possamos averiguar qual a confiança em assumirmos que cada um dos descritores para cada resíduo de aminoácido seja distribuído normalmente.

O teste de D'Agostino calcula os valores de obliquidade  $g_1$  (do inglês *skewness*, relacionado com a simetria da distribuição dos valores) e curtose  $g_2$  (do inglês *kurtosis*, relacionado com o achatamento do pico da distribuição dos valores) de acordo com as equações:

$$g_1 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2\right)^{3/2}} \quad (11)$$

$$g_2 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2\right)^2} - 3 \quad (12)$$

Para uma população normalmente distribuída, a obliquidade e curtose devem ser o mais próximo possível de zero (distribuição simétrica). O teste de D'Agostino calcula os valores da média, variância, obliquidade e curtose dos próprios valores de  $g_1$  e  $g_2$  da amostra, comparando com os valores esperados para uma distribuição normal e retornando um *valor p* sobre a probabilidade de considerarmos a amostra como normal e estarmos errados sobre essa afirmação.

### 2.7.3 Teste não paramétrico de soma de ranques de Wilcoxon

Como alternativa ao uso de testes de normalidade e do teste de Welch, pode-se utilizar testes estatísticos não paramétricos. Esse tipo de estatística não requer que os dados provenham de uma distribuição estatística específica, como a distribuição normal.

O teste não paramétrico de soma de ranques de Wilcoxon (Noether, 1991) é um dos testes mais comuns em estatística não paramétrica e foi utilizado em conjunto com os testes descritos acima. A ideia de ranque consiste em ordenar as entradas do conjunto de dados pelos valores da variável de interesse sem se preocupar com a divisão entre as duas (ou mais) classes analisadas. Dessa forma, a ordem é mantida, mas os valores absolutos são perdidos, ou seja, podemos ainda afirmar quais dois resíduos de aminoácido quaisquer possui maior valor em determinado descritor físico-químico, mas não podemos mais afirmar o quão maior ou menor tal descritor é para cada aminoácido. Uma ideia fundamental nesse processo é a eliminação de possíveis *outliers* (resumidamente definido como entradas com valores bem acima ou bem abaixo da maioria dos indivíduos amostrados).

Com o ranque completo, a soma de todos os ranques para o conjunto de dados utilizado é o mesmo que a soma da progressão aritmética com razão 1:

$$\text{Soma dos ranques} = R = N(N + 1)/2 \quad (13)$$

A soma dos ranques para a amostra 1, denominada  $R_1$ , é feita de forma isolada. O parâmetro  $U$  é definido como:

$$U_1 = R_1 - n_1(n_1 + 1)/2 \quad (14)$$

Como sabemos a soma total dos ranques (equação 13),  $R_2$  é facilmente calculado a partir de  $R_1$ , seguindo para o cálculo de  $U_2$ . O menor valor entre  $U_1$  e  $U_2$  é utilizado para o cálculo do *valor p* utilizando a distribuição normal (para grandes bancos de dados). Para grandes amostras, situação para a qual estamos interessados,  $U$  pode ser aproximado por uma distribuição normal.

#### **2.7.4 Teste não paramétrico de duas amostras de Kolmogorov–Smirnov**

O teste conhecido como Kolmogorov-Smirnov (Noether, 1991) também é outra alternativa como teste estatístico não paramétrico que compara se duas amostras pertencem a mesma distribuição. Nesse teste, calculamos o número de amostras acima e abaixo da média para cada um dos descritores físico-químicos e estruturais utilizando a equação:

$$F_1(\bar{x}) = \frac{1}{n_1} \sum_{i=1}^{n_1} I_{x_i \leq \bar{x}} \quad (15)$$

Onde  $\bar{x}$  é a média do descritor  $x$  para ambas as classes (IFR e FSR), e a função  $I_{x_i \leq \bar{x}}$  vale 1 quando  $x_i \leq \bar{x}$  e 0 caso contrário. O mesma equação é aplicada para a classe 2.

Em seguida, o módulo da diferença entre  $F_1(\bar{x})$  e  $F_2(\bar{x})$  é calculado. A hipótese nula, de que as amostras 1 e 2 (FSR e IFR) provém da mesma distribuição, é rejeitada se:

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} |F_1(\bar{x}) - F_2(\bar{x})| > K_{0,01} \quad (16)$$

Onde  $K_{0,01}$  é o valor tabelado que resulta no cálculo da distribuição de probabilidade de Kolmogorov-Smirnov de errarmos em 1% das vezes quando rejeitamos a hipótese nula, sendo ela verdadeira.

Quando a equação 16 é verdadeira, as amostras são tidas como provenientes de distribuições diferentes, e, portanto, são bons descritores para diferenciar entre IFR e FSR.

### 2.7.5 Análise de variância multivariada

Considerando conjuntamente mais de uma variável, há a necessidade do emprego de técnicas estatísticas multivariadas. Quanto menor o grau de ortogonalidade entre as variáveis utilizadas, maior a necessidade de tais técnicas (Zar, 1999). A ferramenta R, com seu pacote HSAUR2.0 (Everitt e Hothorn, 2008), foi utilizada para o cálculo da análise de variância multivariada (MANOVA, do inglês *Multivariate ANalysis Of VAriance*; Johnson, 1998; Zar, 1999) e dos quatro testes estatísticos multivariados mais presentes na literatura: *Traço de Pillai*, *Traço de Hoetelling*, *Lambda de Wilks* e *Maior Raiz de Roy*.

Em análise de variância univariada (ANOVA, do inglês *ANalysis Of VAriance*; Zar, 1999), dois ou mais grupos são comparados utilizando a variância do descritor em consideração. ANOVA utiliza a variância em cada uma das classes, a variância entre as classes e a variância total sem diferenciação entre as classes. A variância entre as classes é denominada como variância de erro (diferença entre a variância total e variância em cada uma das classes). Com essas grandezas, o parâmetro  $F$  (distribuição  $F$  de Fisher) é estimado e comparado com valores tabelados.

Diferentemente, em MANOVA, a covariância entre as variáveis são utilizadas. A matriz de covariância dos descritores é multiplicada pelo inverso da matriz dos resíduos (erros). A matriz dos resíduos é obtida pela subtração da matriz de variância total da matriz de variância de cada uma das classes.

A matriz resultante da multiplicação da matriz de variância pelo inverso da matriz dos resíduos é utilizada para encontrar autovalores e autovetores. Cada um dos autovalores ( $\lambda_i$ ) é utilizado de forma diferente por cada dos testes citados acima:

1. *Traço de Pillai* calcula a grandeza  $\Lambda_{Pillai} = \sum_{i=1}^P \left( \frac{\lambda_i}{1+\lambda_i} \right)$
2. *Traço de Hoetelling* calcula a grandeza  $\Lambda_{Hoetelling} = \sum_{i=1}^P (\lambda_i)$
3. *Lambda de Wilks* calcula a grandeza  $\Lambda_{Wilks} = \prod_{i=1}^P \left( \frac{1}{1+\lambda_i} \right)$
4. *Maior Raiz de Roy* calcula a grandeza  $\Lambda_{Roy} = \max(\lambda_i)$

A distribuição utilizada para o calculo do *valor p* é conhecida como distribuição lambda de Wilks ( $\Lambda$ ), que é a generalização da distribuição univariada *F* de Fisher (utilizada na ANOVA).

## **2.8 Limitação de testes estatísticos para grandes bancos de dados**

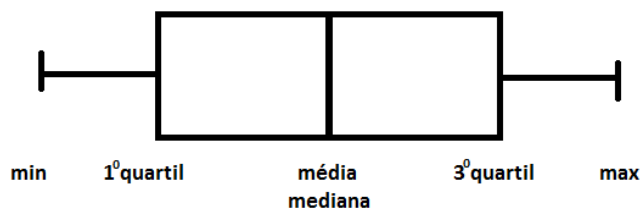
Zar (1999), define *estatística* como sendo uma forma de “análise e interpretação de dados visando uma avaliação objetiva da confiabilidade de conclusões baseadas nos dados.” Dessa forma, um pesquisador deve considerar uma amostra da população de tal forma que quanto mais indivíduos forem amostrados, maior será a confiabilidade dos dados. De fato o próprio teorema do limite central mostra que população que não segue uma distribuição normal tende à normalidade conforme sua amostra seja aumentada.

Os testes estatísticos utilizados e descritos na seção 2.7 são dependentes do número de amostras em cada classe ( $n_1$  e  $n_2$ ). Dessa forma, como o cálculo dos parâmetros que definem o *valor p* pode ser organizado de forma que são diretamente proporcionais ao número de amostras, grandes bancos de dados (lembrando que a base de dados utilizada conta com dezenas de milhares de indivíduos para cada um dos 20 tipos de aminoácido) sempre resultam em baixo *valor p* e, conseqüentemente, rejeitamos a hipótese nula para quase todos os descritores.

Johnson (1998), comenta que nessas situações é mais seguro utilizar métodos gráficos para a análise dos dados. Em um artigo mais recente, van der Laan *et al.* (2010) também discorrem sobre essa limitação de grandes bases de dados. Dessa forma, devido à quantidade de dados para cada um dos tipos de aminoácido (dezenas de milhares de entradas), não podemos confiar apenas nos testes estatísticos utilizados e descritos anteriormente para analisar os dados deste projeto.

## **2.9 Gráficos do tipo boxplot**

Devido à limitação de testes estatísticos descritos na seção 2.7, resolvemos auxiliar a forma de analisar se cada descritor provém de uma distribuição normal (Johnson, 1998) ou não, bem como assegurar quais descritores são mais bem sucedidos na distinção entre IFR e FSR, com a forma gráfica conhecida como *boxplot* (figura 8). Os gráficos utilizados apresentam seis medidas estatísticas para cada variável medida: 1° e 3° quartis, a média, mediana, mínimo e máximo. O primeiro quartil é definido como o valor da variável que separa os dados em 25% dos indivíduos com valores menores e 75% com valores maiores da variável em questão. O inverso é definido para o terceiro quartil, ou seja, o valor da variável que separa os dados em 75% dos indivíduos com valores menores e 25% com valores maiores. A diferença entre o terceiro e o primeiro quartil é chamada de IQR (do inglês *interquartile range*). Com essa definição a mediana também pode ser chamada de segundo quartil, uma vez que ela representa o valor que divide o número de indivíduos em 50% menor e 50% maior. Em uma representação do tipo *boxplot*, os valores de mínimo e máximo podem ser substituídos pelos valores de (1° quartil - IQR) e (3° quartil + IQR), respectivamente. Dessa forma qualquer indivíduo fora dessa faixa pode ser considerado um *outlier*, ou seja, um indivíduo cujo valor de determinada variável é extremamente diferente (acima ou abaixo) de sua população, e que talvez seja um erro associado à medida da variável. Para uma distribuição normal de uma variável, as seis grandezas apresentam simetria como ilustrado na figura 8. Quanto maior o desvio da simetria em *boxplots*, maior o grau de não normalidade da variável em questão.



**Figura 8 – Grandezas estatísticas medidas em um gráfico do tipo boxplot.**

Para distribuições normais (gaussianas) a simetria segue como indicado acima, e quanto maior for o desvio da simetria maior será o desvio de determinada população da normalidade. Os valores de média e mediana não precisam estar sobrepostos, como ilustrado acima.

## 2.10 Modelos classificadores

Um modelo classificador pode ser definido como um mapeamento ou uma função que, dado um conjunto de dados de entrada, retorna ou um valor numérico, uma probabilidade de classificação ou simplesmente associa o indivíduo a uma classe (Fawcett, 2005). Dado um classificador de resíduos de aminoácido formadores de interface e um resíduo de aminoácido a ser classificado (seus descritores), podemos obter quatro resultados possíveis quando há apenas duas classes possíveis de classificação: o resíduo de aminoácido pode ser de interface e ser classificado como tal (*verdadeiro positivo*, ou TP, da sigla em inglês *true positive*), o resíduo de aminoácido pode ser de superfície livre e classificado como tal (*verdadeiro negativo*, ou TN, da sigla em inglês *true negative*), o resíduo de aminoácido pode ser de interface, mas classificado como de superfície livre (*falso negativo*, ou FN, da sigla em inglês *false negative*) ou o resíduo de aminoácido pode ser de superfície livre mas classificado como de interface (*falso positivo*, ou FP, da sigla em inglês *false positive*). Com as quatro saídas possíveis, podemos definir a *matriz de confusão* (também chamada de *tabela de contingência*) utilizada por qualquer métrica de avaliação de modelos classificadores. Como mostra a figura 9, dado um problema de classificação binária, há quatro possíveis resultados para qualquer modelo: (TP), (TN), (FP) e (FN).

Diversas metodologias para a síntese de classificadores podem ser utilizadas e estão disponíveis na literatura. A princípio, não podemos afirmar qual metodologia é superior em termos de desempenho de classificação, e dependendo do problema a ser resolvido podemos ter desempenhos comparativos diferentes entre os modelos classificadores.



		Classe Predit	
		positivo	negativo
Classe Real	positivo	verdadeiro positivo (TP)	falso negativo (FN)
	negativo	falso positivo (FP)	verdadeiro negativo (TN)

Figura 9 – Matriz de confusão para resultados de modelos classificadores binários.

Na primeira etapa deste estudo exploramos o banco de dados avaliando o desempenho de 8 modelos classificadores distintos quanto ao problema de classificação de resíduos de aminoácido formadores de interface dado um conjunto de descritores físico-químicos e estruturais.

### 2.10.1 Modelos de regressão linear multivariados

O primeiro modelo matemático usado para analisar resíduos de aminoácido pertencentes à interface foi desenvolvido por meios da técnica estatística de regressão linear multivariada. Para isso a ferramenta R foi utilizada em conjunto com seu pacote *HSAUR 2.0*.

Modelos lineares exploram as relações entre as variáveis de entrada com a variável de saída (Zar, 1999). Nesse caso dizemos que a variável de saída é dependente da variável de entrada. O modelo matemático linear busca explorar correlações que estejam de acordo com a seguinte equação:

$$Y_i = \alpha + \beta X_i \quad (17)$$

$$\hat{Y}_i = \alpha + \beta X_i + \epsilon_i \quad (18)$$

Onde  $\alpha$  e  $\beta$  são parâmetros a serem estimados de acordo com os dados, enquanto  $\epsilon_i$  é o erro (ou resíduo) associado a diferença entre do valor calculado pelo modelo  $\hat{Y}_i$  e o valor real da variável dependente  $Y_i$ .

Para estimar os valores de  $\alpha$  e  $\beta$  é necessário minimizar os quadrados dos valores preditos e dos valores reais da variável dependente, ou seja, minimizar a expressão:

$$\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (19)$$

Diferenciando a equação acima em relação aos parâmetros  $\alpha$  e  $\beta$  e igualando a zero é possível demonstrar que a melhor estimativa de  $\beta$  é:

$$\beta = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (20)$$

Onde  $\bar{X}$  e  $\bar{Y}$  são os respectivos valores médios das variáveis independente e dependente.

Para  $\alpha$  podemos encontrar seu valor utilizando os valores médios:

$$\bar{Y} = \alpha + \beta\bar{X} \Rightarrow \alpha = \bar{Y} - \beta\bar{X} \quad (21)$$

No entanto, como nesse estudo há diversas variáveis de entrada e apenas uma variável de saída precisamos de um modelo multivariado. A equação da reta acima pode ser generalizada para o caso de N variáveis, resultando em:

$$\hat{Y}_i = \alpha + \sum_{j=1}^N \beta_j X_{ji} + \epsilon_i \quad (22)$$

Para testar a significância de um modelo de regressão podemos usar teste de ANOVA (Zar, 1999). O cálculo de algumas grandezas é necessário, entre elas:

$$\text{soma dos quadrados totais} = \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (23)$$

$$\text{soma dos quadrados da regressão} = \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 \quad (24)$$

$$\text{soma dos quadrados dos resíduos} = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (25)$$

A soma dos quadrados dos resíduos também pode ser definida como a diferença entre a soma dos quadrados totais e a soma dos quadrados da regressão.

O grau de liberdade associado a variabilidade de  $Y_i$  é igual ao número total de amostras subtraído de 1, ou seja,  $N-1$ . Em uma equação de regressão linear simples, o grau de liberdade da variabilidade do modelo é sempre igual a 1 (Zar, 1999). Da mesma forma que a soma dos quadrados, o grau de liberdade dos resíduos é calculado como a diferença entre o grau de liberdade total e o grau de liberdade do modelo linear simples, ou seja,  $N-2$ . A quantidade *quadrado médio* (do inglês *mean squares*) é definida como a razão entre a soma dos quadrados e os grau de liberdade:

$$\text{quadrado médio} = \frac{\text{soma dos quadrados}}{\text{grau de liberdade}} \quad (26)$$

Com essas grandezas calculadas, a hipótese nula  $\beta_j = 0$  para todo  $j$  pode ser testado usando a distribuição *chi-quadrado* e o teste  $F$ :

$$F = \frac{\text{quadrado médio de regressão}}{\text{quadrado médio dos resíduos}} \quad (27)$$

O software R faz essa análise e retorna nível de significância do modelo gerado. Ao final do cálculo do modelo (todos os coeficientes  $\beta_j$  e  $\alpha$ ) o software R também lista o *coeficiente de determinação* definido como o quadrado do *coeficiente de correlação* mostrado acima. O *coeficiente de determinação* é uma medida da percentagem da variabilidade em  $Y_i$  que é explicada pelo modelo linear  $\hat{Y}_i$ , ou seja, da “força” da relação linear imposta pelo modelo.

### 2.10.2 Modelo classificador por análise discriminante linear

O modelo classificador por análise discriminante linear segue o raciocínio apresentado pelo modelo de regressão linear (Johnson, 1998). Nesse caso, no entanto, não queremos prever o valor de alguma variável de saída, apenas a qual classe (IFR ou FSR) cada resíduo de aminoácido pertence.

Para isso, os dados do conjunto de treinamento são utilizados para gerar valores de média e variância para cada descritor para cada classe IFR e FSR. Assim, o conjunto de treinamento gera os vetores de média  $\mu_{IFR}$  e  $\mu_{FSR}$  e os vetores de variância  $\Sigma_{IFR}$  e  $\Sigma_{FSR}$ . Cada classe possui uma probabilidade *a priori* de que um dado resíduo de aminoácido pertença à classe IFR ou FSR, dada

pela frequência de ocorrência em cada classe sem utilizar qualquer informação dos parâmetros físico-químicos e estruturais utilizados na etapa de classificação.

A função de verossimilhança é usada para ambas às classes:

$$f_{IFR} = \frac{1}{(2\pi)^{N/2} |\Sigma_{IFR}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{IFR})' \boldsymbol{\Sigma}_{IFR}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{IFR}) \right] \quad (28)$$

$$f_{FSR} = \frac{1}{(2\pi)^{N/2} |\Sigma_{FSR}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{FSR})' \boldsymbol{\Sigma}_{FSR}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{FSR}) \right] \quad (29)$$

A escolhida é aquela que retorna maior valor da função de verossimilhança.

Para incorporar probabilidades *a priori*, basta multiplicar o resultado de cada função pela frequência de ocorrência de cada tipo de aminoácido como IFR ou FSR, de forma que a soma dos dois termos é igual a 1 (ou 100%):

$$P_{IFR} = freq_{IFR} \cdot f_{IFR} \quad (30)$$

$$P_{FSR} = freq_{FSR} \cdot f_{FSR} \quad (31)$$

Assim, a escolha da classe para cada entrada é feita simplesmente com a comparação:

$$classe\ predita = \max\{P_{IFR}; P_{FSR}\} \quad (32)$$

O modelo de regressão linear pode ser usado como modelo classificador aplicando limiares de classificação para o valor predito. No entanto, com o desenvolvimento de um modelo classificador por análise discriminante linear, resolvemos utilizá-lo apenas como uma ferramenta analítica, uma vez que o valor do *coeficiente de determinação* fornece uma medida da variabilidade da variável binária de resposta explicada pelo modelo linear.

### 2.10.3 Árvore de inferência condicional não enviesada

Árvores de decisão são técnicas de aprendizado de máquina que recursivamente dividem os dados com o objetivo de separar as diferentes classes estudadas, ou ainda podem ser utilizadas para fins de regressão (Murthy, 1998). Há diferentes algoritmos disponíveis na literatura, sendo

que entre os mais conhecidos e utilizados está o algoritmo conhecido como C4.5 (Quinlan, 1993) e sua nova implementação J48 (Witten *et. al*, 2011).

Árvores de decisão são conhecidas por possuírem pelo menos duas características negativas: tendência de favorecimento de escolha de variáveis com maior número de possíveis separações para variáveis categóricas e sobre ajuste aos dados de treinamento (Segal, 1988). O último problema pode ser amenizado aplicando técnicas de poda da árvore construída. No entanto, ao aplicar tais procedimentos, as variáveis categóricas com múltiplas saídas ainda continuam no topo da árvore, afetando a interpretabilidade do modelo resultante.

Para superar ambas as falhas em árvores de decisão, Hothorn *et al.*(2006) descrevem o pacote *party* disponível gratuitamente para o ambiente R. Os autores definem seu algoritmo como uma árvore de inferência condicional não enviesada, ou seja, nenhuma preferência para variáveis categóricas com múltiplas respostas é observada. Esse método consiste em medir a correlação entre cada variável e a variável de saída (para nosso caso, a variável binária 1 = FSR e 2 = IFR) por meio de testes estatísticos (parâmetro  $F$  de ANOVA ou distribuição  $\chi^2$ ; Zar, 1999). A variável com maior correlação é escolhida como a primeira a separar os dados. Usamos a correlação linear simples (equação 33) para correlacionar duas variáveis, uma variável independente e uma variável de resposta binária, nesse caso:

$$r = \frac{\sum y.x_i}{\sqrt{\sum y^2 \sum x_i^2}} \quad (33)$$

Para utilizar o F-escore de ANOVA, podemos calcular  $F$  como:

$$F = \frac{1+|r|}{1-|r|} \quad (34)$$

Para esse estudo, utilizamos o F-escore como grandeza para classificar as variáveis, mas, de forma alternativa, pode-se utilizar o teste  $\chi^2$  calculando a variância de  $r$  como:

$$s_r^2 = \frac{1-r^2}{n-2} \quad (35)$$

Em seguida, definimos um valor arbitrário para a variância  $\sigma_r^2$  esperada da medida de  $r$  entre a variável independente e a variável de resposta, e calculamos  $\chi^2$  com  $\nu$  graus de liberdade da amostra como:

$$\chi^2 = \frac{\nu s_r^2}{\sigma_r^2} \quad (36)$$

Para ambos os parâmetros  $\chi^2$  e  $F$ , podemos associar um *valor p* dos respectivos testes estatísticos, e utilizar cada uma das variáveis  $x_i$  cujo *valor p* é menor que um valor pré-estabelecido (0,05 foi o valor escolhido).

Para evitar a escolha de variáveis categóricas com múltiplas saídas, apenas repartições binárias são permitidas. Como critério de parada, é definido um valor mínimo para a correlação entre as variáveis independentes e a variável de resposta. O algoritmo cessa quando as variáveis ainda não utilizadas estão acima desse valor mínimo.

#### 2.10.4 Ensemble de árvores de inferência condicional via *bagging*

Árvores de decisão são algoritmos com muitas aplicações em diferentes ramos da ciência. No entanto, uma determinada árvore treinada é esperada falhar para certo número de dados de testes. Devido à sensibilidade que o treinamento de uma árvore de decisão possui em relação ao conjunto de treinamento (sobre ajuste aos dados de treinamento), podemos criar árvores diferentes com subconjuntos de treinamento diferentes do mesmo conjunto inicial de treinamento. Dessa forma, árvores específicas podem errar, mas é mais improvável que árvores diferentes errem no mesmo indivíduo, presente no conjunto teste, ou seja, na média um *ensemble* de árvores de decisão vai apresentar a classificação correta para um determinado indivíduo.

Uma forma de criar diferentes árvores de decisão pela manipulação do conjunto de treinamento é conhecida como *bagging* (do inglês *bootstrap aggregating*). Foi utilizado o pacote *party* para o software R. Esse método consiste em amostrar indivíduos do conjunto de dados original com reposição, ou seja, algumas réplicas do mesmo indivíduo estão presentes em um mesmo subconjunto de treinamento. Para um conjunto de treinamento de tamanho  $N$ , a probabilidade de um indivíduo ser selecionado é de  $1/N$ , ou seja, a probabilidade de um indivíduo não ser selecionado é de  $(1 - 1/N)$ . Se cada subconjunto tiver o mesmo tamanho que o conjunto

inicial ( $N$  indivíduos), a probabilidade de um determinado indivíduo não ser selecionado para a criação do subconjunto é de:

$$p_{\text{não}} = (1 - 1/N)^N \approx e^{-1} \approx 0,368 \quad (37)$$

Dessa forma apenas 63,2% dos indivíduos são únicos (aproximação válida para  $N \rightarrow \infty$ ). Cada subconjunto de treinamento possui variância menor do o conjunto inicial. A primeira vantagem desse método é a diminuição do sobre ajuste aos dados de treinamento. Strobl *et. al* (2008) mostraram que o uso de *ensemble* de árvores de decisão construídos por *bagging* supera o desempenho de árvores de decisão simples.

### 2.10.5 Ensemble de árvores de decisão via *random Forest*

O método de *random Forest* foi primeiramente desenvolvido por Breiman (2001). Assim como *bagging* o método de criação de *ensemble* de árvores de decisão via *random Forest* consiste na criação de subconjuntos de treinamento, porém outra variabilidade na construção de árvores individuais é inserida: a limitação do número de variáveis disponíveis para treinamento. Dessa forma, as árvores contidas no *ensemble* são dotadas de maior variabilidade entre si do que as árvores produzidas por *bagging*. Com a incorporação dessa variabilidade extra no treinamento, variáveis que não seriam escolhidas, devido à presença de variáveis mais importantes, aparecem em alguns modelos.

O mesmo pacote *party* utilizado anteriormente possui o algoritmo para esse método, e, portanto, a mesma árvore de inferência condicional é utilizada. Strobl *et al.* (2008) indicam que um valor de referência para o número de variáveis disponíveis para estabelecer cada árvore do *ensemble* seja reduzido de  $p$  no conjunto de dados originais para  $\sqrt{p}$  em cada subconjunto de treinamento na geração do *ensemble*. A escolha de quais variáveis estão disponíveis em cada etapa é feita de forma aleatória.

### 2.10.6 Classificação por regressão logística

Regressão logística possui algumas vantagens sobre o modelo LDA. Diferentemente dos modelos criados por LDA, regressão logística não faz nenhuma inferência sobre a distribuição

associada às variáveis, ou seja, não é necessário que as variáveis sejam distribuídas seguindo uma curva normal (gaussiana). Outra diferença está na necessidade de uma variável de resposta binária para o uso de modelos de regressão logística. Para casos com maior número de classes, grupos de classificadores binários podem ser construídos (Johnson, 1998). Sob a Teoria Linear Generalizada (GLM, do inglês *Generalized Linear Theory*; Nelder and Wedderburn, 1972; McCulloch, 2000), ambas as metodologias LDA e regressão logística são tratadas da mesma forma.

Utilizamos o pacote *stats* (R Development Core Team) para a criação desses modelos.

Na regressão logística, as variáveis também são associadas linearmente de acordo com coeficientes  $\beta_1$ . A probabilidade associada a cada um dos dois valores possíveis para a variável de resposta é dado por:

$$p(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)} \quad (38)$$

É aplicada sobre o modelo resultante a transformação *logit*, que carrega as mesmas propriedades de modelos de regressão linear:

$$g(x) = \log \left\{ \frac{p(y=1|x)}{1-p(y=1|x)} \right\} = \beta_0 + \beta_1 \cdot x \quad (39)$$

Os parâmetros de modelos por regressão linear (inclusive os modelos gerados no desenvolvimento desta tese) são ajustados segundo o método dos quadrados mínimos, enquanto para modelos de regressão logística são usados métodos de máxima verossimilhança (do inglês *maximum likelihood method*; Johnson, 1998).

### 2.10.7 Classificador Naïve Bayes

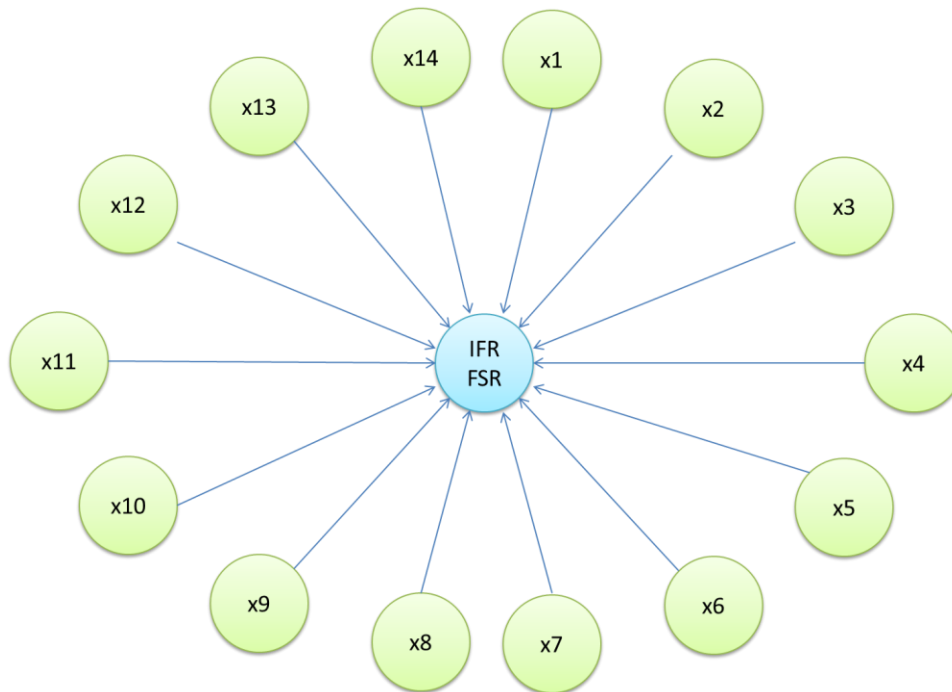
Classificadores do tipo Naïve Bayes são modelos criados com grafos com apenas um nodo parental (de forma semelhante às árvores de decisão), como mostra a figura 10. Esse tipo de modelo é também denominado como rede bayesiana simples (Kotsiantis, 2007). Para a construção desse modelo classificador, cada nodo filho (descriptor) é considerado independente dos outros, ou seja, a etapa anterior de eliminação de variáveis com coeficiente linear simples é fundamental, e, assim, probabilidades condicionais entre o nodo parental e filhos são criados utilizando os dados



disponíveis. Não há probabilidades condicionais entre nodos filhos, ou seja, cada par de variáveis são linearmente independentes entre si.

A decisão ocorre com o cálculo do parâmetro  $R$ :

$$R = \frac{P(FSR|x)}{P(IFR|x)} = \frac{P(FSR)P(x|FSR)}{P(IFR)P(x|IFR)} = \frac{P(FSR) \prod P(x_i|FSR)}{P(IFR) \prod P(x_i|IFR)} \quad (40)$$



**Figura 10 – Representação de uma rede bayesiana simples (Naïve Bayes).**

*Classificação ocorre no nodo parental (em azul) utilizando diversos nodos filhos (variáveis  $x_i$ ). As setas indicam probabilidades condicionais entre as variáveis independentes (filhos) e a variável de resposta (nodo parental).*

Dessa forma, para  $R > 1$  escolhe-se a classe 1 (FSR) e, caso contrário, escolhe-se a classe 2 (IFR). Como a equação (40) é um produtório, o resultado pode ser severamente influenciado caso qualquer  $P(x_i | \dots) = 0$ . Uma forma de controlar é usar estimadores de Laplace, porém para o presente trabalho essa opção não foi utilizada, uma vez que  $P(x_i | \dots) \neq 0$  para todo  $x_i$ .

Para estabelecer esse classificador, utilizamos o pacote *e1071* do software R. Esse pacote calcula as probabilidades condicionais do lado direito da equação (40) a partir da distribuição gaussiana, sendo que os parâmetros ( $\mu_{i,1,2}$  e  $\sigma_{ifr=1,2}$ ) são retirados dos dados de treinamento:

$$P(x_i | ifr = 1,2) = \frac{1}{\sqrt{2\pi}\sigma_{ifr=1,2}} \exp\left(-\frac{(x_i - \mu_{i;1,2})^2}{\sigma_{ifr=1,2}^2}\right) \quad (41)$$

Domingos e Pazzani (1997) compararam os resultados de diferentes classificadores e encontraram que classificadores do tipo Naïve Bayes podem ter resultados iguais ou mesmo superiores quando comparados a algoritmos mais complexos, como árvores de decisão.

### 2.10.8 Classificador por redes-neurais no R

Foi desenvolvido no software R uma rede neural utilizando o pacote *nnet* (Venables e Ripley, 2002). Uma definição mais detalhada de redes neurais é dada na seção 2.13, tendo em vista o sucesso da aplicação desse modelo no problema de classificação de resíduos de aminoácido em IFR e FSR. Nesta seção, no entanto, introduzimos alguns conceitos fundamentais em relação a redes neurais artificiais com alimentação para frente (do inglês *feed-forward*).

A rede neural estabelecida é uma MLP (do inglês *multilayered perceptron*; Rumelhart e McClelland, 1986), ou seja, pode haver uma ou mais camadas de tratamento das variáveis desde a entrada até a saída dos dados, as quais são chamadas de camadas ocultas. Redes neurais artificiais apresentam grande destaque entre os modelos classificadores, tanto no desenvolvimento de modelos para treinamento supervisionado, que é caso deste estudo, como de treinamento não supervisionado visando agrupamento de dados, ou mesmo para fins de otimização (Puma-Villaneuva, 2011). Em parte, o grande sucesso de aplicação se deve ao fato de sua capacidade de aproximação universal, ou seja, ao estabelecer um mapeamento multidimensional e não-linear é possível chegar a resposta do problema de classificação. No entanto, a capacidade de aproximação universal é apenas uma prova teórica, não existindo um método sistemático capaz de retornar a melhor rede neural artificial, ou seja, sua arquitetura com número de neurônios, número de camadas ocultas e com pesos de conexão ótimas. Ainda, com a provável existência de dados com presença de ruído em um conjunto finito de amostras, a arquitetura da rede neural estabelecida pode não ser capaz de sintetizar o melhor mapeamento para o problema estudado.

A MLP utilizada contém uma camada de entrada dos dados, uma camada oculta e uma camada de saída. Para a camada de entrada há tantos neurônios quanto há variáveis. Na camada oculta, há 8 neurônios e apenas um na camada de saída, que é então normalizado entre zero e um, resultando em uma interpretação probabilística para o processo de classificação. Nessa etapa,

não foi explorado a variação do número de neurônios na camada oculta, visando melhor desempenho de classificação.

Treinar uma rede neural consiste em ajustar os pesos que ligam cada variável à camada oculta e, também, os pesos que ligam cada neurônio da camada oculta com a camada de saída, de forma a minimizar os erros de classificação cometidos, ou seja, o modelo aprende com os dados de entrada. Para ajustar os pesos que ligam cada variável de entrada aos neurônios da camada oculta e os pesos que ligam os neurônios da camada oculta até o neurônio da camada de saída, foi utilizado o método BFGS (Broyden–Fletcher–Goldfarb–Shanno) de otimização não linear (Luenberger, 1984), contido no comando *optim* do software R. Dessa forma, o treinamento ocorre até que nenhuma melhora nos pesos é encontrada (levando várias centenas de iterações).

Devido ao melhor desempenho desse modelo classificador em relação aos demais tipos estudados, redes neurais foram aperfeiçoadas para o problema de classificação de resíduos de aminoácido formadores de interfaces, como será mostrado na seção de resultados.

### **2.10.9 Classificador por máquinas de vetores suporte**

Máquinas de vetores suporte (SVM, do inglês *Support Vector Machine*) tentam classificar duas classes maximizando a “margem” entre duas classes no espaço expandido das variáveis de entrada (Vapnik, 1995; Burges, 1998; Kotsiantis, 2007), ou seja, um espaço multidimensional composto pelas variáveis de entrada. Modelos do tipo SVM têm sido aplicados em uma variedade de problemas de reconhecimento de padrão, como reconhecimento de caracteres, de objetos, de rostos e categorização de textos (Burges, 1998). A capacidade de generalização de um modelo por SVM é comparável ou mesmo superior a outros métodos de classificação, como redes neurais, em alguns conjuntos de testes na literatura (Burges, 1998), principalmente quando há poucos indivíduos amostrados no conjunto de treinamento (Kotsiantis, 2007). No caso de classificação de resíduos de aminoácido formadores de interface, Porollo e Meller (2007) desenvolveram e compararam modelos por SVM e redes neurais artificiais, sendo que nos modelos gerados pelos autores (classificadores inespecíficos quanto ao tipo de aminoácido), o modelo do tipo SVM foi avaliado com mesma taxa de acerto que o modelo por redes neurais, porém com maior precisão e menor cobertura.

Para modelos por SVM, as classes binárias são normalmente rotuladas de  $y_i \in \{-1,+1\}$  (diferentemente dos rótulos FSR = 1 e IFR = 2 utilizado nos outros modelos), e buscamos um

hiperplano que divide ambas as classes, ou seja, buscamos resolver as seguintes equações e suas restrições:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad \text{para todo } y_i = +1 \quad (42)$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{para todo } y_i = -1 \quad (43)$$

onde  $\mathbf{x}_i$  é vetor de entrada com as variáveis do  $i$ -ésimo indivíduo,  $\mathbf{w}$  e  $b$  são coeficientes do hiperplano. Ambas as equações podem ser combinadas, resultando:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \text{para todo } i \quad (44)$$

Introduzindo o *Lagrangiano* na inequação acima, temos a redução do problema de otimização original na forma do problema dual (Borges, 1998). Nessa formulação, são incorporados multiplicadores de Lagrange não negativos como uma nova restrição ao problema. Os multiplicadores de Lagrange são incorporados na equação a ser otimizada da seguinte forma:

$$L_p \equiv \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_i \alpha_i \quad (45)$$

$$\alpha_i \geq 0 \quad \text{para todo } i$$

onde  $\alpha_i$  são os multiplicadores de Lagrange (um para cada inequação). A solução do melhor hiperplano que divide ambas as classes estudadas é encontrada minimizando o *Lagrangiano*  $L_p$  com respeito a  $\mathbf{w}$  e  $b$ , ou seja, nos valores de  $\alpha_i$  onde todas as derivadas se anulam. O primeiro termo do lado direito da equação (45) é denominado *função objetivo*, que deve ser minimizada. Esta formulação é conhecida como *problema de programação quadrática* (Borges, 1998), e também pode ser resolvido maximizando  $L_p$  com a condição de que o gradiente de  $L_p$  em relação a  $\mathbf{w}$  e  $b$  se anule. Aplicando as condições de otimização, temos as seguintes restrições para as soluções:

$$\nabla_{\mathbf{w}} L_p = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \rightarrow \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (46)$$

$$\frac{\partial}{\partial b} L_p = -\sum_i \alpha_i y_i = 0 \rightarrow \sum_i \alpha_i y_i = 0.$$

Aplicando essas novas restrições ao Lagrangiano da equação 45, temos:

$$L_p = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (47)$$

As restrições (44), (45) e (46) acima são conhecidas como condições de KKT (Karush-Kuhn-Tucker; Burgues, 1998) e são essenciais para vários problemas de otimização com restrições, e todas as possíveis soluções seguem essas condições (Burgues, 1998). Para completar as condições de KKT, mais uma restrição deve ser adicionada:

$$\alpha_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1) = 0 \quad \text{para todo } i \quad (48)$$

Esta última condição recebe o nome de *condição complementar de KKT*. Após a determinação de  $\mathbf{w}$  pelas outras condições, a restrição da equação (48) pode ser utilizada para encontrar possíveis valores do parâmetro  $b$ . Burgues (1998) comenta que o melhor procedimento é escolher o valor médio de todos os  $b$  encontrados para todos os indivíduos de treinamento cujo multiplicador de Lagrange  $\alpha_i$  é diferente de zero, ou seja, os indivíduos que estão na divisa entre as duas classes.

Para o problema em que erros de classificação são aceitos e possíveis (como é o problema de classificadores de resíduos de aminoácido formadores de interfaces), é necessário modificar as equações, incorporando um custo para todas as instâncias que foram erroneamente classificadas. As equações (42), (43) e (44) são reescritas como:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \varepsilon_i \quad \text{para } y_i = +1 \quad (49)$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \varepsilon_i \quad \text{para } y_i = -1 \quad (50)$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - (1 - \varepsilon_i) \geq 0 \quad (51)$$

$$\varepsilon_i \geq 0 \quad \text{para todo } i$$

A função objetivo pode ser modificada adicionando um termo relacionado com o custo de erro na classificação, de forma que o novo Lagrangiano pode ser escrito como:

$$L_p \equiv \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \varepsilon_i - \sum_i \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_i \alpha_i (1 - \varepsilon_i) - \sum_i \mu_i \varepsilon_i \quad (52)$$

$$\mu_i \geq 0 \quad \text{para todo } i$$

onde os dois primeiros termos do lado direito são chamados de *função objetivo*,  $\mu_i$  são novos multiplicadores de Lagrange para forçar a condição de positividade de  $\varepsilon_i$  e  $C$  é um parâmetro arbitrário relacionado a quanto deve ser penalizado cada classificação errada. O gradiente do novo

Lagrangiano não é modificado em relação às equações (46), mas é necessário derivar  $L_p$  em relação a  $\varepsilon_i$ , resultando na condição:

$$\frac{\partial}{\partial \varepsilon_i} L_p = C - \alpha_i - \mu_i = 0 \quad (53)$$

Novamente, as condições de KKT são necessárias e suficientes para qualquer solução do problema proposto. As duas condições de KKT complementares (relacionadas aos multiplicadores de Lagrange) são:

$$\begin{aligned} \alpha_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - (1 - \varepsilon_i)) &= 0 \quad \text{para todo } i \\ \mu_i \varepsilon_i &= 0 \quad \text{para todo } i \end{aligned} \quad (54)$$

Todas as equações sobre SVM tratadas até então são para separação linear entre as classes. Para o desenvolvimento de SVMs não lineares, é necessário usar uma função *kernel*, que pode ser definida como uma função que transforma os dados para um espaço de maior dimensão. A função *kernel* permite que separações lineares neste espaço de maior dimensão, corresponda a separações não-lineares no espaço das variáveis originais (Kotsiantis, 2007). Novos pontos amostrais são classificados diretamente no espaço de maior dimensão, não sendo necessário explicitar as funções que definem o mapeamento entre o espaço original linear e o espaço de variáveis não-linear. Dado uma lista de funções *kernel* não é possível determinar *a priori* qual é a mais indicado para um dado problema (Burgues, 1998; Kotsiantis, 2007). Utilizamos o pacote *e1071* para os cálculos dos modelos de SVM, e, nesse pacote, há a implementação da função *kernel* radial gaussiana definida como:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \quad (55)$$

onde  $\gamma$  é um parâmetro a ser mapeado manualmente pelo usuário (assim como o parâmetro  $C$  da equação 52). A função *kernel* (53) substitui o produto interno  $(\mathbf{x}_i \cdot \mathbf{x}_j)$  do Lagrangiano da equação (47). Dessa forma, apenas fornecendo os dados de entrada  $\mathbf{x}_i$ , suas classes  $y_i$ , os parâmetros  $C$  e  $\gamma$ , as equações acima são suficientes para o processo de treinamento de uma SVM não linear.

Como nem todos os resíduos de aminoácido presentes no conjunto de treinamento são usados no processo de treinamento da SVM, esse método é menos sensível à dados com ruídos.

## **2.11 Tratamento dos dados por análise de componentes principais**

Johnson (1998) trata assuntos relacionados à aplicação da técnica de análise de componentes principais. O autor destaca que a técnica deve ser usada para o melhor entendimento das variáveis originais e como elas interagem entre si. Podemos usar esse procedimento para a redução do número de variáveis a serem utilizadas, devido à criação de novas variáveis normalmente distribuídas e não correlacionadas entre si, chamadas de *componentes principais*.

A normalidade dos componentes principais é vital para a aplicação de várias técnicas analíticas, como, por exemplo, a criação dos modelos por LDA que exige que as variáveis provenham de uma população normal. Por isso, a análise de componentes principais foi utilizada como uma etapa de pré-processamento dos dados que serão utilizados como entrada para os diversos modelos classificadores apresentados na seção 2.10, mas não no modelo por *ensemble* de redes neurais apresentado na seção 2.13.

Para o cálculo dos componentes principais, precisamos obter a matriz de correlação  $\vec{P}$  do conjunto de dados, ou seja, uma matriz quadrada com  $p$  linhas e  $p$  colunas (onde  $p$  é o número de variáveis originais retiradas do STING\_DB). Na matriz  $\vec{P}$ , seus elementos são formados pelo cálculo do coeficiente de correlação linear  $r$  dado pela equação (33), por exemplo, o elemento  $a_{16}$  é o valor de  $r$  entre as variáveis 1 e 6 do conjunto de dados. A diagonal principal de  $\vec{P}$  é idêntica a diagonal principal da matriz identidade.

De posse da matriz de correlação, é necessário obter os *autovalores* e *autovetores* da matriz. Os autovalores são facilmente identificados resolvendo a equação:

$$\det(\vec{P} - \lambda \vec{I}) = 0 \quad (56)$$

onde  $\det$  representa o determinante,  $\lambda$  representa o autovalor e  $\vec{I}$  a matriz identidade. De posse dos autovalores  $\lambda$ , voltamos ao problema original de autovalores e autovetores e resolvemos a equação:

$$\vec{P} \mathbf{v} = \lambda \mathbf{v} \quad (57)$$

onde  $\mathbf{v}$  representa o autovetor associado ao autovalor  $\lambda$ . Para uma matriz quadrada  $\vec{P}$  de dimensão  $p$ , há  $p$  autovalores e autovetores, sendo que alguns autovalores podem ser iguais e que autovetores diferentes podem ter o mesmo autovalor.

Os autovalores são organizados de forma que  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p$ . Quanto maior a magnitude do autovalor, maior será o comprimento do autovetor multiplicado pelo respectivo autovalor. Dessa forma, ao organizar os autovalores por ordem de magnitude, o primeiro autovetor carrega a maior variabilidade possível presente no conjunto de dados original.

A escolha do número de autovetores a serem utilizados pode seguir diferentes estratégias (Johnson, 1998). Adotamos o critério de variabilidade conjunta, ou seja, a soma da variabilidade presentes em cada componente principal. Escolhemos dessa forma o número de componentes principais que, quando somados, correspondem a 95% de toda a variabilidade presente no banco de dados original.

Cada par autovalor-autovetor resulta em uma nova variável chamada então de componente principal. Para o cálculo de cada componente principal, as grandezas estatísticas de média ( $\mu$ ) e variância ( $\sigma$ ) são estimadas do banco de dados. Em seguida são calculados os escores-Z:

$$z_{i,p} = \frac{x_{i,p} - \mu_p}{\sqrt{\sigma_p}} \quad (58)$$

Nesse caso, cada elemento do vetor  $\mathbf{x}_i$  é subtraído pelo respectivo elemento do vetor de médias  $\mu$  e dividido pela raiz quadrada do mesmo elemento do vetor de variâncias  $\sigma$ . O conjunto de todas as componentes resulta no vetor de escores-Z do  $i$ -ésimo elemento (aminoácido) do banco de dados,  $\mathbf{z}_i$ .

O escore de componente principal é definido como o produto interno entre cada autovetor  $\mathbf{v}_j$  e o escore-Z de cada indivíduo do dataset:

$$y_{i,j} = \mathbf{v}_j \cdot \mathbf{z}_i \quad (59)$$

A análise de componentes principais e o posterior cálculo dos escores foram realizados utilizando o pacote HSAUR2.0 do R (Everitt e Hothorn, 2008). Apenas os dados do conjunto de



treinamento (60% do banco de dados original, escolhidos aleatoriamente) foram usados para a obtenção dos autovalores e autovetores e em seguida os escores foram calculados para os demais dados do conjunto de treinamento.

## 2.12 Comparação entre classificadores por análise ROC

Vários critérios podem ser utilizados para comparar dois ou mais modelos classificadores. Utilizando a matriz de confusão descrita na seção 2.10, podemos estabelecer medidas de taxa de acerto (acurácia), precisão, sensibilidade (também chamado de *recall* ou *cobertura*) e mérito, definidos por:

$$\text{acerto (accuracy)} = A = \frac{TN+TP}{TN+TP+FN+FP} \quad (60)$$

$$\text{sensibilidade (sensitivity)} = S = \frac{TP}{TP+FN} \quad (61)$$

$$\text{precisão (precision)} = P = \frac{TP}{TP+FP} \quad (62)$$

$$\text{mérito (merit)} = M = \frac{S.P}{0,7.S+0,3.P} \quad (63)$$

Nenhuma dessas medidas pode ser considerada definitiva ou melhor do que outra. A taxa de acerto fornece uma medida da percentagem de resíduos de aminoácido corretamente classificados, no entanto utiliza o mesmo peso para resíduos de aminoácido da superfície livre quanto para os resíduos de aminoácido formadores de interface (que deveriam ser priorizados). Sensibilidade fornece uma medida utilizando *falsos negativos* (resíduos de aminoácido de interface que são erroneamente classificados como de superfície livre), e está, portanto, relacionada com a cobertura das interfaces preditas. Olhando apenas para a taxa de sensibilidade, ignoramos a quantidade de *falsos positivos* (resíduos de aminoácido de superfície livre que são erroneamente classificados como de interface). Com a taxa de precisão, temos uma medida utilizando *falsos positivos*, o que fornece uma medida da confiabilidade de classificação quando um resíduo de aminoácido é classificado como de interface. Dessa forma, levando em conta apenas a precisão, a cobertura da região de interface não é utilizada como critério de avaliação. A

taxa de mérito é taxa definida como uma combinação não balanceada entre as taxas de precisão e sensibilidade, favorecendo a taxa de precisão, ou seja, um decréscimo em precisão penaliza mais a taxa de mérito do que o mesmo decréscimo em sensibilidade.

Outra grandeza comumente utilizada pela comunidade de classificadores binários e é o coeficiente de correlação de Matthew (MCC, do inglês *Matthew correlation coefficient*), definido por:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (64)$$

No entanto, todas as medidas de desempenho apresentada nas equações 60-64 são dependentes de um valor limiar para a classificação. A maior parte dos modelos classificadores descritos na seção 2.10 (com exceção de árvores de decisão) retorna um valor que pode ser interpretado como uma probabilidade de determinado resíduo de aminoácido ser classificado como IFR. Quanto maior o limiar escolhido para a classificação (por exemplo, acima de 80% de probabilidade de ser IFR), menor é o número de *falsos positivos*. Da mesma forma, quanto menor o valor do limiar escolhido (por exemplo, acima de 20% de probabilidade), menor será o número de *falsos negativos*. Dependendo da aplicação desejada, o valor de probabilidade pode ser selecionado de forma a priorizar as taxas de precisão ou sensibilidade. Qualquer comparação feita utilizando as taxas descritas, leva em conta a escolha do limiar de classificação.

Por isso, utilizamos a análise ROC (do inglês *Receiver Operating Characteristics*). Esse tipo de análise utiliza gráficos de forma comparativa e é disponível para avaliação de classificadores binários (como os desenvolvidos e descritos na seção 2.10; Fawcett, 2006). De acordo com Fawcett (2006), o uso deste tipo de análise vem crescendo nos últimos anos na comunidade científica, principalmente pela constatação de que simplesmente taxa de acerto é uma forma pobre de comparação entre sistemas classificadores, sendo bastante útil na análise de casos onde há uma grande diferença entre a proporção das duas classes a serem classificadas (como é o caso desse estudo sobre classificação e que o número de FSR corresponde em média a aproximadamente 90% de todos os resíduos de aminoácido).

Curvas ROC são gráficos em duas dimensões onde no eixo *y* representamos a taxa *TP* e no eixo *x* a taxa *FP*. Quanto maior a taxa *TP* e menor a taxa *FP*, melhor é classificado o modelo. Um classificador gera simplesmente uma saída binária, que é a classe predita. Assim, dado um

conjunto de teste, a matriz de confusão resultará em pontos estáticos das taxas FP e TP, e, portanto, um ponto no gráfico ROC.

O limiar de classificação pode ser variado continuamente entre [0,1] e para cada limiar um ponto diferente do gráfico ROC será traçado, resultando em uma curva. Mesmo algoritmos como árvores de decisão, que não resultam em uma probabilidade de classificação, podem ser analisados com gráficos ROC. Em tais algoritmos, cada nodo final da árvore possui, normalmente, ambas as classes observadas (não há classificação perfeita) em uma determinada proporção específica para cada nodo. A fração de cada classe nos nós finais pode ser usada como um score (ou probabilidade) de cada indivíduo classificado pertença a cada uma das classes. Criamos as curvas ROC utilizando o pacote *ROCR* (Sing *et al.*, 2005) do software R.

Para uma curva ROC, oriunda de um classificador, quanto maior a área abaixo da curva (AUC, do inglês *Area Under the Curve*) mais eficiente o classificador pode ser considerado. Como a curva é gerada varrendo os possíveis limiares de classificação, esse critério é independente de qualquer valor limiar específico, sendo, portanto, uma medida do máximo poder de classificação de cada modelo classificador.

O pacote *ROCR* também oferece a possibilidade de cálculo de outras taxas encontradas na literatura, como todas as taxas mencionadas previamente. O critério de AUC fornece uma boa medida do desempenho de um determinado classificador, mas deve-se tomar cuidado, pois um classificador com AUC menor pode ter desempenho melhor para alguns valores do limiar (Fawcett, 2006).

### ***2.13 Classificadores por redes neurais artificiais com processo de seleção de variáveis e formação de ensemble de classificadores***

A comparação entre os modelos classificadores, como mostrado e discutido na seção de resultados, apontou a rede neural como o melhor modelo frente aos dados utilizados para a predição de resíduos de aminoácido formadores de **interfaces. Dessa forma, em colaboração com o Laboratório de Bioinformática e Computação Bio-Inspirada do Departamento de Engenharia da Computação e Automação Industrial da Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, liderado pelo professor Doutor Fernando Von Zuben,**

desenvolvemos um algoritmo baseado em redes neurais com processo de seleção de variáveis descrita nesta seção.

O desenvolvimento de um modelo classificador mais elaborado utilizou uma rede neural artificial hetero-associativa MLP (Rumelhart e McClelland, 1986), cuja estrutura é ilustrada na figura 11. O mecanismo de aprendizado escolhido é do tipo supervisionado, ou seja, como a saída desejada é conhecida, caso a predição para as entradas do conjunto de treinamento não esteja correta, os pesos das conexões sinápticas são alterados minimizando o erro associado. Para isso, foi utilizado o algoritmo de retro-propagação (do inglês *backpropagation*; Werbos, 1974). Este tipo de rede apresenta capacidade de aproximação universal, como descrito na seção 2.10.8.

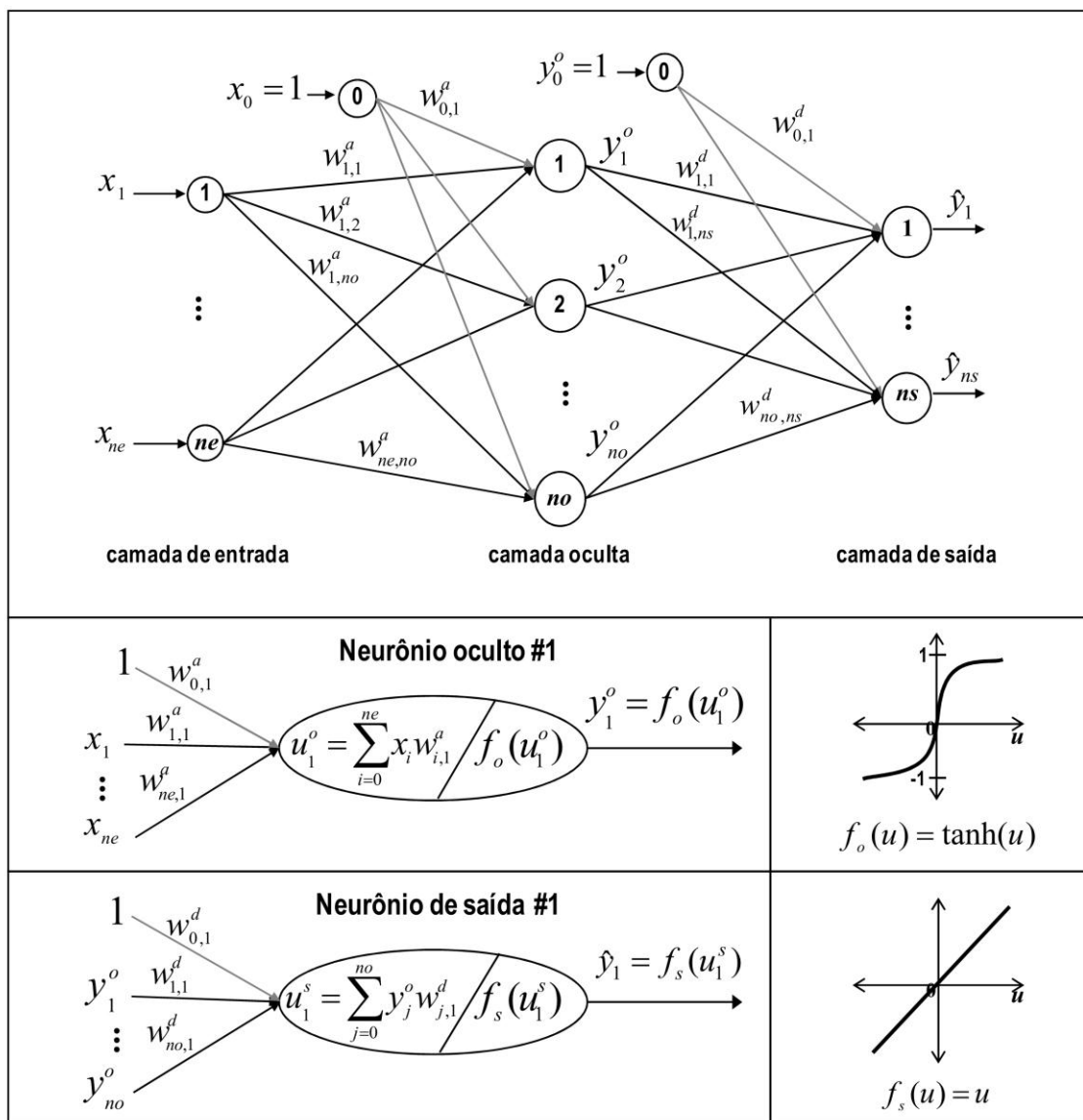
Antes de serem carregadas nos neurônios da camada de entrada ( $x_i$ ), as variáveis descritas na tabela 2 são escalonadas para o intervalo [-1, 1]. Apenas uma camada oculta foi utilizada, e o número de neurônios “no” foi varrido entre 8 e 15 neurônios, com incremento unitário em cada etapa de treinamento. Os atributos de entrada são linearmente combinados nos neurônios da camada oculta com os respectivos pesos sinápticos  $w_{ij}^a$  ( $i=1, 2, 3, \dots, ne$ ;  $j=1, 2, \dots, no$ ). Em cada neurônio da camada oculta, é produzida a ativação interna, dada por:

$$u_j^a = \sum_{i=0}^{ne} x_i w_{ij}^a \quad (65)$$

A entrada  $x_0 = 1$  é adicionada para adicionar flexibilidade no mapeamento não-linear multidimensional. A transferência de informação da camada oculta para a camada de saída é feita utilizando a função de ativação sigmoide *tangente hiperbólica* ( $y_j^o = f_{o,j} = \tanh(u_j^a)$ ). A saída de cada neurônio da camada oculta é ainda multiplicada pelos pesos  $w_{jk}^d$  ( $j=1, 2, \dots, no$ ;  $k=1, 2$ ), onde  $k$  é o número de padrões de treinamento, sendo que  $k = 1$  representa a classe FSR e  $k = 2$  representa a classe IFR. Dessa forma, em cada um dos dois neurônios da camada de saída, é produzida a ativação interna, dada por:

$$u_k^s = \sum_{j=0}^{no} y_j^o w_{jk}^d \quad (66)$$

Novamente, a entrada  $y_0^o = 1$  é adicionada para adicionar flexibilidade no mapeamento não-linear multidimensional. A ativação  $u_k^s$  serve de argumento para a função linear  $\hat{y}_k = f_{sk} = u_k^s$ , que é a saída em cada um dos dois neurônios da rede neural artificial.



**Figura 11 – Arquitetura da rede MLP utilizada com uma camada de entrada (dados provenientes da base de dados BlueStar STING), uma camada oculta e uma camada de saída.**

Fonte: Figura disponível no capítulo 2, pág. 13 da tese de doutorado de Wilfredo Jaime Puma Villanueva (Puma-Villanueva, 2011), colaborador dessa proposta de classificador.

O treinamento da MLP descrita consiste em aperfeiçoar os pesos das conexões sinápticas  $w_{ij}^a$  e  $w_{jk}^d$  de forma supervisionada, ou seja, utilizando a classe conhecida de cada um dos aminoácidos do conjunto de treinamento. Há algumas formas de aperfeiçoar os pesos utilizando o algoritmo de retro-propagação (do inglês *backpropagation*; Werbos, 1974). Para isso, definimos a função erro como a função objetivo que buscamos minimizar no processo de treinamento:

$$E = \frac{1}{2} \sum_p \sum_{k=1}^2 (y_k - \hat{y}_k)^2 \quad (67)$$

onde  $y_k$  é o valor da classe observada e  $\hat{y}_k$  o valor predito pela rede neural, e  $p$  é o número de resíduos de aminoácidos no conjunto de treinamento.

A condição de minimização implica que o gradiente da função erro seja nulo. Para isso, é necessário o cálculo das derivadas de  $E$  em relação aos pesos sinápticos depois da camada oculta, que queremos ajustar, usando a regra da cadeia:

$$\frac{\partial E}{\partial w_{jk}^d} = \sum_p \frac{\partial E}{\partial w_{jk}^d} = \sum_p \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial w_{jk}^d} = \sum_p \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial u_k^s} \frac{\partial u_k^s}{\partial w_{jk}^d} \quad (68)$$

onde  $p$  é o número de padrões de treinamento, ou seja, o número de aminoácidos no conjunto de treinamento. O mesmo procedimento é feito para os pesos antes da camada oculta:

$$\frac{\partial E}{\partial w_{ij}^a} = \sum_p \frac{\partial E}{\partial w_{ij}^a} = \sum_p \frac{\partial E}{\partial y_j^o} \frac{\partial y_j^o}{\partial w_{ij}^a} = \sum_p \frac{\partial E}{\partial y_j^o} \frac{\partial y_j^o}{\partial u_j^a} \frac{\partial u_j^a}{\partial w_{ij}^a} \quad (69)$$

Após o cálculo das derivadas, o método quase-Newton de otimização não linear foi empregado para atualizar os pesos. Nesse método, informação de segunda ordem da matriz Hessiana  $\vec{H}$ , ou seja, segunda derivada da função erro em relação aos pesos sinápticos, é estimada a partir do vetor gradiente calculado pelas equações (68) e (69) (Puma-Villanueva, 2011). Após cada etapa de treinamento  $t$ , cada peso é então atualizado de acordo com as equações:

$$\begin{cases} w_{ij}^a(t+1) = w_{ij}^a(t) - \alpha \vec{H}^{-1} \frac{\partial E}{\partial w_{ij}^a} \\ w_{jk}^d(t+1) = w_{jk}^d(t) - \alpha \vec{H}^{-1} \frac{\partial E}{\partial w_{jk}^d} \end{cases} \quad (70)$$

Onde  $\alpha$  recebe o nome de *passo*, e tem como objetivo modular a variação de cada peso, ou seja, impedir grandes variações em apenas uma etapa de treinamento.

Cada rede neural treinada, isto é, com pesos sinápticos fixos, tem como saída dois valores, um para cada neurônio da camada de saída. No processo de treinamento, os resíduos de aminoácidos pertencentes à classe FSR são utilizados para direcionar o ajuste dos pesos de forma que o primeiro neurônio da camada de saída retorne o valor 1, enquanto que o segundo neurônio

retorne o valor 0. De forma oposta, os resíduos de aminoácidos pertencentes à classe IFR direcionam o ajuste dos pesos sinápticos de forma que o primeiro neurônio da camada de saída seja 0 e o segundo neurônio seja 1. Dessa forma, o processo de classificação de uma rede neural treinada com essa arquitetura, classifica um aminoácido de entrada como FSR se o valor do primeiro neurônio da camada de saída for maior que o valor do segundo neurônio da camada de saída, e classifica como IFR caso contrário.

Atrelado à síntese de redes neurais artificiais treinadas, empregamos uma metodologia para selecionar quais variáveis de entrada (processo de seleção de variáveis) são melhor combinadas para diferenciar os resíduos de aminoácido da superfície livre dos que estão nas interfaces de complexos proteicos. Entre as abordagens disponíveis, utilizamos o método conhecido como *envoltório* (do inglês *wrapper*). Nessa abordagem, o processo de seleção de variáveis interage com o processo de treinamento do modelo classificador (Puma-Villanueva, 2006), o que oferece vantagens sobre outros métodos disponíveis, como o método de *filtro*, porém com maior custo computacional.

No início do processo de seleção de variáveis, todos os descritores escolhidos do STING\_DB são utilizados como entrada para a rede neural. Após a primeira etapa de treinamento, ou seja, ajuste dos pesos sinápticos de acordo com as equações (65)-(70), cada uma das variáveis de entrada é sistematicamente substituída pela sua média, obtida no conjunto de treinamento. Assim, toda a variabilidade presente nesse descritor é eliminada e o desempenho medido em um conjunto de validação com e sem o descritor testado é guardado. Esse processo é repetido com todos os descritores de entrada de forma que uma lista da mudança de desempenho é obtida. Ao testar a “sensibilidade” de predição de cada uma dos descritores, aquele que se mostrar menos sensível, ou seja, aquele que menos interfere na predição do conjunto de validação, é retirado do conjunto de entrada. Como o conjunto inicial é composto por todas as variáveis disponíveis para cada tipo de aminoácido e cada etapa consiste na remoção de um descritor, dizemos que esse é um mecanismo de poda (Guyon e Elisseeff, 2003).

Nesse ponto, uma nova rede neural artificial é treinada sem a variável que foi removida, obtendo assim novos pesos sinápticos. O processo de seleção de variáveis e sucessivos treinamentos de redes neurais é feito até que apenas um descritor permaneça no conjunto de treinamento. O nome *envoltório* relaciona ao fato do processo de treinamento e seleção de variáveis acontecerem simultaneamente.

A ordem de remoção dos descritores implica na importância de cada descritor para o processo de predição. Descritores relevantes para o processo de classificação influenciam no desempenho do modelo classificador. Por outro lado, descritores pouco relevantes não influenciam no desempenho do modelo classificador. Assim, descritores que são removidos nas etapas iniciais são menos relevantes para a distinção entre IFR e FSR, frente aos descritores que ainda permanecem no processo de treinamento. A taxa de mérito, equação (63), é utilizada como índice de desempenho para escolher qual descritor será removido em cada etapa.

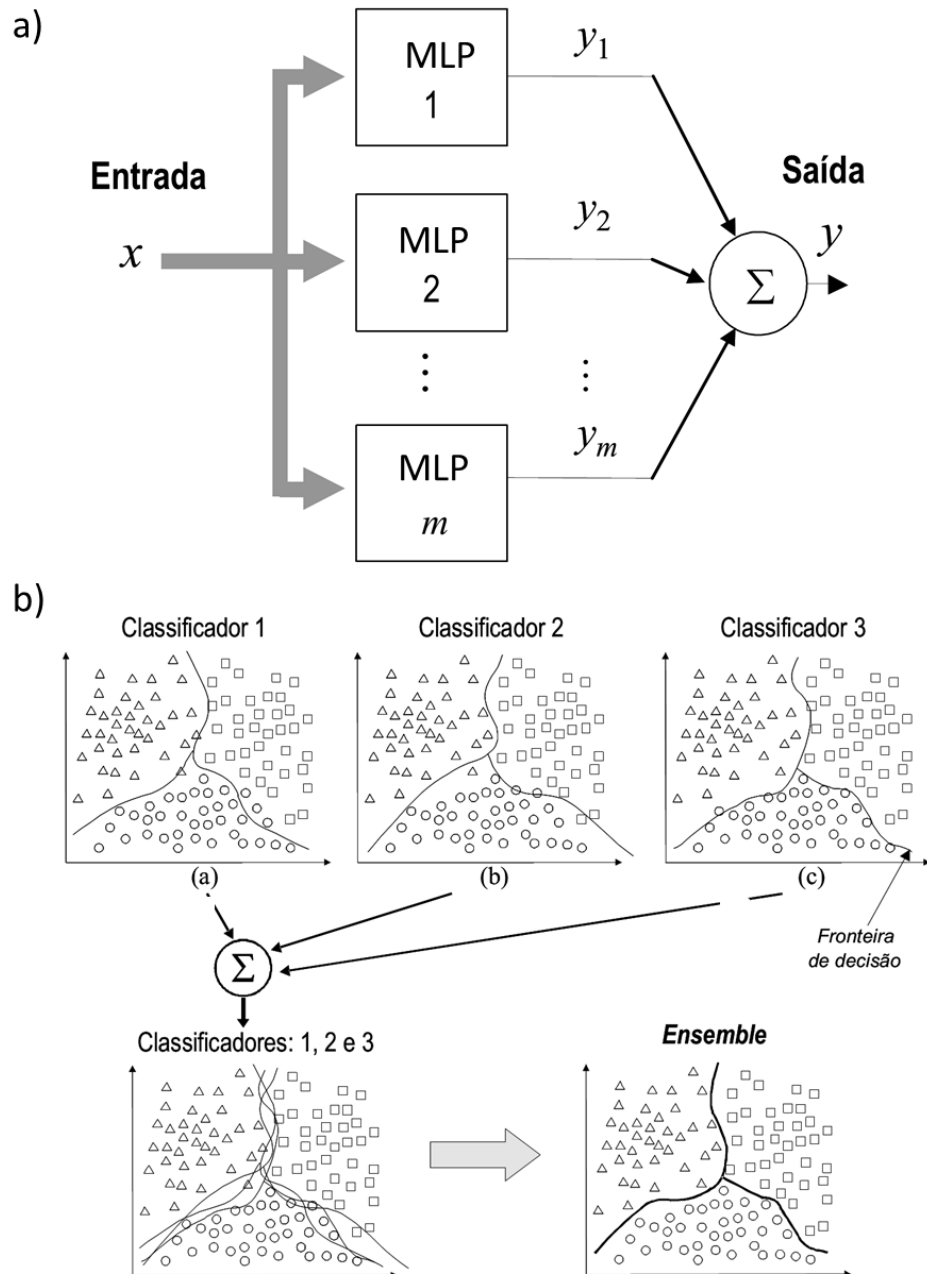
Durante todo processo de seleção de variáveis, centenas de modelos classificadores são criados e armazenados. Dessa forma, foi desenvolvida uma abordagem de criação de *ensemble* de classificadores, ou seja, um comitê de redes neurais diferentes é utilizado em conjunto para classificar cada um dos resíduos de aminoácido. Para isso, foi utilizado o conceito de voto majoritário, onde cada componente do *ensemble* de redes neurais faz a classificação de cada resíduo a ser classificado em uma das classes, IFR ou FSR, e a classe que receber mais votos é a classificação final. Em caso de empate, o resultado da rede neural com melhor desempenho é tido como a classificação final.

Para empregar a metodologia aqui proposta, dividimos o conjunto de *teste* em um conjunto extra, denominado conjunto de *seleção*. O conjunto de *teste*, como descrito na seção 2.2, é composto por 40% do conjunto de dados original. Com a incorporação do novo conjunto, a divisão foi feita de forma que 25% do conjunto original foi utilizado como conjunto de *teste* e 15% como conjunto de *seleção*, que é então utilizado para selecionar o número e a ordem dos componentes do *ensemble*. A rede neural com melhor desempenho é escolhida como primeiro componente do comitê. Em seguida, 10% das centenas de redes neurais criadas no processo de seleção de variáveis são sistematicamente combinadas com a melhor rede neural. A rede neural que resultar no melhor aumento de desempenho, em relação à taxa de mérito, é escolhida como a segunda componente do comitê. Esse processo é repetido até que nenhum aumento de desempenho seja observado por duas inclusões consecutivas de novos componentes.

O uso de *ensemble* de classificadores é escolhido para fins de aumento de desempenho (Puma-Villanueva, 2006). Isso é devido ao fato de que é esperado que classificadores simples falhem para uma certa porção dos dados de testes, mas classificadores diferentes tendem a falhar com entradas diferentes, como ilustrado na figura 12. Ao combinar as respostas de diferentes componentes do comitê de classificadores, é esperado que a classificação contenha menos resíduos de aminoácidos classificados erroneamente. Como a taxa de mérito é escolhida para



determinar o número e a ordem dos componentes do *ensemble* de redes neurais, a taxa de precisão deve ser favorecida e aumentada em relação a classificação pela melhor rede neural isolada.



**Figura 12 – Criação do ensemble de redes neurais como método para evitar sobre-ajuste aos dados de treinamento.**

A saída de cada componente do comitê de classificadores combinam-se para resultar em uma saída de classificação (a). Cada rede neural possui uma fronteira de decisão diferente que, quando combinadas, tendem a reduzir o erro de classificação (b).

**Fonte:** Figuras disponíveis no capítulo 3, páginas 49 e 55 da tese de mestrado de Wilfredo Jaime Puma Villanueva (Puma-Villanueva, 2006), colaborador dessa proposta de classificador.

O teste final dos classificadores é feito com um conjunto de teste que não foi utilizado para treino/validação da rede neural artificial e que também não foi utilizada como conjunto de seleção dos componentes do *ensemble*. O desempenho do *ensemble* de redes neurais nesse conjunto teste é representativo do valor real de classificação dessa proposta de classificador. No entanto, a comparação direta com outros modelos disponíveis na literatura deve ser feita baseada no mesmo banco de dados. Para isso, o conjunto de *benchmark* foi utilizado como conjunto de teste final.

O processo de treinamento da rede neural com seleção de variáveis é de alto custo computacional. Utilizando o conjunto de dados DS30, para cada tipo aminoácido foi necessário uma média de 10 horas de treinamento em uma máquina com processador *quad-core* 3GHz e 2 Gigabits de memória RAM. Como há 20 tipos de aminoácidos, o processo levou cerca de 200 horas para ser finalizado. Ainda, utilizamos a técnica estatística de validação cruzada, ou seja, o processo foi repetido 10 vezes. Por fim, realizamos o treinamento com três conjuntos de dados: sem e com descritores ponderados pela vizinhança e acrescidos dos descritores de conservação. No total, somam-se aproximadamente seis mil horas de cálculo.

A criação do *ensemble* de redes neurais e avaliação de desempenho no conjunto teste é expressivamente menos exigente, finalizando cada uma das tarefas em segundos. O mesmo é válida para a aplicação do classificador final em novos casos de interesse biológico.

# 3 Resultados e Discussão

## 3.1 *Correlação linear entre os descritores*

Antes de executar os testes estatísticos uni e multivariados propostos e criação dos modelos classificadores de IFR e FSR, analisamos os dados para averiguar quais variáveis são consideradas linearmente independentes umas as outras. Algumas variáveis são conhecidas antecipadamente como sendo diretamente correlacionadas, por exemplo, os descritores sobre *número de contatos (usados e não usados)* e suas respectivas *energias*, que são obtidas pela multiplicação de uma constante.

A tabela 5 resume os resultados das variáveis que foram removidas por apresentarem correlação linear acima de 0,85 com outra variável. Apenas as variáveis consideradas não correlacionadas linearmente foram utilizadas nas etapas posteriores.

**Tabela 5 – Avaliação dos descritores físico-químicos e estruturais da base de dados BlueStar STING em relação a normalidade de distribuição dos valores e correlação linear entre si.** As variáveis assinaladas como “x” são excluídas por apresentarem coeficiente de correlação linear maior que 0,85 com outra variável presente no banco de dados. O rótulo “-” foi atribuído para as variáveis que não são definidas para determinados tipos de aminoácidos.

Descritores	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Acessibilidade	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Acessibilidade Relativa	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Energia Total de Contatos																				
Hidrofobicidade																				
CED no Ca (3)																				
CED no Ca (4)				x																
CED no Ca (5)				x		x			x											
CED no Ca (6)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CED no Ca (7)																				
CED no LHA (3)								--												
CED no LHA (4)	x	x		x	x	x		--	x	x	x	x	x	x	x		x	x		x
CED no LHA (5)		x		x	x	x	x	--	x	x	x	x	x	x	x		x	x	x	x
CED no LHA (6)	x	x	x	x	x	x	x	--	x	x	x	x	x	x	x	x	x	x	x	x
CED no LHA (7)								--												
CPO no C-alpha																				

Descritores	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
CPO no C-beta								--												
CPO no LHA								--												
CLO no C-alpha																				
CLO no C-beta	x		x	x	x	x		--	x	x	x	x	x		x	x	x		x	x
CLO no LHA	x				x			--	x	x	x					x				
Densidade no Ca (3)																				
Densidade no Ca (4)						x					x	x	x	x	x	x			x	x
Densidade no Ca (5)	x	x		x			x	x	x	x	x	x	x	x	x	x	x	x	x	x
Densidade no Ca (6)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Densidade no Ca (7)																				
Densidade no LHA (3)								--												
Densidade no LHA (4)								--												
Densidade no LHA (5)					x			--												
Densidade no LHA (6)		x	x	x	x	x	x	--	x	x	x	x	x	x	x	x	x	x	x	x
Densidade no LHA (7)	x							--												
EP no Ca																				
EP no LHA								--												
EP na Superfície																				
EP Médio	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no Ca (3)																				
Esonja no Ca (4)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no Ca (5)	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no Ca (6)		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no Ca (7)	x																			
Esonja no LHA (3)								--												
Esonja no LHA (4)			x	x				--						x					x	
Esonja no LHA (5)	x	x	x	x	x	x	x	--	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no LHA (6)	x	x	x	x	x	x	x	--	x	x	x	x	x	x	x	x	x	x	x	x
Esonja no LHA (7)								--												
Energia Total de Contatos Não-usados																				
PHI																				
PSI																				
CHI 1	--							--							--					
CHI 2	--				--			--							--	--	--			--
CHI 3	--	--	--	--	--			--	--	--	--			--	--	--	--	--	--	--
CHI 4	--	--	--	--	--	--	--	--	--	--	--			--	--	--	--	--	--	--

Descritores	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Energia de Contatos Hidrofóbicos																				
Energia de Carregados atrativos	--				--			--		--	--		--	--	--					--
Energia de Carregados repulsivo	--				--			--		--	--		--	--	--					--
Energia de Ligação de Hidro - MM																				
Energia de Ligação de Hidro - MS																				
Energia de Ligação de Hidro - SS	--							--		--	--		--	--	--					--
Energia de Ligação de Hidro - MWM																				
Energia de Ligação de Hidro - MWS																				
Energia de Ligação de Hidro - SWS	--							--		--	--		--	--	--					--
Energia de Ligação de Hidro - MWWM																				
Energia de Ligação de Hidro - MWWS																				
Energia de Ligação de Hidro - SWWS	--							--		--	--		--	--	--					--
Energia de Contatos Aromáticos	--		--	--	--			--		--	--		--		--	--	--			--
Energia de Ponte Dissulfeto	--	--	--	--		--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Energia de Contatos Hidrofóbicos Não-usada																				
Energia de Carregados atrativos Não-usada	--				--			--		--	--		--	--	--					--
Energia de Carregados repulsivo Não-usada	--				--			--		--	--		--	--	--					--
Ligação de Hidro - MM Energia Não-usada																				
Ligação de Hidro - MS Energia Não-usada																				
Ligação de Hidro - SS Energia Não-usada	--							--		--	--		--	--	--					--
Ligação de Hidro - MWM E. Não-usada																				
Ligação de Hidro - MWS E. Não-usada																				
Ligação de Hidro - SWS Energia Não-usada	--							--		--	--		--	--	--					--

Descritores	ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Ligação de Hidro - MWWW E. Não-usada																				
Ligação de Hidro - MWWS E. Não-usada																				
Ligação de Hidro - SWWS E. Não-usada	--							--		--	--		--	--	--					--
Energia de Contatos Aromáticos Não-usada	--		--	--	--			--		--	--		--	--	--	--	--			--
Energia de Ponte Dissulfeto Não-usada	--	--	--	--		--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

**Rótulo:** CED = Densidade de Energia de Contatos; CPO = Ordem Cross Presence; CLO = Ordem de Cross Link; EP = Potencial Eletrostático; C $\alpha$  = Carbono alfa; LHA = Último átomo pesado da cadeia lateral; C $\beta$  = Carbono Beta; MM = cadeia principal-cadeia principal; MS = cadeia principal-cadeia lateral; SS = cadeia lateral-cadeia lateral; W = uma molécula de água; WW = duas molécula de água;

### 3.2 Testes Estatísticos de normalidade

Realizamos o teste de normalidade D'Agostino (Zar, 1999) para todos os descritores de todas as tabelas separadas por tipo de aminoácido. Apenas uma pequena porção dos descritores utilizados obteve como resultado um *valor p* maior que 1%, sendo que o restante dos descritores apresentou um *valor p* sempre menor que  $10^{-5}$ . Dessa forma, apenas os descritores presentes na tabela 6 seguem uma distribuição normal com menos de 1% de probabilidade de estarmos errados, de acordo com o teste realizado.

**Tabela 6 – Lista de variáveis que obtiveram maiores valores de valor p (acima de 1%) no teste de normalidade de D'Agostino, separados de acordo com o tipo de aminoácido.**

Para o restante de descritores valores abaixo de  $10^{-5}$  foram obtidos. Apenas os descritores abaixo seguem uma distribuição normal, de acordo com o teste de D'Agostino.

	Variável	valor p	Variável	valor p	Variável	valor p
<b>Ala</b>	ep_lha	0,8439				
<b>Arg</b>	chi_4	0,4182				
<b>Asn</b>	psi	0,5311	hb_swws_uc	0,1222		
<b>Gly</b>	ep_ca	0,0194				
<b>Ile</b>	ep_lha	0,3983				
<b>Lys</b>	acc_isolacão	0,6324	acc_relativa	0,6637	hidrofobicidade	0,6524
	chi_3	0,7153	chi_4	0,1243		
<b>Thr</b>	esponja_lha7	0,0115				
<b>Trp</b>	densidade_lha5	0,7316	densidade_lha6	0,2786		

Esse resultado indica que o resultado do teste paramétrico de Welch pode não ser confiável, mesmo com o teorema do limite central sendo válido para nosso conjunto de dados com dezenas de milhares de observações em cada amostra. Em todo caso, os resultados dos testes não-paramétricos são mostrados em conjunto com os resultados do teste paramétrico de Welch.

### 3.3 Testes estatísticos uni e multivariados

A tabela 7 mostra o *valor p* de cada um dos três testes estatísticos univariados para cada um dos descritores do resíduo de aminoácido alanina. Em vermelho, destacamos os descritores que obtiveram um valor *p* maior do que 1%, ou seja, os descritores que falham em diferenciar IFR de FSR com mais do que 1% de estarmos errados em relação a distinção.

**Tabela 7 - Resultados dos testes estatísticos univariados para cada um dos descritores do resíduo de aminoácido alanina relativo ao seu poder de distinguir IFRs dos FSRs.**

*Em vermelho são destacados os descritores que apresentam o valor de mais de 1%, indicando pouco potencial estatístico para distinguir IFR dos FSR.*

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,0011	0	0
Energia de Ligação de Hidro - MWWS	0,3313	0	0,2526
Energia de Ligação de Hidro - MWWM	0,0022	0	0,1483
Energia de Ligação de Hidro - MWS	0,2321	0	0,0643
Energia de Ligação de Hidro - MWM	3,00E-04	0	8,00E-04
Energia de Ligação de Hidro - MS	0,46	0	6,00E-04
Energia de Ligação de Hidro - MM	0	0,8429	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha5	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha6	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca4	0	0	0
Esponja no ca6	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0044	0,1278	0
Energia de Ligação de Hidro – MWWS não usada	0,3345	0,0078	0,0017
Energia de Ligação de Hidro – MWWM não usada	0,0031	0,0785	0,043
Energia de Ligação de Hidro – MWS não usada	0,0252	0,0464	0,0022
Energia de Ligação de Hidro – MWM não usada	4,00E-04	0,0458	7,00E-04
Energia de Ligação de Hidro – MS não usada	0,4707	0,5954	0,361
Energia de Ligação de Hidro – MM não usada	0,0147	0,3416	1,00E-04
Energia de Contatos Hidrofóbicos não usados	2,00E-04	0	0
phi	0,9194	0,2884	0,0979
psi	0,5953	0,8087	0,2512

Para os outros tipos de aminoácido, apresentamos uma tabela similar como apêndice. A maior parte dos descritores com mais do que 1% do valor p nos testes realizados pertencem a categoria *contatos utilizados e não-usados* e *potencial eletrostático no LHA* (em alguns testes). Este resultado indica que para a maioria das variáveis utilizadas para cada tipo de aminoácido tem poder estatístico para discriminar entre IFR e FSR.



Utilizando todas as variáveis, excluindo os descritores considerados linearmente dependentes, a análise de variância multivariada (MANOVA) foi calculada utilizando o comando do software R (pacote HSAUR2.0) ilustrado no quadro 1.

#### Quadro 1 – comando em R para cálculo das matrizes de MANOVA.

```
> ala_manova <- manova(as.matrix(cbind(ala[,1:(ncol(ala)-2)]))~ifr, data=ala)
```

Comando do pacote HSAUR 2.0 do software R.

Para o resíduo de aminoácido alanina, foi utilizado um total de 46 variáveis independentes (não necessariamente ortogonais) e tendo a variável binária “ifr” como referência para as classes IFR e FSR. Os quatro testes estatísticos após a MANOVA foram feitos, tendo como saída as informações no quadro 2.

#### Quadro 2 – comando em R para cálculo dos testes estatísticos multivariados da MANOVA.

```
> summary(ala_manova, test = "Pillai")
      Df Pillai approx F num Df den Df    Pr(>F)
ifr      1 0.4287  1113.3     46 68248 < 2.2e-16 ***
Residuals 68293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ala_manova, test = "Wilks")
      Df Wilks approx F num Df den Df    Pr(>F)
ifr      1 0.5713  1113.3     46 68248 < 2.2e-16 ***
Residuals 68293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ala_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
ifr      1      0.75039  1113.3     46 68248 < 2.2e-16 ***
Residuals 68293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ala_manova, test = "Roy")
      Df Roy approx F num Df den Df    Pr(>F)
ifr      1 0.75039  1113.3     46 68248 < 2.2e-16 ***
Residuals 68293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

Assim como para o resíduo de aminoácido alanina (mostrado acima), todos os tipos de aminoácido obtiveram *valor p* menores  $2,2 \times 10^{-16}$ . Os resultados para os demais tipos de

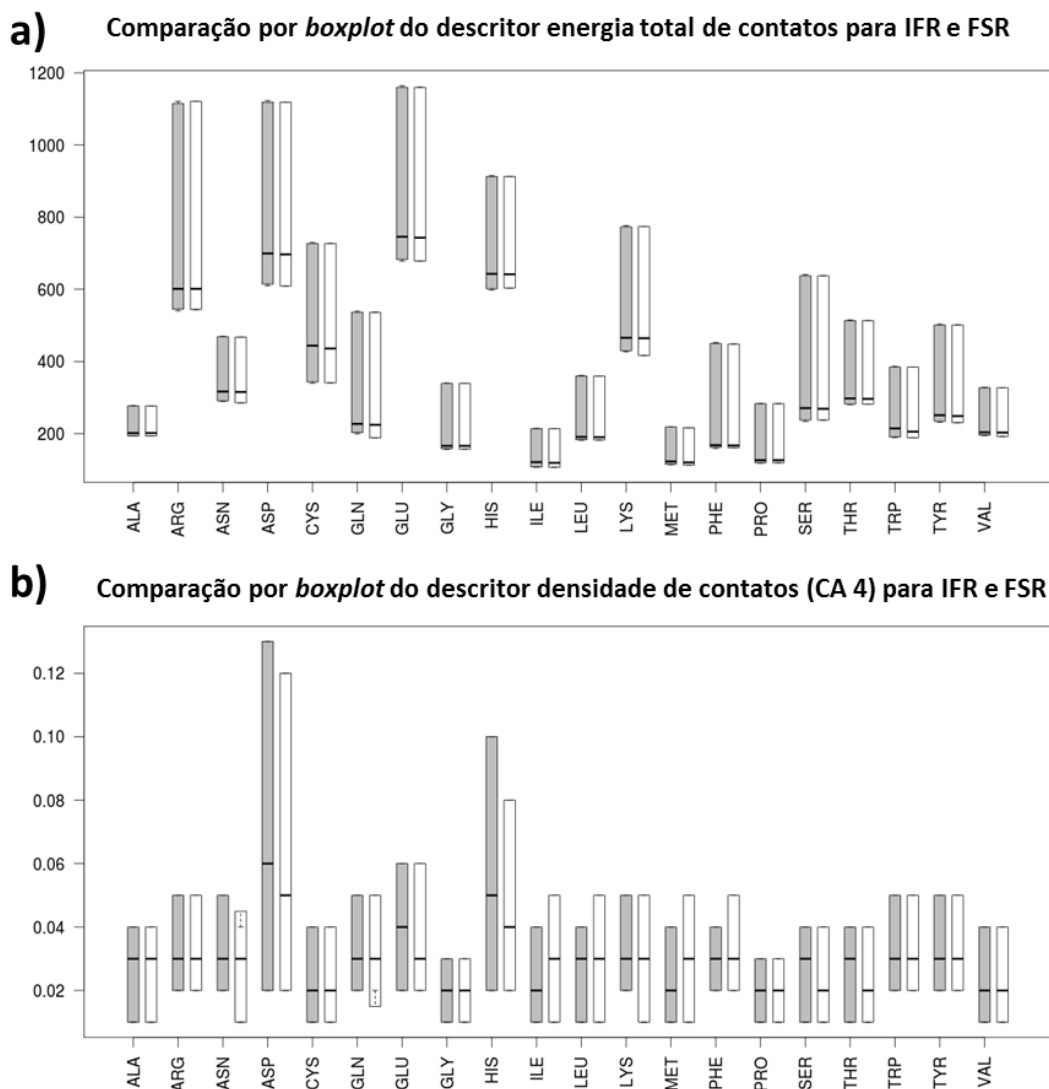
aminoácido são encontrados como apêndice. Um *valor p* tão baixo indica que uma abordagem multivariada é capaz de distinguir entre IFR e FSR.

### **3.4 Gráficos do tipo *boxplot***

Para auxiliar os testes de normalidade, *boxplots* fornecem uma forma visual sobre a distribuição. Utilizando os gráficos *boxplots* foi possível verificar quais variáveis não seguem um padrão de distribuição normal. *Boxplots* também podem ser utilizados para avaliar quais descritores são bem sucedidos em distinguir as classes IFR de FSR, para cada variável no DS95, para cada um dos 20 tipos de aminoácido. A figura 13 mostra o perfil de cada um dos aminoácidos para os descritores *energia total de contatos* e *densidade de energia no CA* (4). Os gráficos dos demais descritores estão disponíveis como apêndice e discutidos a seguir. Alguns descritores apresentam o mesmo padrão para ambas as classes IFR e FSR, como é caso dos descritores *Potencial Eletrostático*, *Energia Contatos Estabelecidos* e *Não Estabelecidos*. Para os descritores de *Densidade de Energia de Contatos* e *Densidade*, observamos uma maior diferença entre as médias das classes, assim como também uma diferença entre os quartis. É interessante observar que apesar de descritores específicos para a energia de contatos não detectarem diferenças entre IFR e FSR o descritor de *Densidade de Energia de Contatos*, que descreve o nano-ambiente ao redor do resíduo de aminoácido, consegue detectar essa diferença. Isso indica que, em média, uma ação cooperativa (da vizinhança) é importante para o estabelecimento de interfaces em complexos proteicos.

No entanto, entendemos que resíduos de aminoácido específicos conhecidos como *hot spot* (resíduos de aminoácido com elevado energia de interação na interface) podem desviar da tendência geral dos demais resíduos de aminoácidos quando olhamos descritores sobre energia de contatos. Isso é devido ao fato que a energia das ligações estabelecidas com resíduos de aminoácidos de outra cadeia proteica mantém o complexo estável. Energias de ligação podem variar de décimos de kcal/mol até algumas unidades de kcal/mol. O limite arbitrário para a energia de ligações através da interface de resíduos de aminoácidos para serem considerados *hot spots* é de pelo menos 2 kcal/mol. Apesar de *hot spots* serem extremamente importantes para a estabilidade de complexos proteicos, nenhuma alusão é feita nesse estudo para esses resíduos de aminoácidos, sendo que em nosso estudo o número de IFR é muito maior do que o número de *hot*

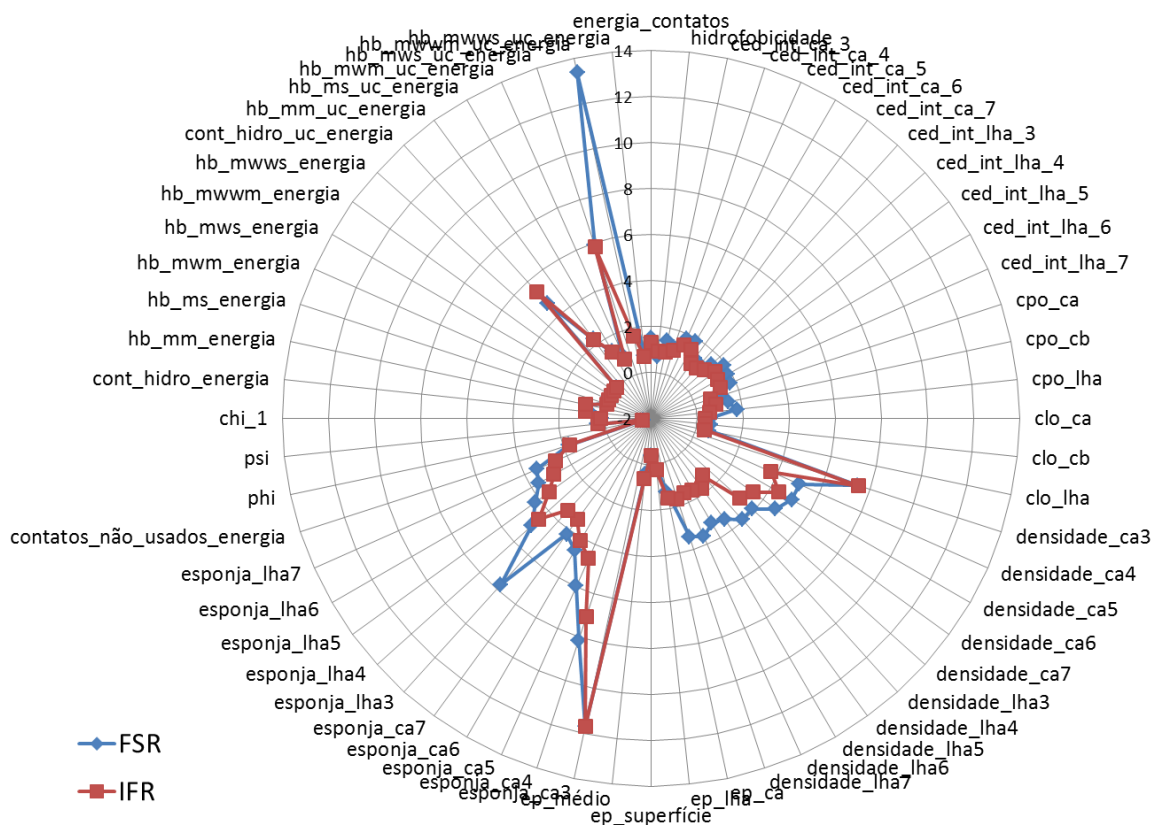
spots, o que deve acobertar o efeito desses resíduos de aminoácido que são estatisticamente sub-representados no conjunto de dados. Mais detalhes sobre *hot spots* são obtidos na dissertação de mestrado de Pereira (2012).



**Figura 13 – Comparação entre a distribuição dos descritores “energia total de contatos” (a) e “densidade de energia de contatos no CA-4” (b) em relação as classes IFR (branco) e FSR (cinza). Barras de mínimo e máximo foram omitidas.**

De forma similar, podemos averiguar o comportamento médio de vários descritores de cada tipo de aminoácido em relação às classes IFR e FSR. A figura 14 ilustra o padrão dos valores médios de cada descritor dividido pelo desvio padrão para o resíduo de aminoácido valina. Apesar de

algumas diferenças serem observadas para os descritores *densidade*, *esponjicidade* e *energia de contatos de ligação de hidrogênio*, observar apenas o comportamento médio de cada descritor mascara a dispersão dos valores de cada parâmetro como observado nos gráficos do tipo *boxplot* apresentados na figura 13 e no apêndice 6.3 (os gráficos em apêndice também ilustram o comportamento da média dos descritores com linhas sólidas e tracejadas).



**Figura 14 – Comparação entre os valores médios divididos pelo desvio padrão de cada descritor físico-químico ou estrutural para as classes FSR (azul) e IFR (vermelho) para o resíduo de aminoácido valina.**

Apesar de diferenças serem observadas, principalmente para o valor médio, os alcances da maior parte dos descritores se interpolam, indicando que em uma tentativa preditiva, vários resíduos de aminoácido seriam classificados erroneamente. Assim, o uso de técnicas multivariadas se faz fundamental.

Apenas para alguns descritores de alguns tipos de aminoácido podem ser observados uma separação efetiva entre as classes. O descritor *Cross Presence Order* separa os IFRs dos FSRs para

os aminoácidos cisteína, histidina, isoleucina, leucina, metionina, fenilalanina, triptofano, tirosina e valina (ver figuras no apêndice). Para o resíduo de aminoácido triptofano a variável *Espanja no LHA (3)* mostra a melhor divisão entre as classes estudadas. Uma divisão parcial é observada para o mesmo descritor no resíduo de aminoácido ácido aspártico.

### **3.5 Análise dos descritores por modelo de regressão linear multivariado**

Após a exclusão das variáveis correlacionadas, 20 modelos de regressão linear, um para cada aminoácido, foram gerados com o software R.

Para toda variável, é mostrado seu coeficiente  $\beta_j$ , assim como o intercepto  $\alpha$ . Ao final é mostrado o *valor p* do teste estatístico utilizando a distribuição *chi-quadrado* (estatística *F*). Para o resíduo de aminoácido alanina, o *valor p* do modelo de regressão mostrou-se menor do que  $2,2 \times 10^{-16}$ . Isso indica que a probabilidade de todos os coeficientes  $\beta_j$  serem simultaneamente iguais a zero é muito baixa.

O software R também mostra os resultados dos testes estatísticos da distribuição *t* de Student para cada um dos coeficientes  $\beta_j$ , indicando o *valor p* de cada coeficiente ser diferente de zero. As seguintes variáveis obtiveram uma probabilidade relativamente alta (acima de 10%) de serem descartadas: *ced no CA 4*, *ced no CA 5*, *clo no CA*, *ep no LHA*, *ep na superfície*, *energia de hb\_ms*, *energia de hb\_mwma*, *energia de hb\_mwwma*, *energia de hb\_mwws*, *energia de cont. hidrofóbicos\_uc* e *energia de hb\_mws\_uc*.

O quadro 3 mostra o valor do *coeficiente de determinação* ( $r^2$  – do inglês *Multiple R-squared*). Para o resíduo de aminoácido alanina, apenas 44,88% da variabilidade das classes IFR e FSR é explicada pelo modelo adotado. Embora esse valor seja um bom indicativo da qualidade do modelo gerado em explicar os dados observados, Zar (1999) comenta que uma melhor grandeza de comparação (principalmente quando se compara modelos com diferente número de variáveis e também diferentes número de entradas) é através do uso da grandeza chamada de *coeficiente de determinação ajustado* (Adjusted R-squared) e definida por:

$$r_a^2 = 1 - \frac{\text{quadrados médio dos resíduos}}{\text{quadrado médio totais}} \quad (70)$$

### Quadro 3 – análise por regressão linear

```

> ala_lm <- lm(ifr~as.matrix(cbind(ala[,1:(ncol(ala)-2)])), data=ala)
> summary(ala_lm)

Call:
lm(formula = ifr ~ as.matrix(cbind(ala[, 1:(ncol(ala) - 2)])), data = ala)

Residuals:
    Min       1Q   Median       3Q      Max
-1.31035 -0.18693 -0.08157  0.04091  1.25286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.527e-01  1.919e-02  34.006 < 2e-16 ***
total_contact_energy  4.877e-04  2.803e-04   1.740 0.081868 .
hb_mwws_energy  2.493e-03  1.220e-03   2.043 0.041024 *
hb_mwmm_energy -2.002e-03  1.305e-03  -1.535 0.124904
hb_mws_energy  2.397e-03  1.110e-03   2.160 0.030768 *
hb_mwm_energy -1.035e-03  1.029e-03  -1.006 0.314570
hb_ms_energy -1.448e-03  8.803e-04  -1.645 0.100029
hb_mm_energy  2.609e-03  3.951e-04   6.603 4.06e-11 ***
hydrophobic_energy -1.131e-02  1.133e-03  -9.981 < 2e-16 ***
hydrophobicity  1.639e-01  1.327e-02  12.347 < 2e-16 ***
ced_ca3        3.312e-01  4.318e-02   7.670 1.74e-14 ***
ced_ca4        8.325e-02  6.092e-02   1.367 0.171754
ced_ca5        1.068e-01  7.050e-02   1.515 0.129653
ced_ca7       -6.166e-02  9.484e-02  -0.650 0.515620
ced_lha3      -9.161e-02  3.973e-02  -2.306 0.021123 *
ced_lha5     -2.520e-01  6.009e-02  -4.194 2.75e-05 ***
ced_lha7     -3.878e-01  1.050e-01  -3.693 0.000221 ***
cpo_ca       -2.466e-03  2.595e-03  -0.950 0.342056
cpo_cb       -3.606e-03  2.724e-03  -1.324 0.185554
cpo_lha     -1.234e-02  1.917e-03  -6.440 1.20e-10 ***
clo_ca      -4.794e-03  2.829e-03  -1.695 0.090164 .
density_ca3  -6.237e-02  1.533e-02  -4.069 4.73e-05 ***
density_ca4  -3.090e-01  2.517e-02 -12.275 < 2e-16 ***
density_ca7  -1.321e-01  3.930e-02  -3.361 0.000776 ***
density_lha3  3.919e-01  1.533e-02  25.568 < 2e-16 ***
density_lha4  4.119e-01  1.494e-02  27.573 < 2e-16 ***
density_lha5  3.702e-01  1.546e-02  23.945 < 2e-16 ***
density_lha6  4.231e-01  1.749e-02  24.189 < 2e-16 ***
ep_ca        1.235e-03  2.321e-04   5.322 1.03e-07 ***
ep_lha     -1.051e-04  2.652e-04  -0.396 0.691923
ep_surf     4.480e-04  2.901e-04   1.544 0.122508
sponge_ca3   5.809e-01  5.346e-02  10.865 < 2e-16 ***
sponge_ca4  -5.685e-01  6.353e-02  -8.948 < 2e-16 ***
sponge_ca6  -1.250e-01  5.858e-02  -2.135 0.032779 *
sponge_lha3  8.386e-01  4.112e-02  20.394 < 2e-16 ***
sponge_lha4 -1.428e+00  3.992e-02 -35.764 < 2e-16 ***
sponge_lha7  5.195e-03  3.280e-02   0.158 0.874175
total_unused_contacts_energy  4.972e-04  9.675e-05   5.139 2.77e-07 ***
hb_mwws_uc_energy  4.179e-05  1.681e-04   0.249 0.803721
hb_mwmm_uc_energy -3.245e-04  1.068e-04  -3.040 0.002370 **
hb_mws_uc_energy  -9.537e-05  2.250e-04  -0.424 0.671723
hb_mwm_uc_energy  4.345e-04  1.991e-04   2.183 0.029073 *
hb_ms_uc_energy  -2.121e-05  7.654e-04  -0.028 0.977897
hb_mm_uc_energy  4.414e-03  3.761e-04  11.734 < 2e-16 ***
hydro_uc_energy  4.574e-04  3.403e-04   1.344 0.178898
phi         -3.204e-04  3.577e-05  -8.957 < 2e-16 ***
psi         8.987e-06  1.878e-05   0.479 0.632269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.316 on 68248 degrees of freedom
Multiple R-squared: 0.4287, Adjusted R-squared: 0.4283
F-statistic: 1113 on 46 and 68248 DF, p-value: < 2.2e-16

Comando do software R.

```

Esse parâmetro também é mostrado no quadro 3, sendo muito próximo ao valor já mencionado do *coeficiente de determinação*. Ainda de acordo com Zar (1999), enquanto a medida de  $r^2$  sempre aumenta com o aumento do número de variáveis, o *coeficiente de determinação ajustado*  $r_a^2$  apenas aumenta quando a incorporação de uma nova variável ao modelo linear, traz alguma melhora para explicar a variabilidade da variável dependente. Para o caso do resíduo de aminoácido alanina, mostrado abaixo, ambos os valores indicam praticamente o mesmo valor, da ordem de 44%.

A tabela 8 mostra os valores de  $r^2$ ,  $r_a^2$  e de *valor p* do teste estatístico da distribuição *chi-quadrado* para todos os modelos lineares calculados para cada um dos 20 tipos de aminoácido.

**Tabela 8 – Medida de desempenho dos modelos lineares para previsão dos IFR, para cada um dos 20 tipos de aminoácido em termos de coeficiente de determinação, coeficiente de determinação ajustado e valor p.**

	$r^2$	$r_a^2$	<i>valor p</i>		$r^2$	$r_a^2$	<i>valor p</i>
<b>ala</b>	0,4488	0,4484	$< 2,2 \times 10^{-16}$	<b>leu</b>	0,459	0,4587	$< 2,2 \times 10^{-16}$
<b>arg</b>	0,348	0,3472	$< 2,2 \times 10^{-16}$	<b>lys</b>	0,2644	0,2636	$< 2,2 \times 10^{-16}$
<b>asn</b>	0,3579	0,3569	$< 2,2 \times 10^{-16}$	<b>met</b>	0,4573	0,4556	$< 2,2 \times 10^{-16}$
<b>asp</b>	0,555	0,5545	$< 2,2 \times 10^{-16}$	<b>phe</b>	0,4844	0,4836	$< 2,2 \times 10^{-16}$
<b>cys</b>	0,4492	0,4463	$< 2,2 \times 10^{-16}$	<b>pro</b>	0,4372	0,4367	$< 2,2 \times 10^{-16}$
<b>glu</b>	0,32	0,3194	$< 2,2 \times 10^{-16}$	<b>ser</b>	0,3913	0,3906	$< 2,2 \times 10^{-16}$
<b>gln</b>	0,3425	0,3414	$< 2,2 \times 10^{-16}$	<b>thr</b>	0,4376	0,4369	$< 2,2 \times 10^{-16}$
<b>gly</b>	0,1184	0,118	$< 2,2 \times 10^{-16}$	<b>trp</b>	0,6221	0,6202	$< 2,2 \times 10^{-16}$
<b>his</b>	0,3993	0,3977	$< 2,2 \times 10^{-16}$	<b>tyr</b>	0,4202	0,4191	$< 2,2 \times 10^{-16}$
<b>ile</b>	0,4622	0,4616	$< 2,2 \times 10^{-16}$	<b>val</b>	0,9606	0,9605	$< 2,2 \times 10^{-16}$

Surpreendentemente, o modelo linear para o resíduo de aminoácido valina mostrou-se com alto valor do *coeficiente de determinação ajustado* e destaca-se entre todos os outros modelos lineares. Para esse resíduo de aminoácido os dados dispostos em um modelo linear explicam a variabilidade observada para as duas classes de resíduos (IFR e superfície livre) ao nível de 96%. Apesar do alto sucesso do modelo por regressão linear em explicar a variabilidade presente nos dados do resíduo de aminoácido valina, resultados posteriores mostram que modelo não lineares como redes neurais artificiais e *ensemble* de árvores de decisão são mais bem sucedidos para a tarefa de classificação de IFR e FSR.

Em seguida os resíduos de aminoácidos triptofano e ácido aspártico possuem modelos lineares com  $r_a^2$  acima de 50%. Os modelos lineares para os demais aminoácidos apenas explicam

uma baixa percentagem da variabilidade das duas classes estudadas, em especial o modelo para o resíduo de aminoácido glicina possui *coeficiente de determinação ajustado* próximo de 12%.

Esse resultado indica que apesar do baixo *valor p* associado ao modelo de regressão linear, os *coeficientes de determinação* indicam que uma pequena percentagem é explicada, e, portanto, modelos não-lineares devem ser considerados para a predição de IFR.

Ressaltamos que o modelo de regressão linear não é um modelo classificador de IFR. Utilizamos esse modelo como uma etapa preliminar de análise para considerarmos a importância de cada variável utilizada, além de entendermos como a variabilidade presente em cada variável de entrada se associa com a variabilidade de variável binária de saída (classe IFR ou FSR).

### **3.6 Desenvolvimento de novos modelos classificadores**

O desempenho de cada tipo de classificador para cada um dos 20 tipos de aminoácido é avaliado utilizando os critérios de MCC e AUC. A figura 15 mostra o gráfico para cada modelo, ordenado em ordem decrescente de AUC.

Triptofano e ácido aspártico revezam entre os dois primeiros lugares em todos os métodos. Os piores resultados são observados para os resíduos de aminoácido glicina, lisina, ácido glutâmico, glutamina e arginina, nessa ordem. A figura 16 mostra a distribuição de cada tipo de aminoácido entre IFR e FSR.

Em relação ao classificador por SVM, os resíduos de aminoácido alanina, glutamina, isoleucina, prolina, serina, triptofano (este mostrado na figura 17), tirosina e valina, obtiveram uma curva simétrica ao normalmente observado, porém no triângulo inferior do gráfico ROC. Como descrito por Fawcett (2006), um classificador que gera pontos no gráfico ROC dessa forma indica que o modelo é capaz de utilizar os dados, porém está aplicando de forma errada. Dessa forma, se criarmos um classificador que é a negação do classificador “problemático”, obtemos uma curva ROC simétrica em relação a diagonal do gráfico ROC. Portanto, para esses classificadores re-calculamos o critério de AUC simplesmente subtraindo o valor dado pelo software R da unidade (que é o valor máximo da área do gráfico), obtendo assim a área “acima” da curva que, por sua vez, é o valor da AUC do classificador negação. Apesar do fato ser intrigante, não foi possível chegar a uma conclusão do porque esses tipos de aminoácidos obtiveram esse comportamento para o modelo classificador por SVM.



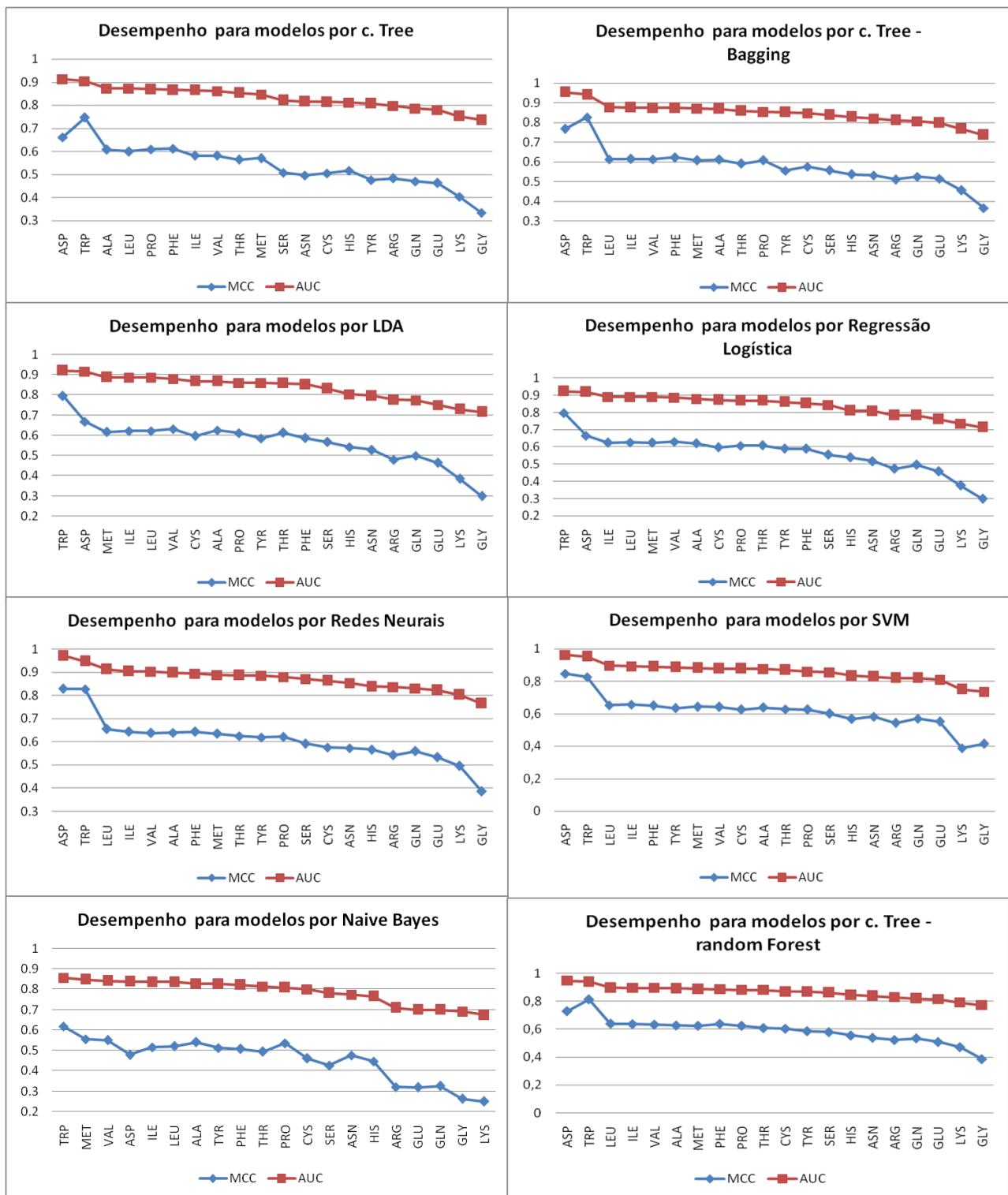
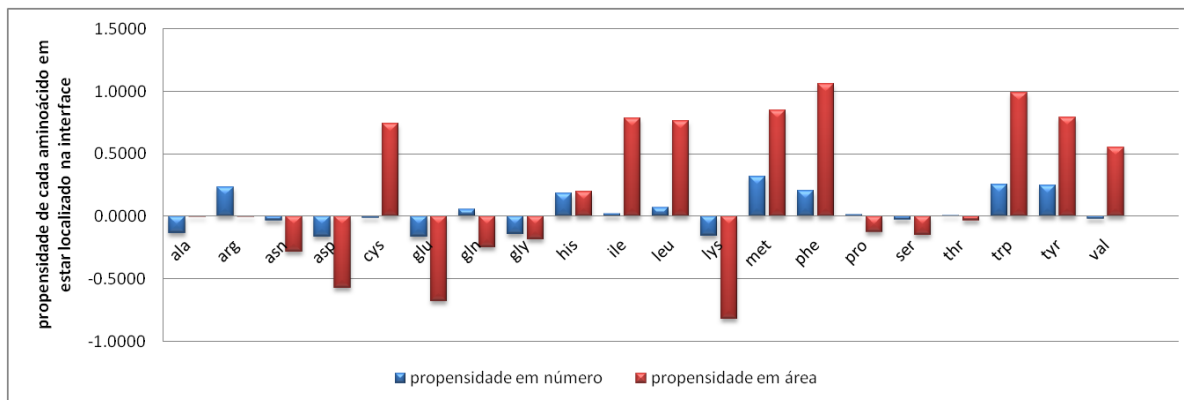


Figura 15 – Avaliação de desempenho de cada modelo classificador em relação aos 20 tipos de aminoácido.

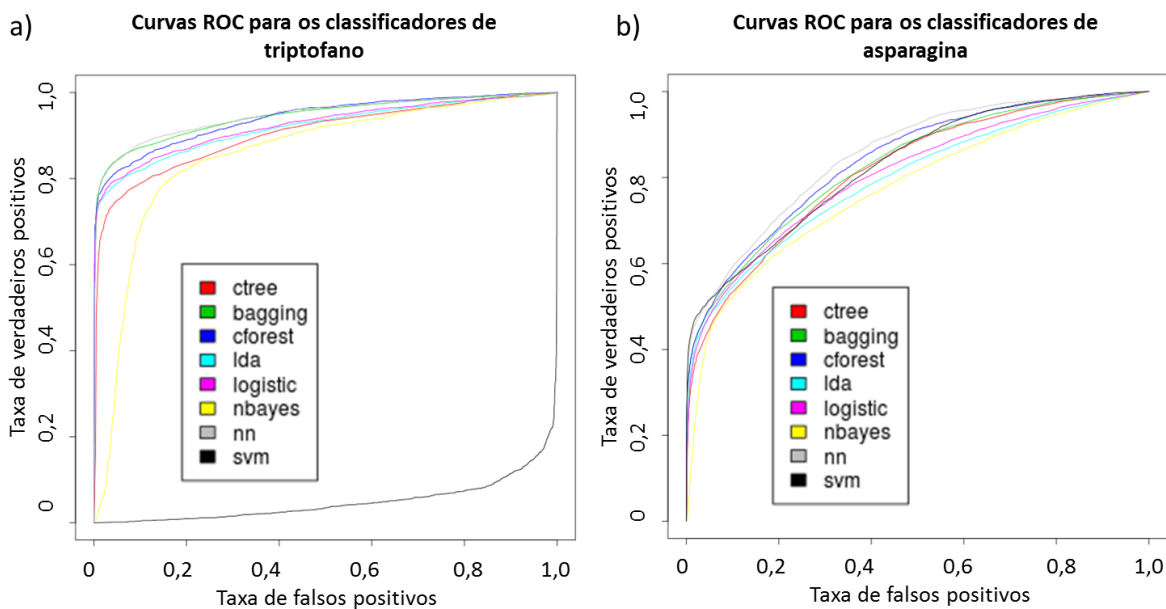
Os tipos de aminoácido estão ordenados de forma decrescente em relação ao critério de AUC.



**Figura 16 – Propensidade de cada tipo de aminoácido de estar localizado na superfície livre (FSR) ou na interface (IFR).**

Propensidade calculada como o logaritmo do número (azul) ou área (vermelho) total de resíduos de aminoácido na interface dividido pelo número ou área total de resíduos de aminoácido na superfície livre.

De forma comparativa entre os modelos, podemos sobrepor para o mesmo tipo de aminoácido as curvas ROC dos 8 modelos analisados. A figura 17 mostra as curvas ROC para os resíduos de aminoácido triptofano (figura 17-a) e asparagina (figura 17-b). Os demais resíduos de aminoácido serão apresentados como apêndice.



**Figura 17 – Comparação do desempenho dos modelos classificadores.**

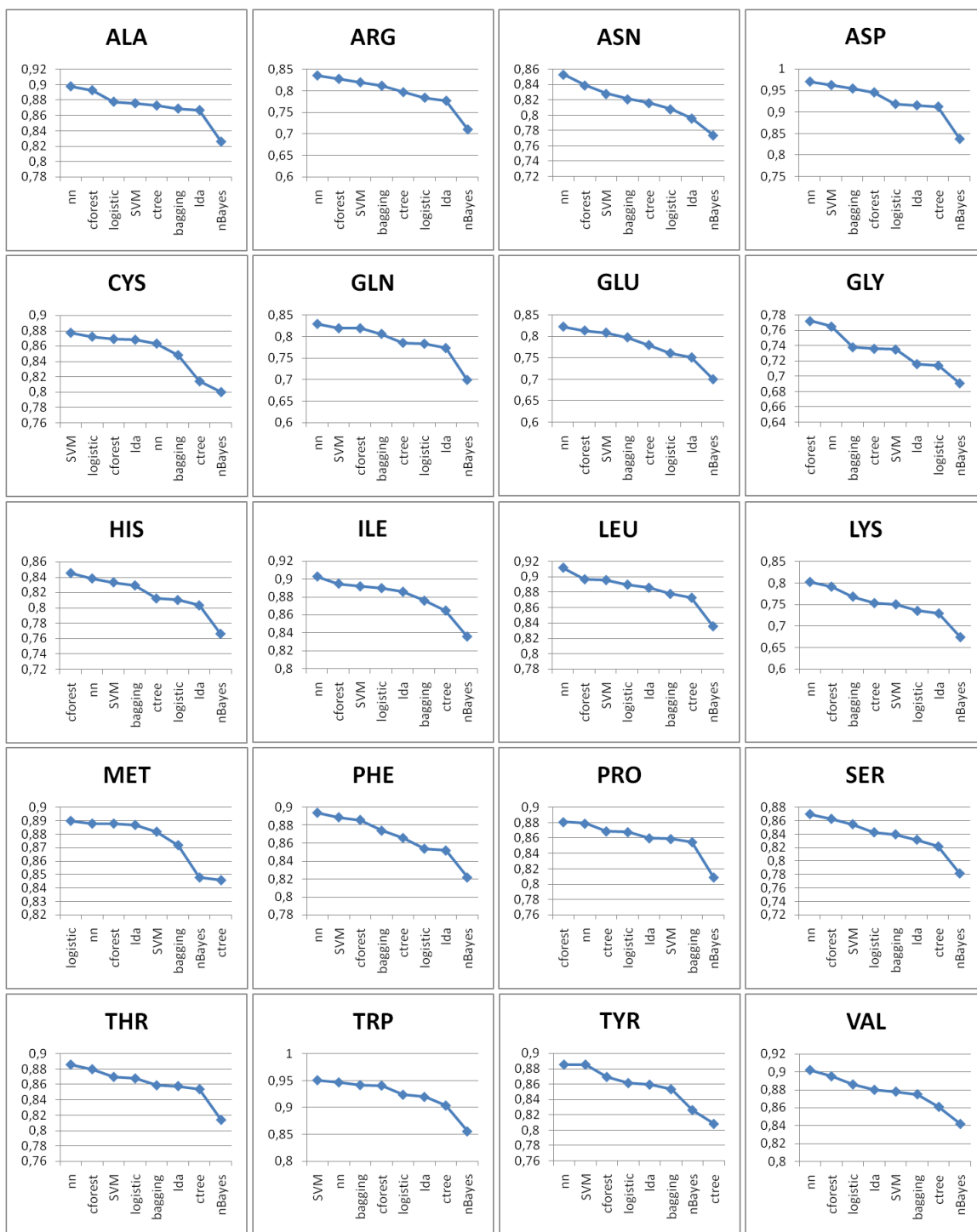
Árvore de decisão (vermelho), ensemble de árvores de decisão via bagging (verde) ou via random Forest (roxo), por LDA (magenta), por regressão logística (azul), por rede Bayesiana simples (amarelo), rede neural (cinza; abreviado como “nn”, do inglês neural network) e SVM (preto) de para os resíduos de aminoácido triptofano (a) e asparagina (b).

Para o resíduo de triptofano, o melhor modelo avaliado é o classificador via SVM, seguido da rede neural. Para o asparagina, a rede neural foi o método mais bem avaliado, sendo seguido pelo *ensemble* de árvores de decisão condicionais por *random Forest* e pela SVM.

Outra forma de visualizar o resultado, é utilizando os valores do critério de AUC para os oitos modelos separados por cada tipo de aminoácido, como mostra a figura 18, ordenando os classificadores pelo seu desempenho para cada um dos 20 tipos de aminoácido.

Para a maioria dos tipos de aminoácido (14 dos 20) o melhor modelo avaliado é a rede neural. Para o resíduo de aminoácido cisteína, no entanto, os modelos com melhor desempenho são SVM, regressão logística, *random Forest*, LDA e enfim a rede neural, nessa ordem. Para metionina, o modelo via regressão logística foi mais bem avaliado, sendo seguido pela rede neural e *random Forest*. Para glicina, o modelo via *random Forest* também ganhou destaque, obtendo melhor desempenho do que a rede neural que ficou em segundo lugar.

Entre os modelos com menor desempenho, encontramos os classificadores do tipo Naïve Bayes, exceto para os resíduos de aminoácido metionina (segundo pior classificador) e tirosina (segundo pior classificador), cujo modelo classificador por árvore de decisão condicional simples obteve o de desempenho menos satisfatório.



**Figura 18 – Comparação segundo o critério de AUC para os oito modelos classificadores (ordenados de forma decrescente, no eixo x) divididos para cada um dos 20 tipos de aminoácido. O modelo de redes neurais (abreviado como “nn”) foram os mais bem avaliados para 14 tipos de aminoácido, sendo seguido do modelo de ensemble de árvores de decisão condicional via random Forest (abreviado como “cforest”), que foi o melhor avaliado para 3 tipos de aminoácido e pelo modelo de SVM que foi o melhor avaliado em dois casos. O modelo de regressão logística ficou em primeiro lugar para o resíduo de aminoácido metionina.**

### 3.7 Modelo classificador linear por LDA

O modelo classificador por LDA foi desenvolvido e testado de forma extensiva em relação a validação cruzada, se há vantagem na separação por tipo de aminoácido do conjunto de dados e para comparar com outros modelos.

Para realizar a validação cruzada com 10 pastas, alocamos cada arquivo PDB em um de dez grupos distintos que foram sistematicamente divididos em 9 grupos de treino e 1 grupo de teste. Como o processo é repetido 10 vezes, todas as entradas são utilizadas como teste uma única vez.

A figura 19 mostra o desempenho de cada classificador específico de aminoácido relativo a indicação de classificação como IFR ou FSR ao qual o resíduo predito pertence durante o processo de validação cruzada, ordenado por melhor desempenho. Observamos que a variação do desempenho em relação a validação cruzada é pequena (dentro de 0,05) para a maioria dos tipos de aminoácido, sendo que a maior distribuição de valores de desempenho está relacionado com os resíduos de aminoácido isoleucine, cisteína e metionina. Quando medimos o desempenho pelo critério AUC, a variabilidade do desempenho é menor do que quando comparamos os 10 classificadores da validação cruzada pelo MCC.

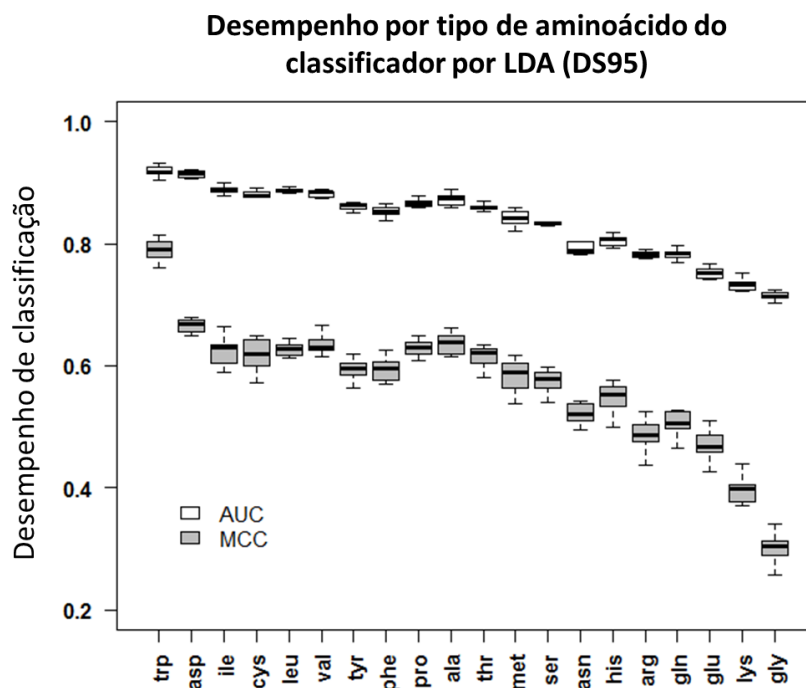


Figura 19 – Desempenho de cada classificador específico de aminoácido durante processo de validação cruzada, avaliados pelo critério AUC (branco) e MCC (cinza). Tipos de aminoácido ordenados por maior desempenho.

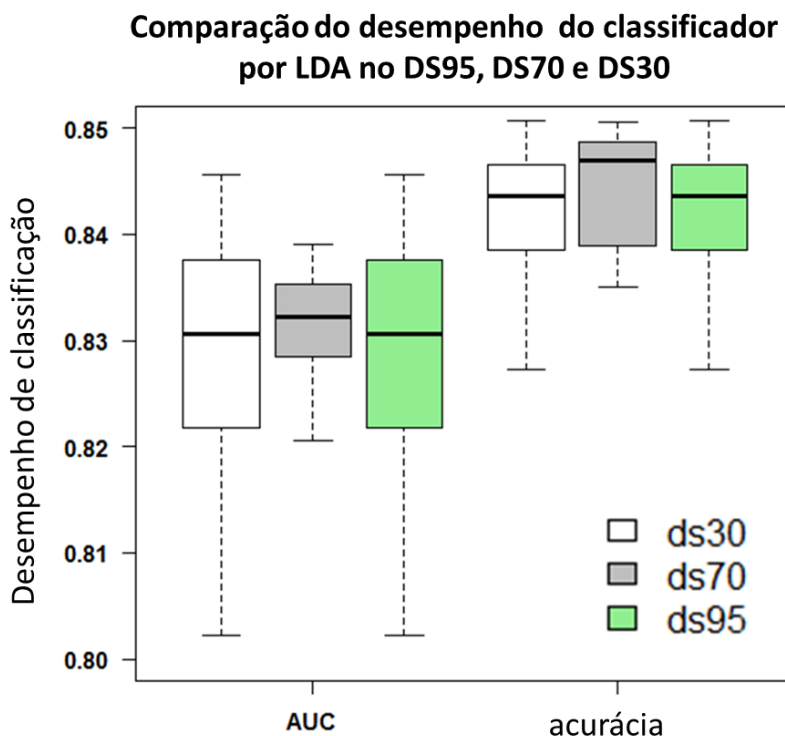
Diferentes tipos de aminoácido possuem estrutura e composição química específica que, em conjunto, definem o nano ambiente em que cada resíduo de aminoácido está inserido. Dessa forma, ao definirmos 20 classificadores específicos buscamos obter como o nano ambiente de cada tipo de aminoácido pode estar correlacionado com o nano ambiente da interface em complexos proteína-proteína. Isso explica, por exemplo o porquê o classificador do resíduo de aminoácido glicina configura como o último colocado na figura 19, uma vez que a falta de cadeia lateral impossibilita a definição de um nano ambiente específico para esse tipo de aminoácido cuja cadeia principal é a mesma para todos os tipos de aminoácido, com exceção da prolina.

Observamos também que oito entre os dez tipos de aminoácido com melhor poder de classificação são resíduos de aminoácido do tipo hidrofóbico, ou seja, com constante de hidrofobicidade de Radzicka positivas. Portanto, o nano ambiente gerado pelos descritores físico-químicos e estruturais dos resíduos de aminoácido hidrofóbicos estabelecem as características fundamentais no processo de formação dos complexos entre proteínas.

Apenas os classificadores dos dois resíduos de aminoácido hidrofílicos, ácido aspártico e tirosina, figuram entre os dez classificadores com melhor desempenho na figura 19. A figura 16 indica que o resíduo de aminoácido de ácido aspártico apresenta propensidade de localização na superfície livre. A combinação de seus descritores fornece alta diferenciação quando este resíduo de aminoácido está na interface do que quando está na superfície livre, mais ainda do que os resíduos de aminoácido lisina e ácido glutâmico que possuem maior propensidade de estar na superfície livre do que o resíduo de aminoácido ácido aspártico (figura 16). Apesar de ser hidrofílico, a constante de hidrofobicidade do resíduo de aminoácido tirosina é próxima a zero indicando que o grau de hidrofilicidade desse resíduo de aminoácido é próxima dos resíduos de aminoácido hidrofóbicos, tendo alta propensidade de ser encontrado na interface de complexos proteína-proteína.

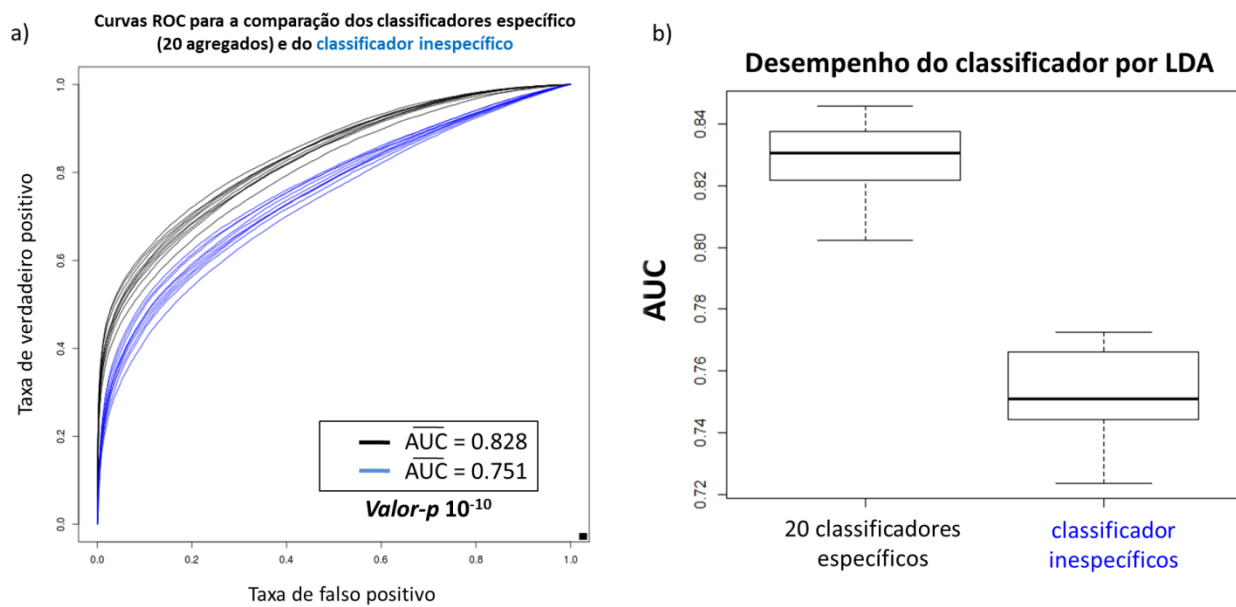
Após a síntese 20 classificadores específicos para aminoácidos, a obtenção do classificador de interface final é obtida agregando os vinte classificadores em apenas um classificador final com 20 componentes. A figura 20 compara a variação de desempenho quando utilizamos os conjuntos de dados DS95, DS70 e DS30 por *boxplots* calculados por validação cruzada com 10 pastas. Nessa figura, os 20 classificadores específicos de aminoácidos são agrupados resultando no classificador final de IFR. Observamos que pouca variabilidade é observada quando restringimos a seleção do banco de dados entre os valores 95%, 70% e 30% de similaridade sequencial. Por isso, os

resultados apresentados para o restante do classificador por LDA e para o classificador por *ensemble* de redes neurais, apresentado na próxima seção, utilizam o conjunto DS30.



**Figura 20 – Comparação do o desempenho do classificador por LDA em relação a escolha do limiar de similaridade sequencial na criação do banco de dados. Boxplots gerados por validação cruzada.**

Para comparar se a divisão em 20 classificadores específicos para cada tipo de aminoácido pode beneficiar a classificação de IFR, construímos, por validação cruzada, 10 modelos classificadores inespecíficos de aminoácidos. A figura 21 compara os 20 modelos classificadores específicos de aminoácidos combinados (agregados) com o modelo classificador inespecífico (em azul). Nota-se que o melhor desempenho do classificador inespecífico (maior AUC) está aquém do pior desempenho dos classificadores específicos de aminoácidos agregados em um classificador final. O mesmo é observado na figura 21-b através da representação por *boxplot*. O conjunto de dados DS30 foi utilizado para a geração dos modelos classificadores inespecíficos de aminoácido e comparado com a mesma versão dos modelos específicos.



**Figura 21 – Comparação entre os 20 classificadores específicos de aminoácidos agregados (em preto) com o classificador inespecífico de aminoácidos (em azul) por curvas ROC (a) e pelo critério de AUC utilizando boxplots (b).**

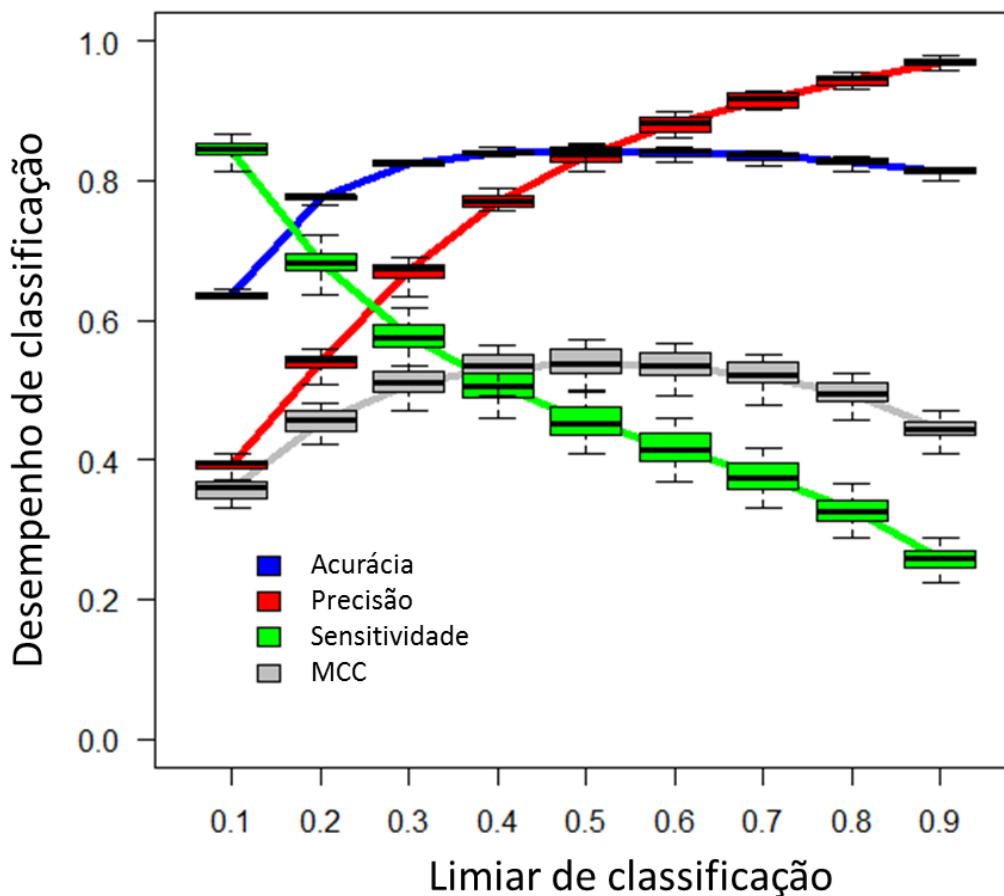
Cada curva representa uma das rodadas do processo de validação cruzada com 10 pastas.

De acordo com a figura 21, a divisão por tipo de aminoácido e criação de classificadores específicos trazem vantagens para a classificação de IFR, confirmando nossa hipótese. O teste estatístico de Welch (descrito acima para outro propósito) conclui que a diferença no desempenho dos classificadores é estatisticamente relevante, com um *valor p* da ordem de  $10^{-10}$ .

Como o resultado do modelo classificador por LDA é uma probabilidade *a posteriori*, podemos varrer o limiar de classificação e observar o que acontece com os valores de cada unidade de medida de desempenho. A figura 22 mostra as curvas para as taxas de acerto (acurácia), precisão, sensibilidade e MCC, conforme o limiar é varrido entre 0,1 e 0,9 para o conjunto teste de cada etapa da validação cruzada. Os *boxplots* mostrados são referentes aos valores da validação cruzada com 10 pastas.



## Dependência do desempenho de classificação em relação ao limiar de classificação (DS30)

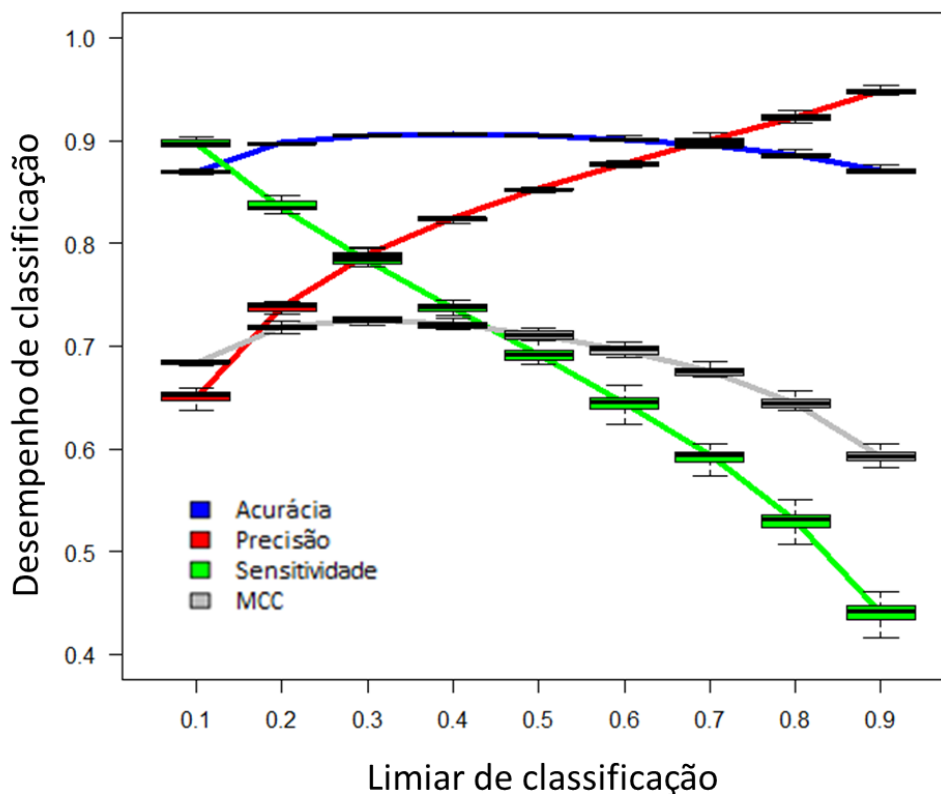


**Figura 22 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido.**

*Boxplots para cada limiar de classificação são construídos com resultados da validação cruzada.*

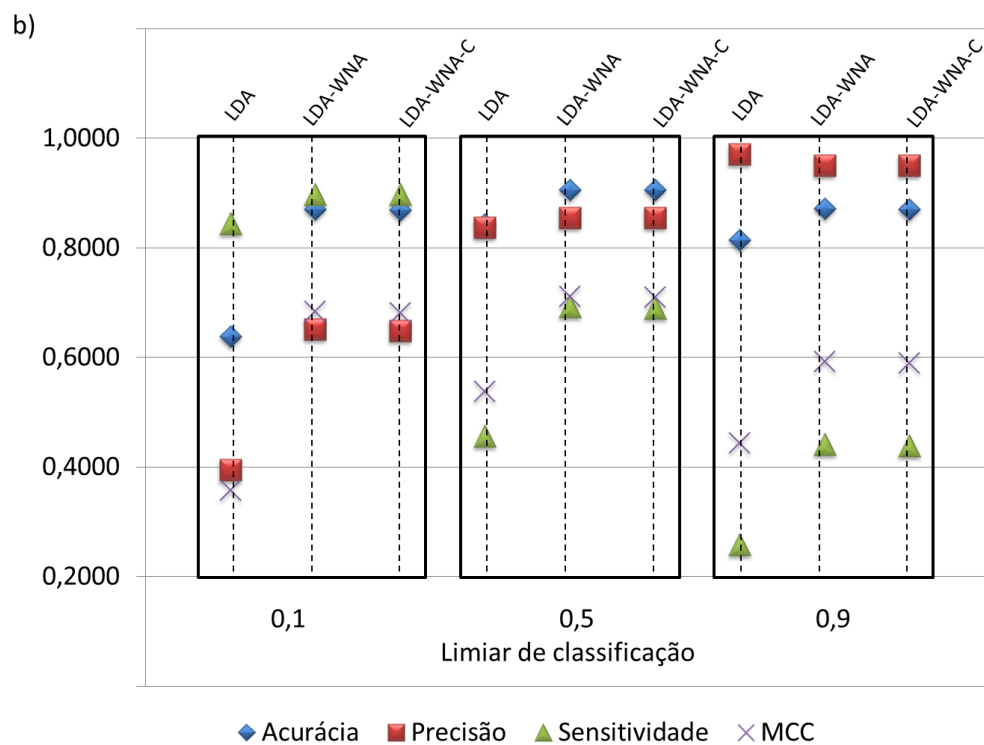
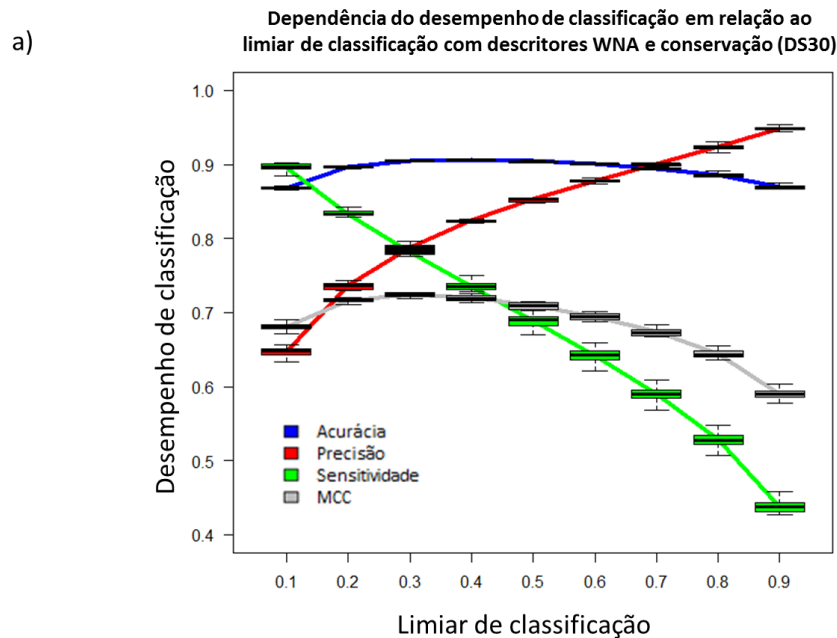
A incorporação de descritores ponderados pela vizinhança (WNA) resultou em um aumento em todas as medidas de desempenho utilizadas (figura 23). Para o limiar de classificação 0,5 observamos um aumento em 10 pontos percentuais para a taxa de acerto (acurácia), 5 pontos percentuais para a precisão, 25 pontos percentuais em sensibilidade e 0,15 em MCC.

**Dependência do desempenho de classificação em relação ao limiar de classificação com descritores WNA (DS30)**



**Figura 23 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido após a incorporação dos descritores ponderados pela vizinhança (WNA).**  
*Boxplots para cada limiar de classificação são construídos com resultados da validação cruzada.*

Ao acrescentar os descritores de conservação de aminoácidos no conjunto de dados de treinamento para o classificador por LDA, não observamos nenhuma mudança no poder preditivo do modelo classificador final, como ilustrado na figura 24-a. Todas as taxas de desempenho permanecem dentro dos limites do *boxplot* para cada limiar de classificação. Com isso, os descritores físico-químicos e estruturais acrescidos de seus respectivos valores ponderados pela vizinhança fornecem toda a informação necessária para a classificação de resíduos de aminoácido como IFR ou FSR de forma que a informação sobre o quão conservado um resíduo de aminoácido é, não altera o potencial do classificador. A figura 24-b compara os três modelos classificadores por LDA com descritores simples (LDA), acrescidos dos descritores de vizinhança (LDA-WNA) e com a inclusão de descritores de conservação de aminoácidos (LDA-WNA-C) para três valores do limiar de classificação.

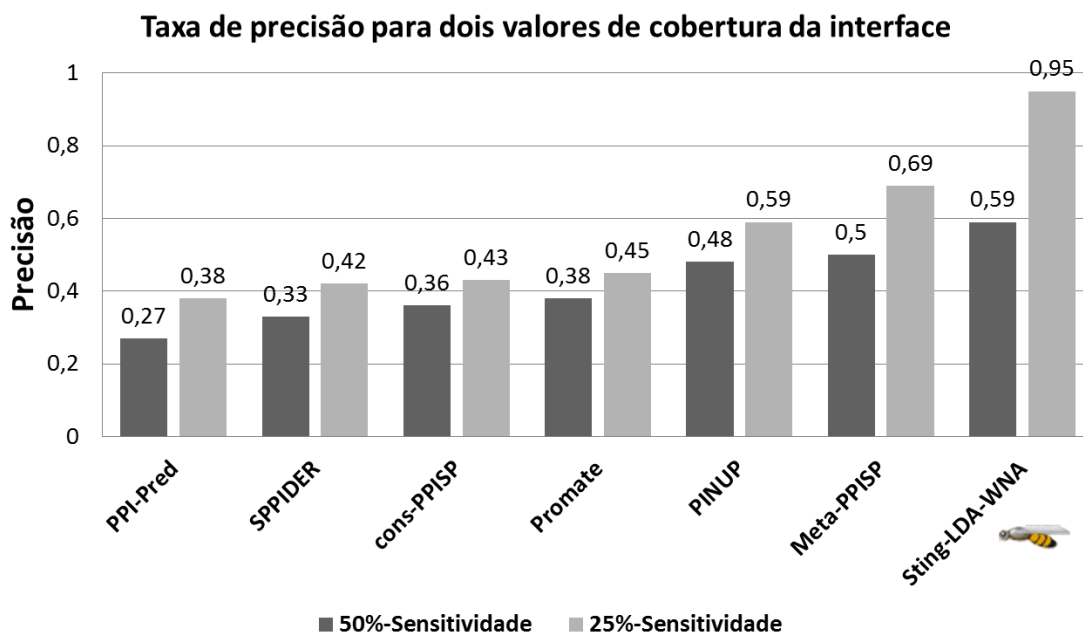


**Figura 24 – Dependência do desempenho de classificação em relação ao limiar de classificação escolhido após a incorporação dos descritores ponderados pela vizinhança (WNA) e conservação de aminoácidos (a) e comparação entre os três modelos classificadores por LDA (b).**

Boxplots para cada limiar de classificação são construídos com resultados da validação cruzada. Utilizando três valores para o limiar de classificação é mostrado que o aumento de desempenho com a adição dos descritores ponderados pela vizinhança (LDA-WNA) não é superado pela adição dos descritores sobre conservação de aminoácidos.

Para comparar com o classificador por LDA com outros métodos, utilizamos o banco de dados 35Enz. Como avaliado no artigo por Zhou e Qin (2007-a), os critérios de precisão e sensibilidade foram utilizados como métricas comparativas. Nessa comparação utilizamos todos os 10 grupos de dados em um novo treinamento do classificador por LDA, batizado de Sting-LDA-WNA, uma vez que o classificador com descritores ponderados pela vizinhança mostrou-se com melhor desempenho.

A figura 25 mostra a comparação do classificador Sting-LDA-WNA com os outros métodos avaliados por Zhou e Qin (2007-a). Os valores de desempenho para os demais métodos foram diretamente retirados do artigo por Zhou e Qin (2007-a), ou seja, não implementamos cada um dos métodos. Seguindo o método de comparação, varremos as taxas de desempenho em relação ao limiar de classificação em que todos os métodos comparados tenham um total de 50% em sensibilidade, ou seja, metade do total de resíduos de aminoácido que compõe a interface sendo preditos corretamente. Na figura 25, o valor da taxa de precisão para qual o valor de sensibilidade alcança 50 e 25% é mostrado. Com ambas as condições, Sting-LDA-WNA possui precisão acima de todos os métodos PPI-Pred (Bradford e Westhead, 2005), SPPIDER (Porollo e Meller, 2007), cons-PPISP (Chen e Zhou, 2005), Promate (Neuvirth *et al.*, 2004), PINUP (Liang *et al.*, 2006) e Meta-PPISP (Qin e Zhou, 2007-b). Para a condição de 25% em sensibilidade, Sting-LDA-WNA obteve o maior ganho em relação a condição de 50% de sensibilidade.



**Figura 25 – Comparação entre o classificador por LDA desenvolvido (Sting-LDA) e outros métodos com base no conjunto teste 35Enz.**

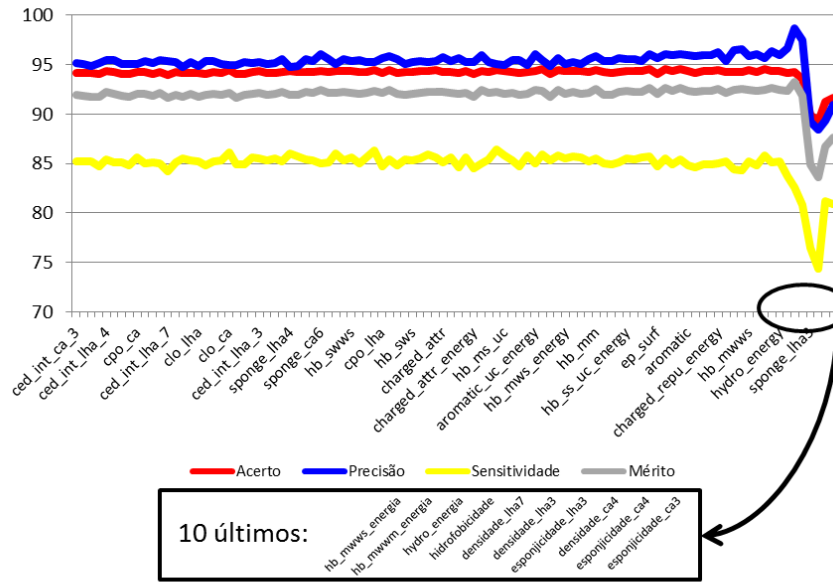
### **3.8 Classificadores por redes neurais artificiais**

Diferentes tipos de aminoácido utilizam combinações ótimas compostas por diferentes descritores. A busca exaustiva e sistemática realizada no processo de seleção de variáveis por redes neurais e a distinção entre os 20 tipos de aminoácido naturais representam um conjunto de resultados ainda não explorado na literatura.

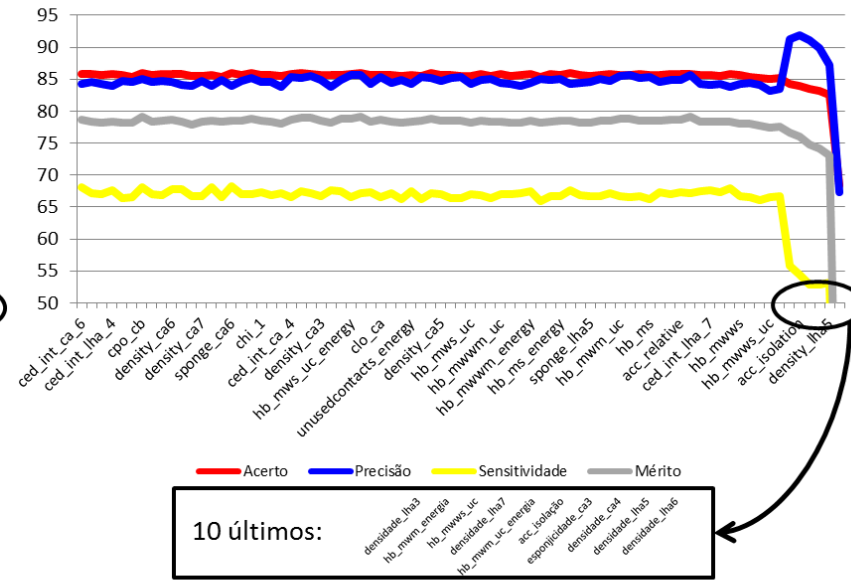
O não uso de descritores referentes à conservação de aminoácidos é outra diferença importante entre a metodologia proposta e as demais reportadas. Apesar de pouca exploração na literatura sobre o efeito de se considerar parâmetros de conservação, Caffrey *et al.* (2004) relatam que o uso desse parâmetro não tem poder suficiente para prever regiões de interação. O uso de conservação quando combinados com outros descritores melhora a capacidade preditiva de sistemas classificadores, como ilustrado por Liang *et al.* (2006). Como explicado anteriormente, nossa abordagem preza por parâmetros que mapeiam o nano-ambiente físico-químico e estrutural do resíduo de aminoácido a ser classificado, e o uso do parâmetro conservação inviabiliza essa interpretação. Mas apenas para fins de comparação de desempenho de classificação, incorporamos o descritor de conservação em uma nova etapa de treinamento do *ensemble* de redes neurais, que será mostrado posteriormente nesta seção.

O processo de seleção de variáveis através de um mecanismo de poda (no início todas as variáveis são utilizadas) foi realizado para todos os tipos de aminoácido separados. Primeiramente utilizamos a abordagem de *holdout* para avaliar o classificador por rede neural. A figura 26 mostra o processo de retirada de variáveis para os resíduos de aminoácido de melhor (figura 26-a – triptofano) e pior (figura 26-b – metionina) desempenho, para um resíduo de aminoácido intermediário (figura 26-c – arginina) e também para o classificador inespecífico quanto ao tipo de aminoácido (figura 26-d). Destacamos que devido ao elevado número de descritores utilizados no processo de seleção de variáveis não é possível listar todos no eixo horizontal da figura 26. A listagem completa da ordem de remoção é apresentada como apêndice. Os resultados para os demais tipos de aminoácido são apresentados no apêndice.

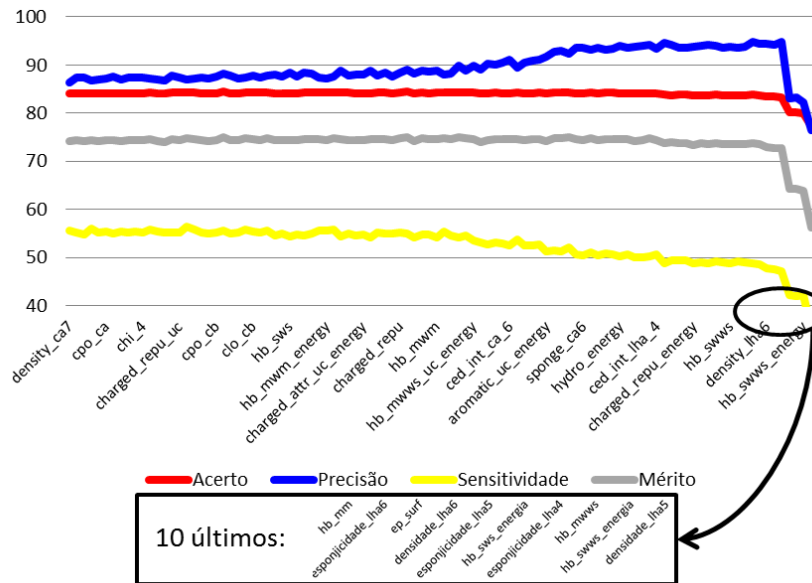
**a) Seleção de variável para o estabelecimento do classificador de triptofano**



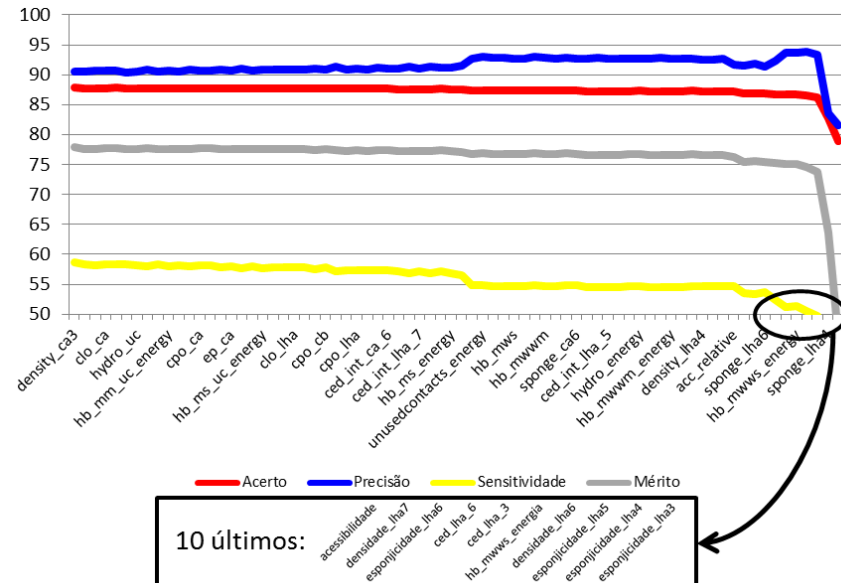
**b) Seleção de variável para o estabelecimento do classificador de metionina**



**c) Seleção de variável para o estabelecimento do classificador de arginina**



**d) Seleção de variável classificador inespecífico quanto ao tipo de aminoácido**



**Figura 26 – Processo de seleção de variável utilizando a rede neural para síntese de classificadores do tipo envoltório (wrapper).**

*No início todos os descritores são combinados e seguindo uma análise de sensibilidade os descritores são removidos um a um, de acordo com a variação da taxa de mérito. Para o classificador específico de triptofano (a) a taxa de acerto manteve-se acima de 93% até restarem cinco descritores usados para predição de IFR. Da mesma forma a precisão oscilou com média aproximada em 95% e a sensibilidade com média próxima a 85%. O classificador específico de metionina (b) manteve o mesmo comportamento, porém com taxa de acerto, precisão e sensibilidade em torno de 85%, 84% e 67%, respectivamente. O aumento em precisão na etapa final desse processo é comum entre os dois gráficos, porém para o aminoácido metionina esse aumento é mantido por mais quatro retirada de variáveis, enquanto que para triptofano esse aumento dura apenas duas variáveis. Para o classificador de arginina (c), a etapa final do processo de seleção de variáveis aumentou a precisão de classificação. Durante a síntese do classificador inespecífico para os tipos de aminoácidos (d) a taxa de acerto, precisão e sensibilidade mantêm-se em 87%, 91-93% e 57-54%, respectivamente. Na metade do processo, observa-se um aumento de dois pontos percentuais para precisão e um decréscimo de três pontos percentuais para sensibilidade. As taxas de acerto e mérito mantêm-se aproximadamente constantes durante todo o processo, sofrendo apenas uma queda no final quando há apenas três variáveis. No eixo horizontal não é possível listar todos os descritores devido a sobreposição do elevado número de variáveis utilizadas no processo de seleção.*



No início, todas as variáveis combinadas geram valores de desempenho que seguem relativamente constantes até que grande parte das variáveis tenha sido retirada. O classificador específico para o resíduo de aminoácido triptofano mostrou-se com melhor desempenho, atingindo 93% de taxa de acerto (acurácia), 95% de precisão e 85% de sensibilidade. Uma taxa de precisão alta indica um baixo número de *falsos positivos*, ou seja, poucos resíduos de aminoácido presentes na superfície livre são classificados como de interface. A taxa de sensibilidade ficou com o menor patamar entre as taxas utilizadas, e sua interpretação está relacionada com o número de *falsos negativos*. Portanto, quanto menor seu valor maior o número de resíduos de aminoácido de interfaces que são classificados como de superfície livre. Para os classificadores de triptofano, essa taxa mantém-se relativamente alta, indicando que poucos resíduos de triptofano são classificados de superfície livre quando são de fato interfaces. Avaliando a figura 26-b para o resíduo de aminoácido metionina, as taxas de acerto, precisão e sensibilidade são 85%, 84% e 67%, respectivamente. Ambas as taxas de acerto e precisão mantêm-se relativamente alta em relação ao classificador de triptofano, indicando alto número de resíduos de aminoácido classificados

corretamente e com baixo número de *falsos positivos*. No entanto a taxa de sensibilidade está situada bem abaixo indicando um maior índice de *falsos negativos* para metionina quando comparado com os classificadores de triptofano. O classificador de arginina mostrou desempenho intermediário entre o classificador de triptofano e de metionina, com a peculiaridade de aumentar a taxa de precisão nas etapas finais do processo de seleção de variáveis. Taxas intermediárias são encontradas para o classificador inespecífico quanto ao tipo de aminoácido (figura 26-d), atingindo 87% de taxa de acerto.

As taxas de precisão e sensibilidade sofrem uma variação na metade do processo de seleção de variáveis, indicando o compromisso entre elas: valores baixos de *falsos positivos* são acompanhados de valores altos de *falsos negativos*. Para esse último classificador há um salto de 91% para 93% na taxa de precisão ao mesmo tempo em que um decréscimo de 57% para 54% é observado para sensibilidade. No entanto, esses valores referidos acima não podem ser aceitos como os verdadeiros valores de desempenho dos classificadores gerados. Eles são referentes à predição no conjunto de dados de validação logo após a utilização dos dados de treinamento.

No apêndice, é apresentada a listagem com a ordem de remoção das variáveis para cada resíduo de aminoácido e para o classificador inespecífico. A tabela 9 mostra apenas os dez últimos descritores para cada um dos classificadores específicos, uma vez que esse número de descritores consegue manter o desempenho que diminui entre as últimas 6 remoções de variáveis, ou seja, os descritores mais relevantes para distinguir entre os resíduos de aminoácido formadores de interface e os resíduos de aminoácido da superfície livre. Os nomes das variáveis estão abreviados de acordo com a descrição presente na tabela 5.

Os gráficos dos resíduos de aminoácidos triptofano, metionina e do classificador inespecífico, todos ilustrados na figura 26, além dos gráficos do apêndice 5 para os resíduos de aminoácido glicina, isoleucina, leucina, fenilalanina, tirosina e valina, indicam uma alta da taxa de precisão com uma queda da taxa de sensibilidade nas etapas finais do processo de seleção de variáveis. Como ambas as medidas de desempenho dependem do número de IFR corretamente preditos, podemos inferir que nessa etapa há um aumento do número de IFR erroneamente preditos (*falsos negativos*) com uma redução do número de FSR erroneamente preditos (*falsos positivos*).



**Tabela 9 – lista dos dez últimos descritores removidos no processo de seleção de variáveis no modelo envoltório usando a taxa de mérito como critério de poda.**

*Quanto mais abaixo o descritor se encontra maior sua relevância*

Ala	Arg	Asn	Asp	Cys
hidro_energia	hb_mm	hb_mwws_energia	densidade_lha4	ss_bond_energia
ced_int_lha_4	esponjicidade_lha6	ced_int_lha_3	hb_ms	hb_mwws_uc
esponjicidade_lha7	ep_surf	densidade_ca4	hb_ss_energia	densidade_lha3
densidade_lha7	densidade_lha6	esponjicidade_ca3	hb_mwwm	hb_mwm_energia
esponjicidade_ca3	esponjicidade_lha5	densidade_lha7	densidade_lha3	densidade_lha5
esponjicidade_ca4	hb_sws_energia	densidade_lha4	esponjicidade_ca3	ced_int_lha_5
ced_int_lha_5	esponjicidade_lha4	carregado_repu	esponjicidade_ca4	esponjicidade_ca3
hydro	hb_mwws	esponjicidade_lha6	densidade_lha7	esponjicidade_ca4
esponjicidade_lha4	hb_swws_energia	densidade_lha6	esponjicidade_lha6	hb_mws
densidade_lha5	densidade_lha5	densidade_lha5	densidade_lha6	densidade_lha6
Glu	Gln	Gly	His	Ile
esponjicidade_ca4	ced_int_lha_6	hb_mws_uc_energia	hb_mwws_energia	densidade_ca4
carregado_repu_energia	ep_surf	contact_energia_score	acc_isolação	esponjicidade_ca4
contact_energia_score	carregado_repu_energia	hb_mws_energia	densidade_ca4	esponjicidade_ca5
densidade_lha7	hb_mm_energia	esponjicidade_ca3	densidade_lha7	hb_mwwm
hydro_energia	densidade_lha5	esponjicidade_ca4	ep_surf	hb_mwwm_energia
esponjicidade_lha3	hb_mwwm_energia	densidade_ca4	densidade_lha4	hb_mm_energia
densidade_lha6	esponjicidade_lha7	hydro	esponjicidade_ca4	ced_int_lha_3
carregado_attr	densidade_lha7	hb_mwws_energia	esponjicidade_ca3	densidade_lha5
esponjicidade_lha5	esponjicidade_lha4	hb_mwws	densidade_lha5	esponjicidade_lha3
densidade_lha5	esponjicidade_lha3	hydro_energia	esponjicidade_lha6	esponjicidade_lha4
Leu	Lys	Met	Phe	Pro
hb_mm	hb_swws	densidade_lha3	hb_mwws_energia	hb_mwws_energia
hb_mws	hb_ms_energia	hb_mwm_energia	densidade_lha7	esponjicidade_ca7
hb_mws_uc_energia	esponjicidade_lha6	hb_mwws_uc	hb_mwwm	hb_mm_energia
hb_mwwm	esponjicidade_lha3	densidade_lha7	esponjicidade_ca3	densidade_lha4
esponjicidade_ca3	densidade_lha4	hb_mwm_uc_energia	densidade_ca4	densidade_ca4
esponjicidade_ca4	densidade_lha5	acc_isolação	esponjicidade_ca4	hydro
ced_int_ca_3	carregado_repu_energia	esponjicidade_ca3	hb_mwm_uc_energia	esponjicidade_lha7
densidade_lha5	hb_mwwm_energia	densidade_ca4	hidrofobicidade	densidade_lha7
ced_int_lha_5	densidade_lha6	densidade_lha5	esponjicidade_lha3	densidade_lha3
densidade_lha6	esponjicidade_lha4	densidade_lha6	densidade_lha4	densidade_lha5
Ser	Thr	Trp	Tyr	Val
esponjicidade_lha6	contact_energia_score	hb_mwws_energia	esponjicidade_lha5	hb_mws_energia
contact_energia_score	esponjicidade_lha6	hb_mwwm_energia	densidade_lha5	esponjicidade_ca3
hb_mwm_energia	densidade_lha6	hydro_energia	hb_mm	esponjicidade_ca5
hb_mwws	densidade_lha4	hidrofobicidade	hb_mwwm_energia	densidade_lha7
esponjicidade_ca3	ced_int_lha_5	densidade_lha7	esponjicidade_lha3	hidrofobicidade
densidade_ca4	densidade_ca4	densidade_lha3	hb_mm_energia	contact_energia_score
densidade_lha6	esponjicidade_ca3	esponjicidade_lha3	hb_ss_energia	hydro_energia
ced_int_lha_5	hb_swws_energia	densidade_ca4	densidade_lha4	ced_int_lha_4
esponjicidade_lha5	densidade_lha7	esponjicidade_ca4	densidade_lha7	densidade_lha4
densidade_lha4	acc_relativo	esponjicidade_ca3	acc_isolação	esponjicidade_lha5

*Densidade e esponjicidade* configuram-se entre os descritores que estão sempre presentes na listagem da tabela 9 dos classificadores específicos para cada tipo de aminoácido. Como

mostrado no teste estatístico da diferença entre duas amostras e nos gráficos do tipo *boxplot* (figura 13 e apêndice 6.3), regiões de interface possuem maior densidade quando comparadas a superfície livre. Regiões com maior densidade podem indicar regiões mais estáveis mecanicamente. Essa diferença mostrada por nossos resultados levanta a hipótese de que IFRs são menos susceptíveis, em média, a variações em suas posições. Tal hipótese indica que a distância relativa para resíduos de aminoácido chaves para o estabelecimento do complexo é importante durante o processo de formação dos complexos proteína-proteína, limitando a região de interface favorável para essa associação entre as subunidades proteicas. Estudos adicionais de dinâmica molecular para proteínas específicas poderiam ajudar a testar a hipótese levantada.

Outros descritores são mais variáveis para diferentes tipos de aminoácido. Hidrofobicidade sempre foi considerado um dos descritores mais importantes para distinguir IFRs, sendo que resíduos de aminoácido mais hidrofóbicos (menor número de átomos de alta eletronegatividade ou carregados na cadeia lateral) tendem a ser mais numerosos na interface. No entanto, esse descritor encontra-se presente na lista dos 10 melhores descritores de apenas três resíduos de aminoácido (fenilalanina, triptofano e valina). Nossos resultados não invalidam o conceito geral de que IFRs são mais hidrofóbicos do que resíduos de aminoácido na superfície livre, uma vez que na tabela 9 não há comparações entre tipos diferentes de aminoácidos. Nesse caso a pergunta a ser feita não é “*resíduos de aminoácido da interface são mais hidrofóbicos?*”, mas sim “*dado um determinado tipo de aminoácido (alanina, por exemplo), há diferença na hidrofobicidade ponderada pela área relativa acessível ao solvente desse tipo específico de aminoácido quando ele está embutido em um nano-ambiente de interface em relação quando ele se situa na superfície livre?*”. Para o classificador inespecífico, as comparações continuam sendo feitas entre as duas classes estudadas, mas como há todos os tipos de aminoácidos (com exceção de glicinas) a primeira pergunta passa a ser válida. A listagem das variáveis mais importantes para o classificador inespecífico indica novamente que esse descritor é de baixo poder discriminatório para resíduos de aminoácido isolados, sendo a 35ª variável a ser retirada (estando na metade do processo de seleção de variáveis). O descritor de *energia de contatos hidrofóbicos* mostrou-se mais importante na tarefa de distinguir interfaces, sendo o descritor número 55 a ser eliminado. Esse mesmo descritor é encontrado na lista dos dez melhores parâmetros (tabela 9) para os classificadores de alanina, ácido glutâmico e glicina. Desta forma, concluímos que o conceito de que uma região de interface é mais hidrofóbica só deve ser levado em conta como uma característica da interface como um todo, ou seja, vários resíduos de aminoácido formando uma interface, e não como uma

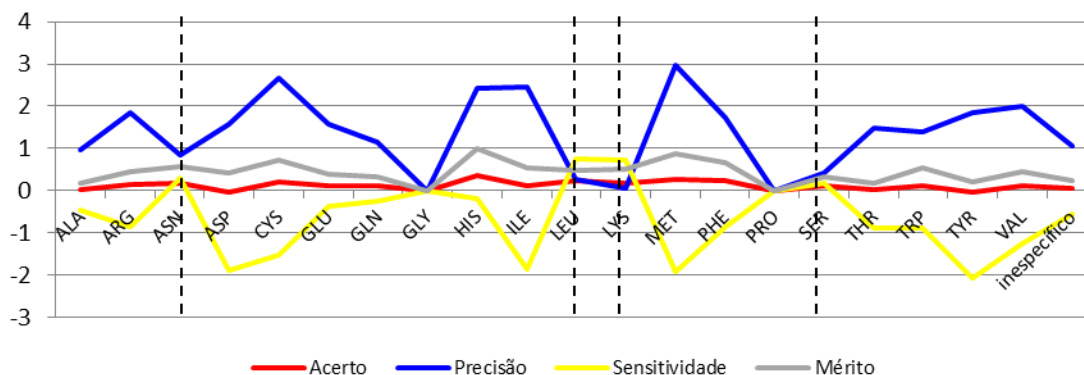
propriedade atribuída ao resíduo de aminoácido específico, como é o caso da análise desenvolvida nesta seção.

Além de *hidrofobicidade*, o estabelecimento de ligações de hidrogênio com resíduos de aminoácido da mesma cadeia proteica mostra-se importante para quase todos os classificadores específicos, com exceção de alanina e ácido glutâmico. De acordo com os resultados dos testes estatísticos univariados mostrados previamente, os IFRs possuem em média um número de ligações de hidrogênio estabelecidas internamente superior aos resíduos de aminoácido da superfície livre. Esse resultado levanta outra hipótese que pode ser testada por experimentos posteriores. Da mesma forma que o descritor de densidade, estabelecendo mais contatos do tipo ligação de hidrogênio pode fazer com que os resíduos de aminoácido da interface fiquem mais estáveis propiciando a formação do complexo das unidades monoméricas, sendo mais um dos fatores que definem resíduos de aminoácido formadores de interface.

Para alguns resíduos de aminoácido polares (asparagina, ácido glutâmico, glutamina e lisina) descritores de *energia de contatos carregados* foram considerados importantes na lista apresentada na tabela 9. Arginina e histidina (polares) não apresentaram tal descritor entre os dez melhores, mas sim o parâmetro de *potencial eletrostático na superfície*. O classificador de glutamina também apresentou esse parâmetro.

O processo de seleção de variáveis do tipo *envoltório* baseado em poda gerou centenas (até milhares em alguns casos) de classificadores. A ideia de combinação dos classificadores em *ensembles* traz benefícios para o sistema classificador, como aumento de desempenho e redução do sobre ajustes aos dados do conjunto de treinamento. Partindo de uma análise construtiva (descrita acima) foram construídos 21 *ensembles*, dos quais 20 são específicos para tipos de aminoácido e um classificador inespecífico. A figura 27 mostra a diferença entre o desempenho do *ensemble* de classificadores e do melhor classificador simples (sem *ensemble*) para as quatro taxas de medida. Apenas os classificadores dos resíduos de aminoácido glicina e prolina não obtiveram ganho algum com a incorporação de um novo classificador no *ensemble*, e por esse motivo apenas o melhor classificador simples foi mantido. Os *ensembles* com maior número de componentes foram criados para os resíduos de aminoácido asparagina, histidina e para o classificador inespecífico quanto ao tipo de aminoácido, atingido oito, sete e sete componentes, respectivamente. Os melhores aumentos de desempenho (linha tracejada vertical preta) foram observados para os classificadores de asparagina, leucina, lisina e serina, pois não houve nenhum decréscimo em nenhuma taxa. Em todos os casos a taxa de acerto variou pouco, sendo que para

ácido aspártico e tirosina houve um pequeno decréscimo. A taxa de precisão foi observada com o maior benefício no processo de estabelecimento de *ensembles*, ou seja, a diminuição do número de *falsos positivos* é priorizada.



**Figura 27 – Variação do desempenho dos classificadores após a formação do ensemble de classificadores.**

Valores positivos indicam desempenho maior para o ensemble enquanto valores negativos indicam que o melhor classificador simples é superior. Apenas para os resíduos de aminoácido glicina e prolina não houve ganho, em termos de taxa de mérito, ao testar uma segunda componente para o ensemble, ficando apenas os classificadores simples. Para os classificadores de ácido aspártico e tirosina a taxa de acerto sofreu uma pequena queda. A taxa de precisão foi aumentada em todos os casos, assim como a taxa de mérito (usada para selecionar as componentes do ensemble). Asparagina, leucina, lisina e serina (linha tracejada vertical preta) obtiveram aumento também na taxa de sensibilidade, e nenhum decréscimo em nenhuma taxa foi observado para esses quatro classificadores.

O processo de síntese de *ensembles* classificadores é considerado computacionalmente barato, principalmente quando comparado com o longo processo de seleção de variáveis. Os benefícios de aumento de precisão e, em muitos casos, de aumento de taxa de acerto (acurácia), faz com que essa nova abordagem em classificadores de IFR seja satisfatória. O ganho em acerto não é tão expressivo quanto o de precisão, no entanto a grande vantagem para uso de *ensembles* está relacionada à diminuição do sobre ajuste aos dados de treinamento, pois que cada componente impõe limites diferentes entre as classes analisadas.

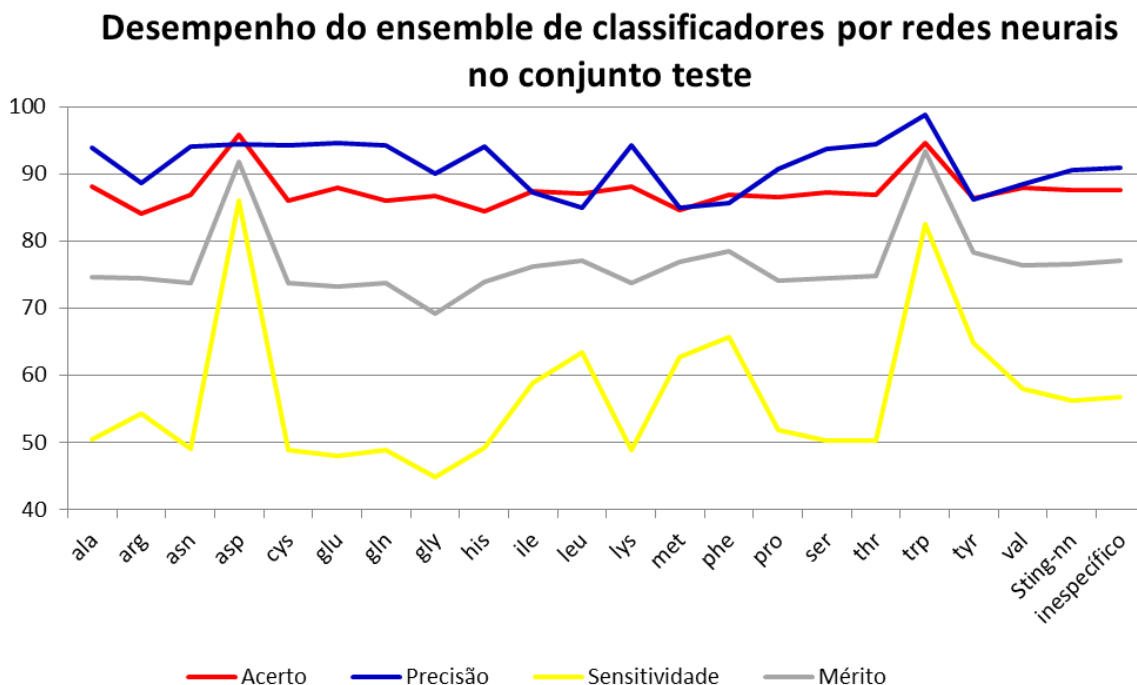
Com os *ensembles* de classificadores prontos, o conjunto de teste foi utilizado para avaliar a capacidade preditiva em um conjunto de dados independente. No total, 1705 cadeias não redundantes foram utilizadas, e no apêndice consta uma lista extensa dos indicadores de desempenho para cada uma das cadeias. Na figura 28, é mostrado o desempenho total para o

conjunto teste de cada um dos 20 classificadores específicos de aminoácidos, da combinação de todos os 20 (classificador agregado com os classificadores específicos para cada aminoácido) e também do classificador inespecífico quanto ao tipo de aminoácido.

Alguns classificadores específicos obtiveram desempenho superior do que o classificador inespecífico. Considerando taxa de acerto (acurácia), os classificadores de alanina, ácido aspártico, ácido glutâmico, lisina, triptofano e valina mostraram-se superiores. Em relação à precisão, onze classificadores foram melhores: alanina, asparagina, ácido aspártico, cisteína, ácido glutâmico, glutamina, histidina, lisina, serina, treonina e triptofano. Os classificadores de ácido aspártico, isoleucina, leucina, metionina, fenilalanina, triptofano, tirosina e valina foram superiores em sensibilidade. Evidenciando o compromisso existente entre as taxas de sensibilidade e precisão, a maior parte dos classificadores melhores em sensibilidade (em relação ao classificador inespecífico), são inferiores em relação a taxa precisão. Em relação à taxa de mérito, os classificadores de ácido aspártico, leucina, fenilalanina, triptofano e tirosina se destacaram.

Quando comparados em relação ao conjunto de teste, os classificadores específicos combinados (denominado *Sting-nn* na figura 28) e o classificador inespecífico, uma ligeira vantagem é percebida para o classificador inespecífico, diferente do que havia sido observado para o modelo por LDA. Há um aumento de 0,1% para taxa de acerto, 0,4% em precisão, 1% em sensibilidade e 0,7% em taxa de mérito.

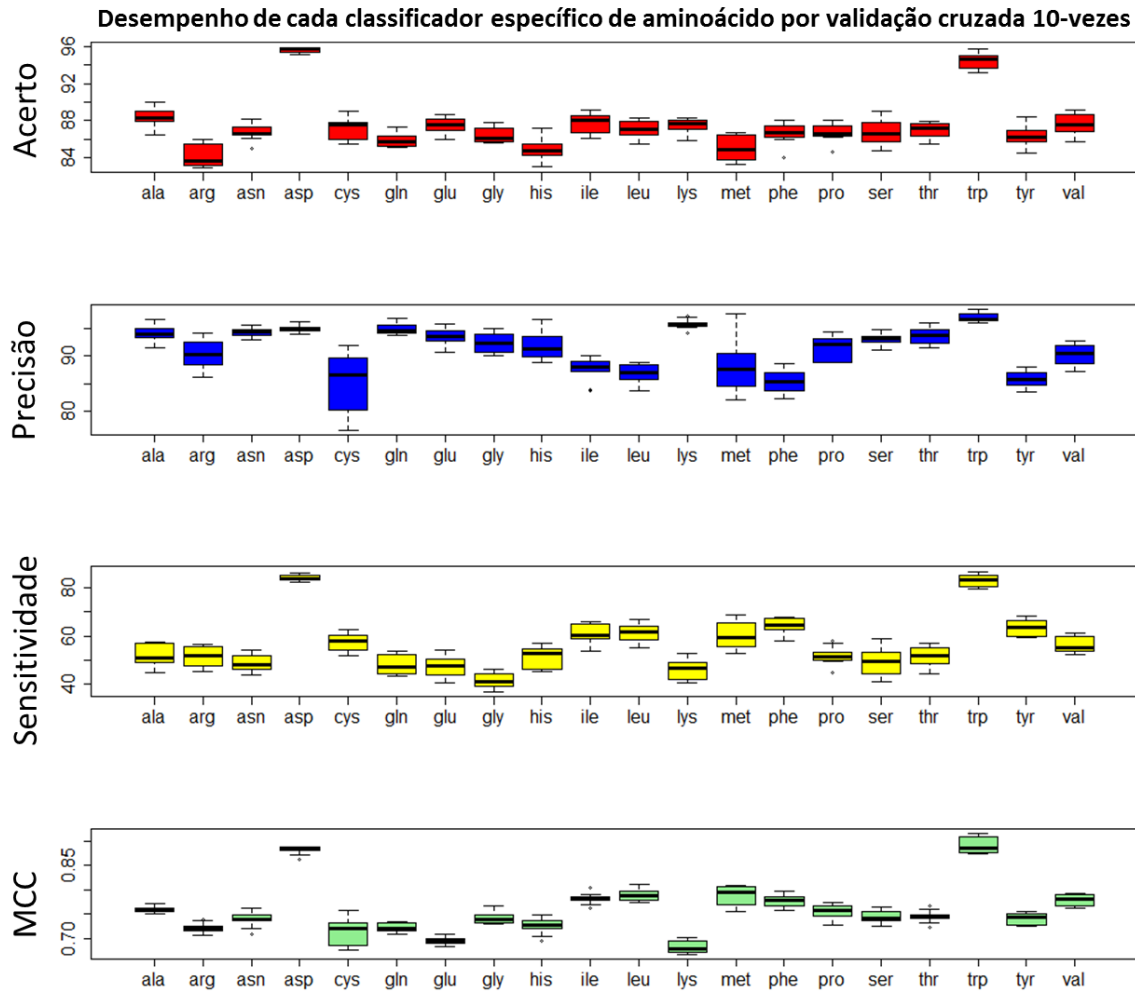
Todo o processo descrito e discutido para o classificador de redes neurais com processo de seleção de variáveis foi repetido utilizando a técnica estatística de validação cruzada com 10 pastas. Para validação cruzada o objetivo é mostrar que o desempenho obtido com o método de *holdout* é estatisticamente significativo e não sofre com grandes variações de acordo com a escolha do conjunto de *teste*. Em cada etapa, a extensa lista de seleção de variáveis apresentou algumas variações, no entanto, na média a importância de cada descritor se manteve. Em especial, destacamos que os descritores de *densidade* e *esponja* sempre ocuparam as últimas posições da lista (indicativo de grande importância no processo de distinção entre IFR e FSR). No apêndice, mostramos a lista detalhada de cada processo de seleção de variáveis específicos para todos os tipos aminoácidos em relação a validação cruzada. O classificador inespecífico não foi recalculado em validação cruzada devido ao alto custo computacional para a síntese desse classificador.



**Figura 28 – Desempenho para o conjunto de teste dos ensembles classificadores específicos para cada aminoácido, da combinação dos 20 (Sting-nn) e do inespecífico quanto ao tipo do aminoácido.**

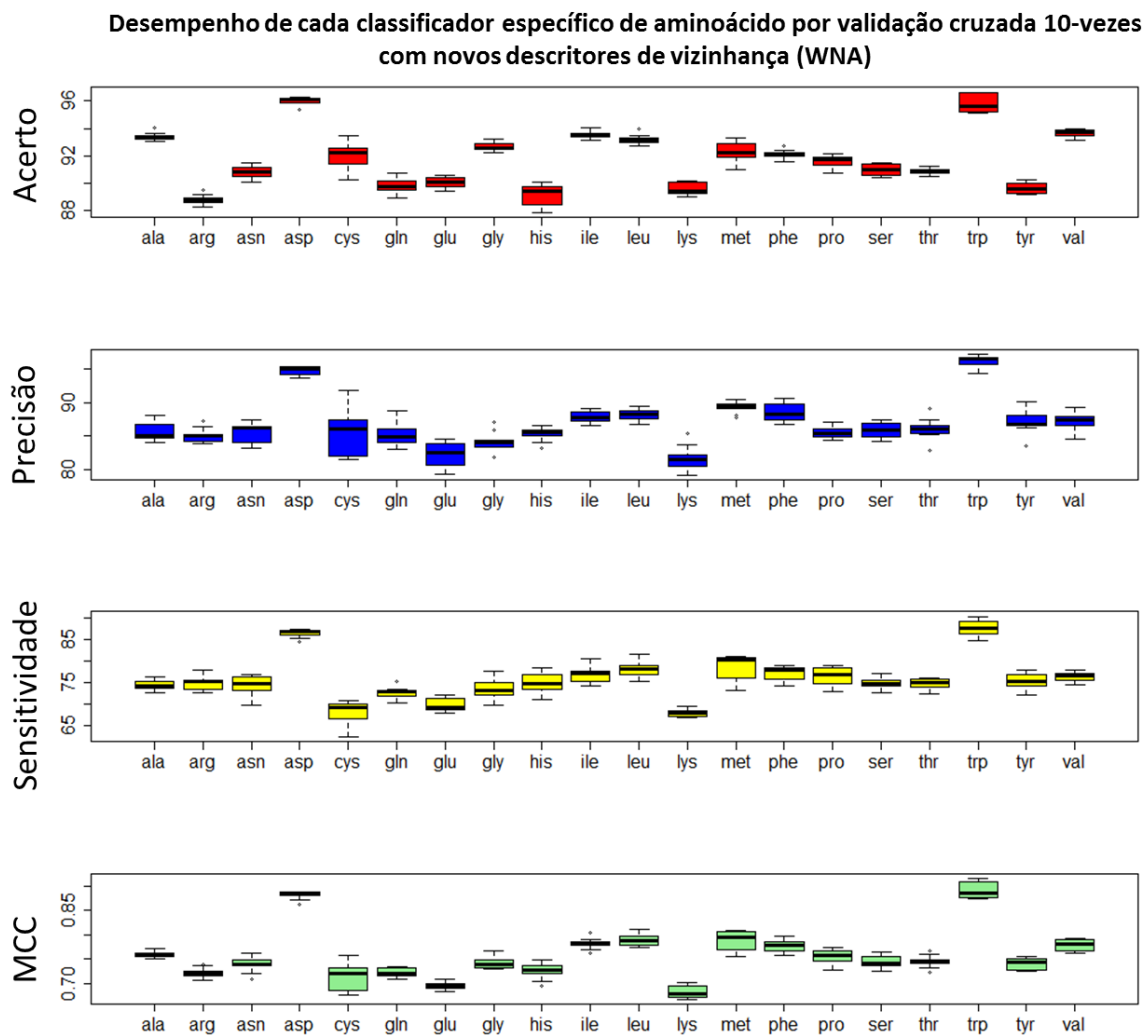
Seis classificadores específicos (alanina, ácido aspártico, ácido glutâmico, lisina, triptofano e valina) mostraram-se com taxa de acerto superior ao classificador inespecífico. Onze classificadores específicos mostraram-se superior quanto à precisão em relação ao inespecífico. Oito classificadores específicos mostraram-se superior quanto a sensibilidade em relação ao inespecífico. Cinco classificadores específicos mostraram-se com taxa de mérito superior ao classificador inespecífico. Comparando os 20 classificadores combinados com o classificador inespecífico, o conjunto de teste mostra que ambos obtiveram desempenho próximo, sendo que o classificador inespecífico é ligeiramente superior (0,1% em acerto, 0,4% em precisão, 1% em sensibilidade e 0,7% em taxa de mérito)

A figura 29 mostra que o desempenho de cada classificador específico de aminoácido mantém o desempenho durante a validação cruzada, ou seja, o desempenho é robusto quanto a escolha do conjunto teste, apresentando desempenho similares para 10 conjuntos não redundantes. Os resíduos de aminoácido cisteína e metionina apresentaram a maior dispersão em relação a taxa de precisão e MCC, este último em uma escala reduzida. O mesmo não é observado quando comparado com as taxas de acerto (acurácia) e sensibilidade.



**Figura 29 – Desempenho de cada classificador específico de aminoácido por validação cruzada.** As barras mostram que a maior parte dos classificadores específicos de aminoácidos apresentam desempenho próximo quando testados em 10 conjuntos diferentes e não redundantes.

Repetimos o processo de validação cruzada com 10 pastas, acrescentando os descritores ponderados pela vizinhança espacial. Os mesmos grupos utilizados anteriormente foram reutilizados nessa etapa, ou seja, não houve uma nova distribuição aleatória em 10 grupos diferentes. A figura 30 mostra o resultado obtido para as taxas de acerto, precisão, sensibilidade e MCC. A tendência para esse novo grupo de variáveis é de aumentar a taxa de acerto entre 4 e 6 pontos percentuais, mantendo-se sempre acima de 88%, e diminuir a precisão entre 5 e 8 pontos percentuais, mantendo-se sempre acima de 80%. A grande vantagem desse novo grupo de classificadores foi aumentar a taxa de sensibilidade entre 20 e 25 pontos percentuais. A taxa de MCC ficou basicamente inalterada.



**Figura 30 – Desempenho de cada classificador específico de aminoácido por validação cruzada com parâmetros ponderados pela vizinhança.**

As barras mostram que a maior parte dos classificadores específicos de aminoácidos apresentam desempenho próximo quando testados em 10 conjuntos diferentes e não redundantes.

Dessa forma, é indicado que o uso de descritores ponderados pela vizinhança (Sting-*nn*-WNA) é satisfatório, reduzindo em mais de 20% o número de falsos negativos, e acrescentando cerca de apenas 8% de falsos positivos. A área de interface corretamente predita aumenta enquanto a confiabilidade permanece comparativamente alta.

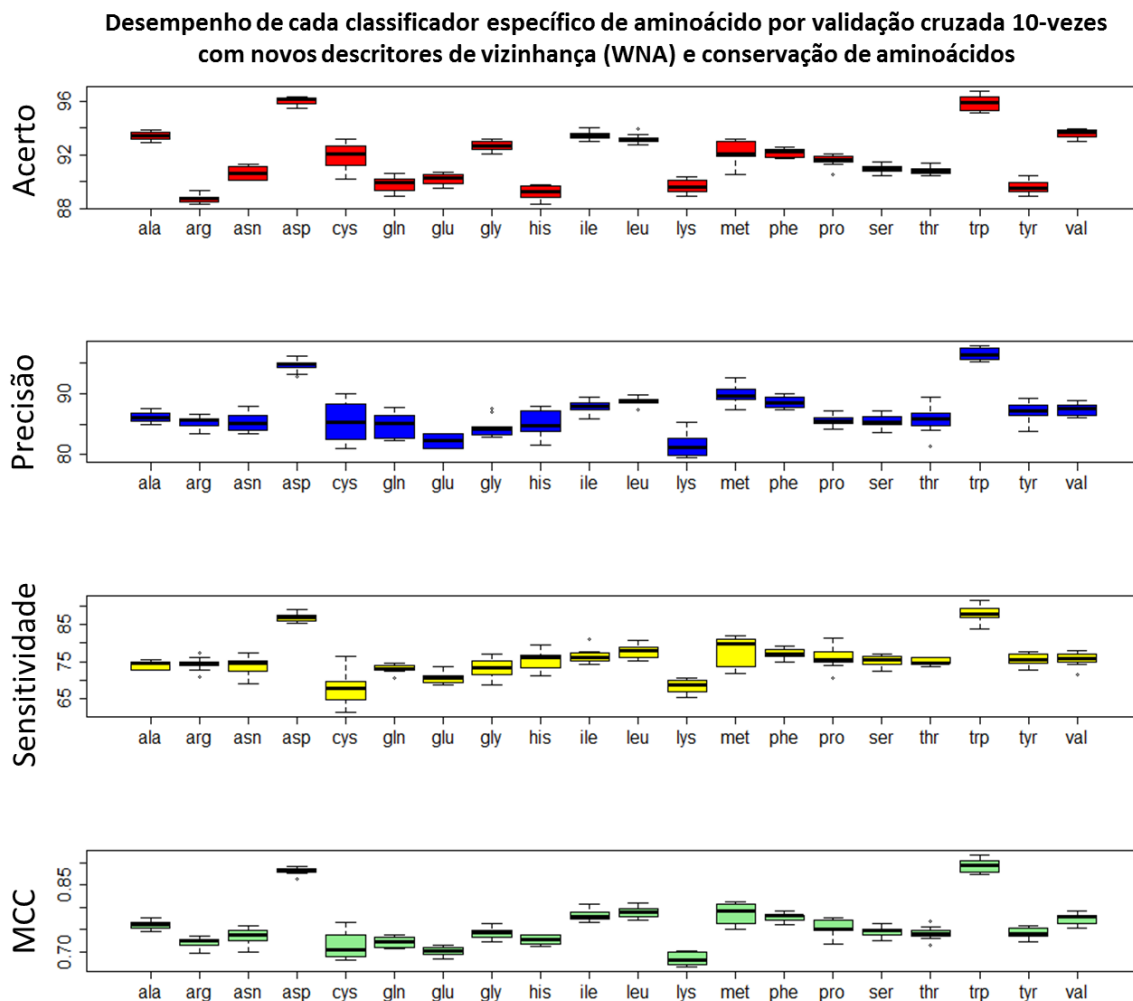
Como todos os métodos presentes na literatura utilizam descritores de conservação aminoácidos, resolvemos incorporar tais descritores, disponíveis no STING\_DB, e observar o desempenho em relação ao classificador com parâmetros físico-químico e estruturais e de vizinhança. A figura 31 mostra que não houve mudança significativa para nenhum dos



classificadores treinados utilizando o mesmo protocolo de validação cruzada com 10 pastas. Isso indica que ao combinar os descritores de conservação com os demais descritores físico-químico e estruturais não há nenhuma vantagem para diferenciar resíduos de aminoácido que fazem parte da interface dos resíduos de aminoácido da superfície livre, ou seja, a descrição do nano-ambiente de cada resíduo de aminoácido consegue deter a quantidade de informação necessária para a conservação de aminoácidos não desempenhe nenhum papel fundamental no processo de classificação. Esse resultado é o mesmo encontrado para o classificador por LDA mostrado na seção 3.7.

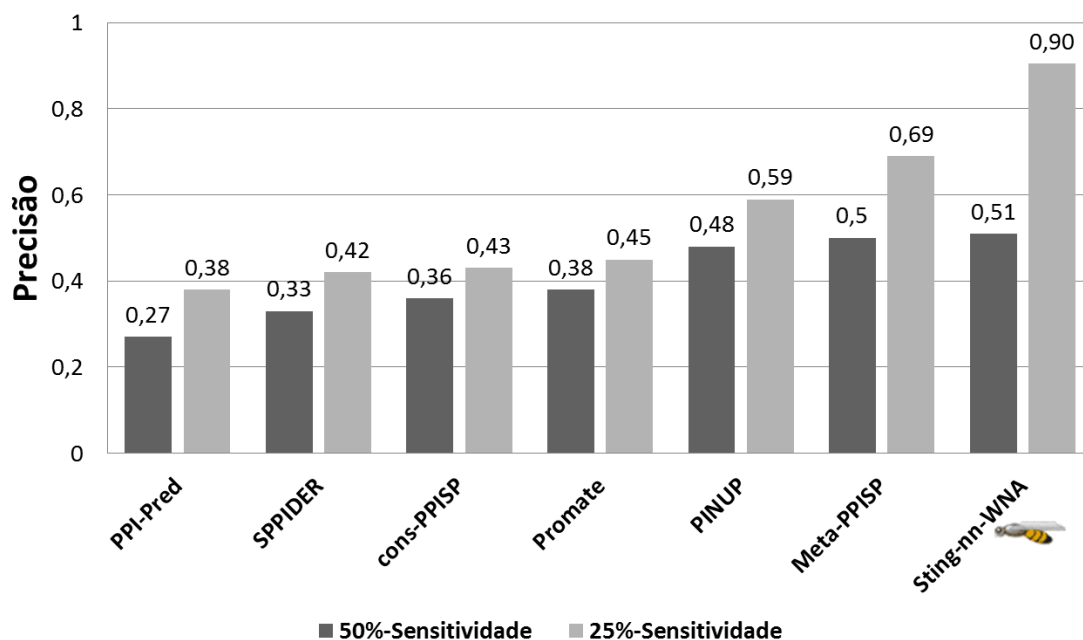
Para prosseguir com a comparação de outros métodos disponíveis na literatura, a saída da rede neural passou pela função de ativação *Softmax* (Sutton e Barto, 1998). Dessa forma, o resultado dos dois neurônios de saída de cada uma das componentes do *ensemble* de classificadores passa pela função de ativação *Softmax*, a média desses valores é calculada e, por fim, a função de ativação *Softmax* é aplicada novamente. Com esse protocolo, obtemos saídas correlacionadas com soma 1, ou seja, uma interpretação probabilística é considerada, com a possibilidade de varrer o limiar de classificação para o *ensemble* de redes neurais.

Para comparar com outros métodos disponíveis na literatura, é preciso utilizar o valor de precisão para o qual a sensibilidade, ou seja, cobertura da interface, é de mesmo patamar entre os modelos classificadores (figura 32). Utilizando o mesmo conjunto de dados que Zhou e Qin (2007-a), Sting-nn-WNA é o melhor em precisão para ambos os valores de cobertura da interface (sensibilidade). Para uma taxa de 50% de sensibilidade, Sting-nn-WNA está apenas um ponto percentual à frente do classificador Meta-PPISP em precisão. Porém, ao reduzirmos a cobertura da interface predita para 26%, Sting-nn-WNA alcança 90% de confiabilidade (precisão) para a classificação de IFR (todos os tipos de aminoácidos), enquanto que o segundo colocado (Meta-PPISP) atingi apenas 70% de confiabilidade.



**Figura 31 – Desempenho de cada classificador específico de aminoácido por validação cruzada com parâmetros ponderados pela vizinhança e com conservação de aminoácidos.** As barras mostram que a maior parte dos classificadores específicos de aminoácidos apresentam desempenho próximo quando testados em 10 conjuntos diferentes e não redundantes.

O modelo classificador Sting-*nn*-WNA ficou ligeiramente abaixo do classificador Sting-LDA-WNA, apresentado na seção 3.7, para o conjunto de dados 35Enz. No entanto, para o desempenho do conjunto teste por validação cruzada com 10 pastas, o classificador por *ensemble* de redes neurais é superior em taxa de acerto, sensibilidade e MCC. Ainda, a obtenção de forma detalhada e objetiva dos fatores que definem a interface de complexos entre proteínas é uma grande vantagem do método por *ensemble* de redes neurais com processo de seleção de variáveis.

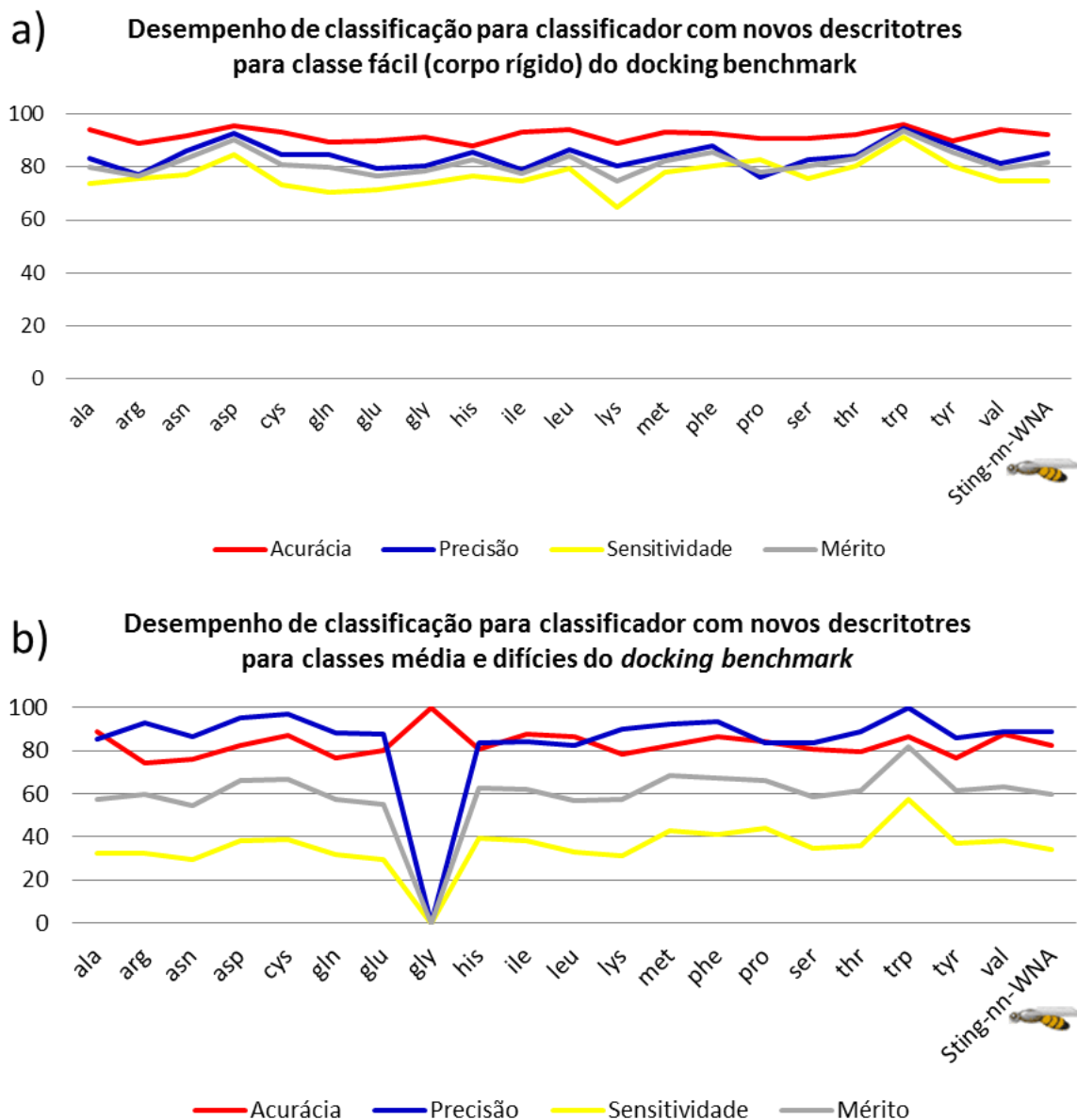


**Figura 32 – Comparação entre o classificador Sting-nn-WNA, desenvolvido nesta tese, e outros métodos disponíveis na literatura utilizando o mesmo conjunto de dados para teste (35Enz).**

Aplicando o modelo de rede neural artificial no *docking benchmark* obteve-se o resultado mostrado na figura 33. O melhor desempenho foi observado para o conjunto teste composto pelos casos fáceis (corpo rígido, figura 33-a) do *docking benchmark*, uma vez que pouca mudança conformacional mantém o nano-ambiente de cada aminoácido. Para os casos de média e alta dificuldade a precisão para o classificador Sting-nn-WNA (agregado dos 20 classificadores específicos de aminoácido) é maior, mas todas as outras taxas são reduzidas. Para o resíduo de aminoácido glicina, há apenas uma entrada no conjunto de média e alta dificuldade, sendo que essa entrada é referente à classe FSR. Apesar de essa entrada ser corretamente classificada, o valor de verdadeiros positivos é zero, anulando as medidas de precisão e sensibilidade, além da taxa de mérito.

Os resultados para a classe média e difícil podem ser considerados como a pior estimativa do resultado do classificador Sting-nn-LDA quando usado para prever interfaces de complexos ainda desconhecidos, uma vez que trata dos casos em que há mudança conformacional significativa no processo de formação do complexo proteico. De forma semelhante ao teste no conjunto 35Enz, a precisão observada é em torno de 90%, com 80% de todos os resíduos de aminoácido sendo previstos corretamente como IFR ou FSR. Isso indica que, mesmo na presença de alta flexibilidade no processo de formação do complexo entre proteínas, a predição de

aminoácidos formadores de interface com classificadores treinados a partir de descritores extraídos de estruturas estática é de alta confiabilidade.



**Figura 33 – Avaliação do classificador por rede neural em relação ao docking benchmark, utilizando as taxas de desempenho acurácia (acerto), precisão, sensibilidade e mérito, para as classes de baixa (a) e média e alta dificuldade (b).**

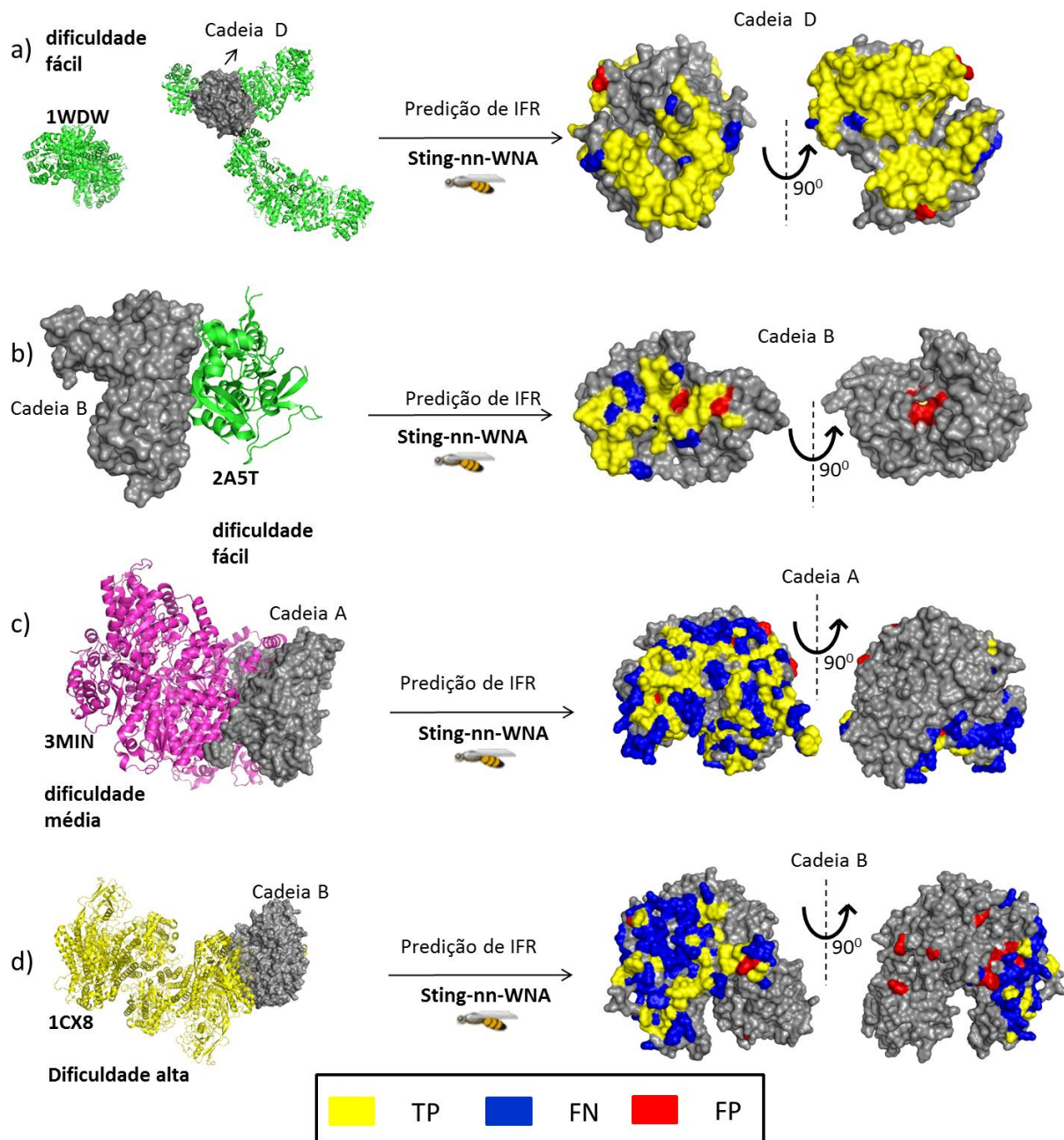
A obtenção de um modelo classificador com alto nível de precisão favorece o uso da ferramenta desenvolvida para a aplicação em *docking* proteína-proteína para o estudo em biologia

de sistemas e desenho racional de drogas. Ao prevermos em média um quarto do total de resíduos de aminoácido da interface com alta confiabilidade, algoritmos de *docking* fornecem o restante da região de interface em complexos proteicos para parceiros específicos de cada proteína. Nesse sentido, ferramentas com alta precisão possuem vantagens sobre ferramentas com alta cobertura, ou sensibilidade. Ao permitir um número maior de falsos positivos, várias interfaces erradas serão testadas por algoritmos de *docking* aumentando a chance de falsas interfaces serem testadas.

Por outro lado, caso o usuário de uma ferramenta preditiva de interface em complexos proteína-proteína esteja interessado em experimentos de mutagênese, um modelo classificador com maior número de resíduos de aminoácido previstos na interface seja mais interessante, uma vez que será possível varrer uma quantidade maior de possíveis resíduos de aminoácido importantes para estabelecer ligações entre duas ou mais proteínas.

Como estudos de casos, escolhemos aleatoriamente quatro exemplos do docking benchmark, dois da classe fácil, um da classe de média dificuldade e um da classe de alta dificuldade. A figura 34 mostra a predição de IFR para a proteína triptofano sintetase do organismo hipertermófilo *Pyrococcus furiosus* (figura 34-a), para o receptor de N-metil-D-aspartato de *Rattus norvegicus* (figura 34-b), para a proteína nitrogenase de molibdênio e ferro de *Azotobacter vinelandii* (figura 34-c) e para o receptor de transferina humano (figura 34-d).

Para os casos da classe fácil (figura 34-a e 34-b) observamos que a maior parte da interface é corretamente predita, como ilustrados pelos pontos em amarelo. Alguns pontos em azul ilustram os IFR que o classificador não conseguiu identificar corretamente. Importante notar o baixo número de pontos vermelhos, ou seja, dos FSR que foram classificados como IFR. Esse resultado indica que se essa saída fosse utilizada para a predição de estrutura quaternária por técnicas de *docking*, poucos falsos positivos iriam ser testados, reduzindo o número de interfaces proteína-proteína testadas e biologicamente irrelevantes. Para os casos de média e alta dificuldade (figura 34-c e 34-d, respectivamente), observamos um aumento do número de pontos em azul, consequência de uma redução do número de pontos em amarelo. Apesar da diminuição dos IFR corretamente preditos, notavelmente o classificador Sting-*nn*-WNA não aumenta consideravelmente o número de FSR erroneamente classificados, ou seja, o número de falsos positivos.



**Figura 34 - Predição de IFR para quatro exemplos do docking benchmark.**

Proteína triptofano sintetase do organismo hipertermófilo *Pyrococcus furiosus* (a), receptor de *N*-metil-*D*-aspartato de *Rattus norvegicus* (b), proteína nitrogenase de molibdênio e ferro de *Azotobacter vinelandii* (c) e receptor de transferina humano (d).

## 4 Conclusões

As interfaces proteicas foram objeto de estudo de vários grupos de pesquisa em todo o mundo. Nós do Grupo de Pesquisa em Biologia Computacional da Embrapa Informática Agropecuária identificamos uma lacuna no sentido de estudo embasado em descrever objetivamente o nano-ambiente que abriga os resíduos de aminoácido formadores de interface. Arelado à descrição quantitativa da importância de uma série de descritores físico-químico e estruturais previamente calculados e armazenados no banco de dados do software BlueStar STING, desenvolvemos um modelo classificador que se mostrou de grande competitividade em relação a outras ferramentas disponíveis na literatura.

A coleta de dados separados por cada tipo de aminoácido demonstra que é possível melhorar o desempenho e comparar dados que seriam perdidos na tentativa de classificar um resíduo de aminoácido como presente ou não na interface. Utilizando uma base de dados maior sobre descritores físico-químico de estruturas de proteínas (provenientes do STING\_DB) foi possível comparar os valores calculados para cada um dos 20 tipos diferentes de aminoácidos entre duas classes: resíduos de aminoácido formadores de interface (IFR) e resíduos de aminoácido que se situam na superfície livre (FSR). Esse resultado gerou uma ferramenta que está sendo integrada ao banco de dados BlueStar Sting para uso livre e pré-disponível para todas as proteínas com estrutura conhecidas e depositadas no banco de dados público PDB.

A análise estatística apresentada fornece indícios de que há diferenças estatisticamente relevantes entre IFR e FSR, que podem ser utilizadas para a classificação de regiões de interação entre proteínas. Ainda, mostramos que metodologias multivariadas fornecem melhor desempenho do que quando cada descritor físico-químico e estrutural é utilizado isoladamente. O amadurecimento da análise estatística revelou que grandes bancos de dados, como o utilizado neste estudo, necessitam de auxílio por métodos gráficos como, por exemplo, o uso de *boxplots*.

Mostramos que diversos modelos classificadores de técnicas de mineração de dados, entre modelos lineares, árvores de decisão e modelos não lineares, como SVM e redes neurais, apresentam poder de classificação diferente quanto ao problema estudado. Ainda, mostramos que os diferentes métodos quando aplicados nos vinte distintos tipos de aminoácidos apresentam poder de generalização diferente, ou seja, ao aprender com os dados disponíveis sobre verdadeiros complexos proteína-proteína os diversos métodos aplicam de forma diferente a

informação disponível para minimizar o erro de classificação. Todos os métodos são considerados eficientes quanto à tarefa de aprendizado supervisionado e aplicação em predições em um conjunto de dados diferente sobre descritores físico-químicos e estruturais. Mostramos ainda que o modelo classificador de redes neurais apresentou o melhor desempenho com os dados disponíveis para a maioria dos tipos de aminoácido. Devido a este resultado, abordamos o problema com a utilização do processo de seleção de variáveis, resultando em uma listagem dos descritores mais importantes para diferenciar resíduos de aminoácido que estão na superfície livre em estruturas tridimensionais de proteínas. A lista dos descritores mais importantes sugere que a densidade local e espaço vazio, descritor esponjicidade, desenham o nano-ambiente de IFR. De forma interessante, outros descritores, como por exemplo, *ligações de hidrogênio, potencial eletrostático* etc., conferem a cada tipo de aminoácido uma listagem única como fatores que propiciam a interação entre proteínas. A técnica de construção de *ensemble* de classificadores mostrou-se útil no sentido de aumentar a precisão do classificador final, tornando assim o desempenho mais confiável quanto à classificação de IFR. A incorporação de descritores físico-químicos e estruturais da vizinhança de cada resíduo de aminoácido (descritores WNA) mostrou-se extremamente útil para fins de classificação, reduzindo em torno de 20% o número de falsos negativos e melhorando o desempenho geral do classificador final.

Ao comparar os resultados obtidos com outros métodos disponíveis na literatura, o classificador Sting-nn-WNA mostrou-se como o mais preciso. Até mesmo o classificador utilizando o modelo simples de LDA atingiu pontuações satisfatórias quanto à classificação de IFR, quando comparado com modelos mais sofisticado utilizando SVM e redes neurais.

Esse trabalho se insere no contexto pós-genômico e visa contribuir para o desenvolvimento de uma abordagem estrutural relacionada à biologia de sistemas. Recentemente, estudos que reportam milhares de interações observadas entre proteínas (Havugimana *et al.*, 2012) estão disponíveis. Tais estudos podem ser utilizados como entrada para cálculos mais sofisticados que utilizam a informação sobre a estrutura das proteínas (resolvidas ou modeladas), como é o caso da abordagem discutida nesta tese. Ao conhecer possíveis parceiros de interação por técnicas experimentais, o uso dos descritores calculados sobre a estrutura de cada um dos parceiros de interação prevê a região de interface que serve então como entrada para algoritmos de *docking* e/ou dinâmica molecular.

A integração de diversas metodologias computacionais, entre elas, alinhamento sequencial com homólogos distantes, modelagem por homologia, predição de regiões de interação entre



proteínas, DNA e outras macromoléculas, *docking*, dinâmica molecular etc., é tida como o objetivo final para um melhor entendimento dos processos que acontecem dentro e fora da célula. Especificamente, esperamos que a ferramenta desenvolvida ajude na importante tarefa de prever regiões de comunicação entre proteínas, útil para o desenho racional de drogas e agroquímicos e estudos bioquímicos sobre os resíduos de aminoácido importantes para o estabelecimento da função proteica.



## 5 Referências

- Acharya C, Coop A, Polli JE e MacKerell Jr. AD (2011) Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr Comput Aided Drug Des.* March 1; 7(1): 10–22.
- Alloy P e Russell RB (2006) Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology* 7, 188-197.
- Altschul SF, Gish W, Miller W, Myers EW e Lipman DJ (1990) Basic local alignment search tool *J. Mol. Biol.* 215:403-410.
- Bahadur RP, Chakrabarti P, Rodier F, Janin J. (2004) A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol.* Feb 27;336(4):943-55.
- Beeley LJ e Duckworth DM (1996) The impact of genomics on drug design. *Drug Discovery Today*, 1, 11, , pp. 474-480(7)
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. *Nucleic Acids Res.* Jan 1;28(1):235-42.
- Bradford JR, Westhead DR. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*; 21:1487-1494.
- Breiman L (2001) Random Forests. *Machine Learning*, 45:5-32.
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery.* 2(2):1-47.
- Caffrey DR, Somaroo S, Hughes JD, Mintseris J e Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* Jan;13(1):190-202.
- Chen H e Skolnick J (2008) M-TASSER: An algorithm for protein quaternary structure prediction. *Biophysical Journal* 94: 918-928.
- Chen H e Zhou HX. (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins.* Oct 1;61(1):21-35.
- Chothia C e Janin J (1975) "Principles of protein-protein recognition" *Nature* 256: 705–708.
- da Silveira CH, Pires DE, Minardi RC, Ribeiro C, Veloso CJ, Lopes JC, Meira W Jr, Neshich G, Ramos CH, Habesch R, Santoro MM. (2009) Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins.* Feb 15;74(3):727-43.

- Domingos P e Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29: 103-130.
- Everitt BS e Hothorn T, (2008) *Handbook of Statistical Analysis using R*. Chapman & Hall/CRC.
- Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters*. 27(8): 861-874.
- Gruber J, Zawaira A, Saunders R, Barrett CP, Noble ME. (2007) Computational analyses of the surface properties of protein-protein interfaces. *Acta Crystallogr D Biol Crystallogr*. Jan;63(Pt 1):50-7.
- Guyon I e Elisseeff A (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182.
- Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ER, Paccanaro A, Marcotte EM, Emili A. A census of human soluble protein complexes. *Cell*. 2012 Aug 31;150(5):1068-81.
- Hothorn T, Hornik K e Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.
- Hwang H, Vreven T, Janin J e Weng Z (2010) Protein-Protein Docking Benchmark Version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78: 3111–3114.
- Janin J e Wodak S (2007) The Third CAPRI Assessment Meeting *Structure* 15: 755-759.
- Johnson DE (1998) *Applied Multivariate Methods for Data Analysis* Brooks/Cole Publishing Company.
- Jones S e Thornton JM (1996) Principles of protein-protein interactions. *Proc. Nati. Acad. Sci.* 93: 13-20.
- Kitano H (2002) Computational systems biology. *Nature*. Nov 14;420(6912):206-10.
- Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31:249-268.
- Krissinel E e Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372, 774—797.
- Lee B e Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biology* 55: 379-400.
- Liang S, Zhang C, Liu S, Zhou Y. (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* 2006 Aug 7;34(13):3698-707.
- Luenberger DG (1984) *Linear and Nonlinear Programming*, 2 ed Addison-Wesley Publishing Group.

- Lybrand TP (1995) Ligand-protein docking and rational drug design. *Curr Opin Struct Biol.* Apr;5(2):224-8.
- Martin AC (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics.* Dec 1;21(23):4297-301.
- McCulloch CE (2000) Generalized Linear Models. *Journal of the American Statistical Association,* 95(452):1320-1324.
- Murthy SK (1998) Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery* 2: 345–389.
- Nelder JA and Wedderburn RWM (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* 135(3):370-384.
- Neshich G, Rocchia W, Mancini AL, Yamagishi ME, Kuser PR, Fileto R, Baudet C, Pinto IP, Montagner AJ, Palandrani JF, Krauchenco JN, Torres RC, Souza S, Togawa RC, Higa RH. (2004) JavaProtein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Res.* Jul 1;32(Web Server issue):W595-601.
- Neshich G, Mancini AL, Yamagishi MEB, Kuser-Falcão PR, Fileto R, Pinto IP, Palandrani JF, Krauchenco JN, Baudet C, Montagner AJ e Higa RH (2005) STING Report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database. *Nucleic Acid Research* 33:D269-D274.
- Neshich G, Mazoni I, Oliveira SRM, Yamagishi MEB, Kuser-Falcão PR, Borro LC, Morita DU, Souza KRR, Almeida GV, Rodrigues DN, Jardine JG, Togawa RC, Mancini AL, Higa RH, Cruz SAB, Vieira FD, Santos EH, Melo EH e Santoro EH (2006) The Star STING server: a multiplatform environment for protein structure analysis. *Genet. Mol. Res.* 5(4): 717 – 722.
- Neuvirth H, Raz R, Schreiber G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *JMolBiol,* 338:181-199.
- Noether GE (1991) *Introduction to statistics : the nonparametric way.* New York, N.Y.: Springer, c. 1ª Ed.
- Ofran Y, Rost B. (2003) Analysing six types of protein-protein interfaces. *J Mol Biol.* Jan 10;325(2):377-87.
- Oliveira SRM, Almeida GV, Souza KRR, Rodrigues DN, Kuser-Falcão PR, Yamagishi MEB, Santos EH, Vieira FD, Jardine JG e Neshich G (2007) Sting\_rdb: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. *Genet. Mol. Res.* 6 (4): 911 – 922.
- Parrill AL (1996) Evolutionary and genetic methods in drug design. *Drug Discovery Today,* Vol 1, 12, December, pp. 514-521(8).

- Pereira JGC (2012) Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir dos dados estruturais. Dissertação de mestrado em Genética e Biologia Molecular. IB-Unicamp.
- Ponstingl H, Kabir T, Gorse D e Thornton JM (2005) Morphological Aspects of Oligomeric Protein Structures. *Progress in Biophysics and Molecular Biology* 89: 9–35.
- Porollo A e Meller J (2007) Prediction-based fingerprints of protein-protein interactions. *Proteins*, 66:630-645.
- Puma-Villanueva WJ (2006) Comitê de Máquinas em Predição de Séries Temporais. Dissertação de mestrado em Engenharia Elétrica. FEEC-Unicamp.
- Puma-Villanueva WJ (2011) Síntese automática de redes neurais artificiais com conexões à frente arbitrárias. Tese de doutorado em Engenharia Elétrica. FEEC-Unicamp.
- Radzicka A e Wolfenden R (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* 27: 1664–1670.
- Reichmann D, Rahat O, Cohen M, Neuvirth H e Schreiber G (2007) The Molecular Architecture of protein-protein binding sites. *Current Opinion in Structural Biology* 17: 67-76.
- Ribeiro C, Togawa RC, Neshich IAP, Mazoni I, Mancini AL, de Melo Minardi RC, da Silveira CH, Jardine JG, Santoro MM, Neshich G (2010) Analysis of binding properties and specificity through identification of the interface forming residues (IFR) for serine proteases in silico docked to different inhibitors. *BMC Structural Biology*, 10:36
- Rocchia, W e Neshich, G (2007) Electrostatic Potential Calculation for biomolecules - creating a database of pre-calculated values reported on a per residue basis for all PDB protein structures. *Genet. Mol. Res.* 6 (4): 923-936.
- Rumelhart D. E. e McClelland J.L. (1986) Parallel distributed processing: Exploration in the microstructure of cognition. MIT Press, Cambridge, Massachusetts, vol. 1,.
- Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. (2004) A structural perspective on protein-protein interactions. *Curr Opin Struct Biol.* Jun;14(3):313-24.
- Schneider M, Fu X e Keating AE (2009) X-ray vs NMR structures as templates for computational protein design. *Proteins* Oct;77(1): 97-110.
- Schrauber H, Eisenhaber F, Argos P (1993) Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol.* Mar 20;230(2):592-612.
- Segal MR (1988) Regression Trees for Censored Data. *Biometrics*, 44, 35–47.
- Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21(20):3940-3941.

- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. (2008) Conditional variable importance for random forests. *BMC Bioinformatics* 11; 9:307.
- Sutton RS e Barto AG (1998) Reinforcement Learning: An Introduction. The MIT Press, Cambridge, MA.
- Tsuchiya Y, Kinoshita K e Nakamura H (2006) Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Protein Engineering, Design & Selection* 19(9): 421–429.
- UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acid Res.* jan; 37 (Database issue): D 169-74.
- van der Laan, M, Jiann-Ping Hsu/Karl E. Peace, KE, e Rose S (2010) Next Generation of Statisticians Must Build Tools for Massive Data Sets. September 1<sup>st</sup>, *Statistics Ready for a Revolution - Amstat News*.
- Vapnik V (1995) *The Nature of Statistical Learning Theory*. Springer Verlag. 1<sup>a</sup> Ed.
- Venables WN e Ripley BD (2002) *Modern Applied Statistics with S*. Springer, New York, 4a Ed.
- Xenarios I e Eisenberg D (2001) Protein interaction databases. *Curr Opin in Biotech*, 12 : 334–339.
- Xu Q, Canutescu A, Wang G, Shapovalov M, Obradovic Z e Dunbrack Jr R (2008) Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins. *Journal of Molecular Biology*, 381 (2): 487-507.
- Wade RC (1997) 'Flu' and structure-based drug design. *Structure*. Sep 15;5(9):1139-45.
- Werbos P (1974) *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, PhD thesis, Harvard University.
- Witten IH, Frank E, Hall MA (2011) *Data Mining: Practical Machine Learning Tools and Techniques* Morgan Kaufmann, 3<sup>a</sup> Ed.
- Zar JH (1999) *Biostatistical Analysis*. Prentice Hall Inc. 4<sup>a</sup> Ed.
- Zhou HX e Qin S (2007-a) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* 23, 2203-2209.
- Zhou HX e Qin S (2007-b) meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* 23, 3386-3387.
- Zsoldosa Z, Szaboa I, Szaboa Z, Johnson AP (2003) Software tools for structure based rational drug design. *Journal of Molecular Structure: THEOCHEM*, 666-667, Pages 659-665.





## 6 Apêndice

### 6.1 *Apêndice 1 – Resultados dos testes estatísticos univariados*

Resultados dos testes estatísticos univariados para cada um dos descritores de cada aminoácido. Descritores que apresentam o valor de mais de 1% indicam pouco potencial estatístico para distinguir IFR dos FSR. Complemento da seção 3.3.

Arginina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0	0	0
Energia de Contatos Aromáticos	1,00E-04	0	0
Energia de Ligação de Hidro - SWWS	0,0049	0,6937	1
Energia de Ligação de Hidro - MWWS	0,0017	0,5762	0,9744
Energia de Ligação de Hidro - MWWM	0,0069	0,9097	1
Energia de Ligação de Hidro - SWS	0,0829	0,1349	0,9998
Energia de Ligação de Hidro - MWS	0,7442	0	0,0019
Energia de Ligação de Hidro - MWM	0,0259	0,4462	0,9919
Energia de Ligação de Hidro - SS	0,0917	0	0,001
Energia de Ligação de Hidro - MS	0	0	0
Energia de Ligação de Hidro - MM	0	0,0568	0
Energia de contatos carregados repulsivos	0,1574	0	0
Energia de contatos carregados atrativos	0	0	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no C-beta	0	0	0
CLO no LHA	0	0	0

Densidade no ca3	1,00E-04	0,0266	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0,3431	8,00E-04	0
EP no LHA	0	2,00E-04	0
EP na Superfície	0	0	0
EP Médio	0	0,0062	0
Esponja no ca3	0	0	0
Esponja no ca5	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,6348	0	0
Energia de Contatos Aromáticos não usados	0,0064	0	0
Energia de Ligação de Hidro – SWWS não usada	1,00E-04	0	0
Energia de Ligação de Hidro – MWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0
Energia de Ligação de Hidro – SWS não usada	2,00E-04	0	0
Energia de Ligação de Hidro – MWS não usada	0,3834	0,0123	1,00E-04
Energia de Ligação de Hidro – MWM não usada	0,0023	0,0013	0,0043
Energia de Ligação de Hidro – SS não usada	0,0136	0,0288	0,0032
Energia de Ligação de Hidro – MS não usada	0,7679	3,00E-04	0
Energia de Ligação de Hidro – MM não usada	0,0055	0,4858	1,00E-04
Energia de Contatos carregados repulsivos não usados	0,1014	0,4225	0,0033
Energia de Contatos carregados atrativos não usados	2,00E-04	0	0
Energia de Contatos Hidrofóbicos não usados	0,0031	0	0
phi	0,1165	0,7799	0,653
psi	0,8629	0,2226	0,2335

chi_1	0,0318	0,0809	0,1732
chi_2	0,1357	0,0767	0,1701
chi_3	0,1252	0,1533	0,0035
chi_4	0,0181	0,0143	1,00E-04

Asparagina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,9289	0	0
Energia de Contatos Aromáticos	0,5919	0,326	1
Energia de Ligação de Hidro - SWWS	0,3598	0,4608	1
Energia de Ligação de Hidro - MWWS	0,9684	0,1492	0,9832
Energia de Ligação de Hidro - MWWM	0,1543	0,4375	1
Energia de Ligação de Hidro - SWS	0,2317	0,9136	1
Energia de Ligação de Hidro - MWS	0,4786	0	0,0068
Energia de Ligação de Hidro - MWM	0,7405	0,004	0,5643
Energia de Ligação de Hidro - SS	0	0	0
Energia de Ligação de Hidro - MS	0	0	0
Energia de Ligação de Hidro - MM	0	0,7304	0
Energia de contatos carregados repulsivos	0,2592	0,0465	1
Energia de contatos carregados atrativos	0,3728	0,2993	1
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no C-beta	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0

Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0,0698	0,4383	0,0307
EP na Superfície	0	0	0
EP Médio	0	0	0
Esponja no ca3	0	0	0
Esponja no ca5	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0244	0,8893	0,0124
Energia de Contatos Aromáticos não usados	0,4067	0,1809	1
Energia de Ligação de Hidro – SWWS não usada	0,0071	0,0026	0,0509
Energia de Ligação de Hidro – MWWS não usada	0,1043	0,0367	0,0078
Energia de Ligação de Hidro – MWWM não usada	9,00E-04	0,0011	0,0099
Energia de Ligação de Hidro – SWS não usada	0,0029	0,0023	0,0062
Energia de Ligação de Hidro – MWS não usada	0,0425	0,1311	0,006
Energia de Ligação de Hidro – MWM não usada	0,0052	0,0415	0,0163
Energia de Ligação de Hidro – SS não usada	0,0222	0,4482	5,00E-04
Energia de Ligação de Hidro – MS não usada	0,094	0,011	2,00E-04
Energia de Ligação de Hidro – MM não usada	0,7694	0,8202	0,0263
Energia de Contatos carregados repulsivos não usados	0,4748	0,5452	1
Energia de Contatos carregados atrativos não usados	0,559	0,5535	1
Energia de Contatos Hidrofóbicos não usados	0,1398	0	0
phi	0,2309	0,0586	1,00E-04
psi	0,1869	0,0182	0
chi_1	0,3794	0,0081	5,00E-04
chi_2	0,4897	0,7791	0,3723

chi_3	0,899	0,9943	1
chi_4	0,5224	0,7885	1

### Ácido aspártico:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	2,00E-04	0	0
Energia de Ligação de Hidro - SWWS	0,0475	0,3155	1
Energia de Ligação de Hidro - MWWS	0,227	0,073	0,9325
Energia de Ligação de Hidro - MWWM	2,00E-04	0,0357	0,9958
Energia de Ligação de Hidro - SWS	0,0018	0,615	0,9398
Energia de Ligação de Hidro - MWS	0,9455	0	0,001
Energia de Ligação de Hidro - MWM	0,0325	0,0463	0,8769
Energia de Ligação de Hidro - SS	0,3535	0	0
Energia de Ligação de Hidro - MS	0,2284	0	0
Energia de Ligação de Hidro - MM	0	0,4311	0
Energia de contatos carregados repulsivos	0,8991	0	0
Energia de contatos carregados atrativos	0	0	0
Energia de Contatos Hidrofóbicos	0,0119	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0,2184	0,012	0
Densidade no lha5	0,3979	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0

Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,007	0,2235	0
Energia de Ligação de Hidro – SWWS não usada	0,0093	0,0018	0,1122
Energia de Ligação de Hidro – MWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0
Energia de Ligação de Hidro – SWS não usada	0,1202	0,0028	0,0644
Energia de Ligação de Hidro – MWS não usada	0,0256	0,0171	4,00E-04
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – SS não usada	0,0503	0,0143	0
Energia de Ligação de Hidro – MS não usada	0,0167	0,851	2,00E-04
Energia de Ligação de Hidro – MM não usada	0,0051	0,3277	1,00E-04
Energia de Contatos carregados repulsivos não usados	0,4602	0,2694	0,0367
Energia de Contatos carregados atrativos não usados	0,9013	0,0028	0
Energia de Contatos Hidrofóbicos não usados	0,0085	0,4473	0
phi	0,386	0,4587	0,0049
psi	0,0468	0,0177	4,00E-04
chi_1	0,6803	0,2712	0,0015
chi_2	0,1571	0,0267	0,0042

#### Cisteína:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	1,00E-04	0	0
Energia de ligação dissulfeto	1,00E-04	0	0
Energia de Ligação de Hidro - MWWS	0,0354	0,0184	0,982
Energia de Ligação de Hidro - MWWM	0,821	0,0936	0,9996
Energia de Ligação de Hidro - MWS	0,708	0,0576	0,9763
Energia de Ligação de Hidro - MWM	0,6692	0,0022	0,4271
Energia de Ligação de Hidro - MS	0,4211	0,1046	0,7962
Energia de Ligação de Hidro - MM	0,2639	0,002	0
Energia de Contatos Hidrofóbicos	0,0267	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0

CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca5	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0,056	0,0134	0
EP na Superfície	0,0092	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,105	0,5189	0,0061
Energia de ligação dissulfeto não usada	0,1134	0,2304	0,0186
Energia de Ligação de Hidro – MWWS não usada	0,0058	0,0053	0,004
Energia de Ligação de Hidro – MWWM não usada	0,7946	0,0646	0,0492
Energia de Ligação de Hidro – MWS não usada	0,5569	0,1091	0,0137
Energia de Ligação de Hidro – MWM não usada	0,758	0,1416	0,0838
Energia de Ligação de Hidro – MS não usada	0,3231	0,1806	0,1811
Energia de Ligação de Hidro – MM não usada	0,9459	0,6053	0,0919
Energia de Contatos Hidrofóbicos não usados	0,1089	0,2573	0,0326
phi	0,0495	0,0569	0,039
psi	0,4294	0,4366	0,4094

chi_1	0,565	0,6293	0,8241
-------	-------	--------	--------

**Glutamina:**

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	2,00E-04	1,00E-04	0
Energia de Contatos Aromáticos	0,831	0,4874	1
Energia de Ligação de Hidro - SWWS	0	0,0309	0,9407
Energia de Ligação de Hidro - MWWS	0,2661	0,798	1
Energia de Ligação de Hidro - MWWM	0,0135	0,877	0,9997
Energia de Ligação de Hidro - SWS	0,0042	0,8125	0,9999
Energia de Ligação de Hidro - MWS	0,0153	0,5482	0,9903
Energia de Ligação de Hidro - MWM	0,2397	0,2923	1
Energia de Ligação de Hidro - SS	0,309	0	0
Energia de Ligação de Hidro - MS	0,0582	0	0
Energia de Ligação de Hidro - MM	0	0,3473	0
Energia de contatos carregados repulsivos	0,6973	0,0912	1
Energia de contatos carregados atrativos	0,2752	0,036	1
Energia de Contatos Hidrofóbicos	4,00E-04	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha4	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no C-beta	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	2,00E-04	0,2219	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0



EP no Ca	4,00E-04	1,00E-04	0
EP no LHA	0,0839	0,2548	0,1214
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	3,00E-04	0,2012	8,00E-04
Energia de Contatos Aromáticos não usados	0,7073	0,9659	1
Energia de Ligação de Hidro – SWWS não usada	0,4054	0,0095	0,0116
Energia de Ligação de Hidro – MWWS não usada	4,00E-04	3,00E-04	0,0048
Energia de Ligação de Hidro – MWWM não usada	2,00E-04	1,00E-04	0,0021
Energia de Ligação de Hidro – SWS não usada	0,001	0,0017	0,0051
Energia de Ligação de Hidro – MWS não usada	0,0662	0,0024	0,009
Energia de Ligação de Hidro – MWM não usada	5,00E-04	0,0017	0,0027
Energia de Ligação de Hidro – SS não usada	0,3697	0,184	0,0884
Energia de Ligação de Hidro – MS não usada	0,4102	0,4487	0,0064
Energia de Ligação de Hidro – MM não usada	0,0407	0,6118	0,0044
Energia de Contatos carregados repulsivos não usados	0,1059	0,1063	1
Energia de Contatos carregados atrativos não usados	0,1322	0,1067	1
Energia de Contatos Hidrofóbicos não usados	0,1156	0	0
phi	0,1775	0,0291	0,0019
psi	5,00E-04	0,0018	6,00E-04
chi_1	0,0021	0,003	0,0066
chi_2	0,0038	0,0085	0
chi_3	0,9489	0,8548	0,483

#### Ácido glutâmico:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,1886	0	0
Energia de Contatos Aromáticos	0,5337	0,0344	1
Energia de Ligação de Hidro - SWWS	0,0033	0,0756	0,9994
Energia de Ligação de Hidro - MWWS	0,4007	0,1818	0,9993

Energia de Ligação de Hidro - MWWM	1,00E-04	0,0498	0,9922
Energia de Ligação de Hidro - SWS	0	0,014	0,6544
Energia de Ligação de Hidro - MWS	0,031	0,0611	0,7936
Energia de Ligação de Hidro - MWM	0,0454	0,1744	0,9994
Energia de Ligação de Hidro - SS	0,438	0	0,0302
Energia de Ligação de Hidro - MS	0,1842	0	0
Energia de Ligação de Hidro - MM	0	0	0
Energia de contatos carregados repulsivos	0,5939	0	0
Energia de contatos carregados atrativos	0	0	0
Energia de Contatos Hidrofóbicos	0,027	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,0048	0,1632	0
Densidade no ca5	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0,3296	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0661	0,2519	7,00E-04
Energia de Contatos Aromáticos não usados	0,938	0,1997	1

Energia de Ligação de Hidro – SWWS não usada	0,8139	0,3802	0,9972
Energia de Ligação de Hidro – MWWS não usada	0	0	2,00E-04
Energia de Ligação de Hidro – MWWM não usada	0,0474	0,0755	0,0199
Energia de Ligação de Hidro – SWS não usada	0,002	3,00E-04	0,0026
Energia de Ligação de Hidro – MWS não usada	3,00E-04	0,0041	0,0012
Energia de Ligação de Hidro – MWM não usada	1,00E-04	2,00E-04	2,00E-04
Energia de Ligação de Hidro – SS não usada	0,0079	0,0543	0,0032
Energia de Ligação de Hidro – MS não usada	0,0392	0,4812	8,00E-04
Energia de Ligação de Hidro – MM não usada	7,00E-04	0,0039	3,00E-04
Energia de Contatos carregados repulsivos não usados	0,9989	0,4831	0,4114
Energia de Contatos carregados atrativos não usados	0,8425	0	0
Energia de Contatos Hidrofóbicos não usados	0,8175	0	0
phi	0,4697	0,0476	0,0164
psi	4,00E-04	3,00E-04	3,00E-04
chi_1	0,0206	0,0429	0,001
chi_2	0,9755	0,7541	0,1313
chi_3	0,4472	0,1466	0,0042

#### Glicina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,4172	0	0
Energia de Ligação de Hidro - MWWS	0,3957	0,0532	0,9905
Energia de Ligação de Hidro - MWWM	0,5884	1,00E-04	0,4614
Energia de Ligação de Hidro - MWS	0,9013	0	0,0159
Energia de Ligação de Hidro - MWM	0,0015	0	1,00E-04
Energia de Ligação de Hidro - MS	0,0242	0	0
Energia de Ligação de Hidro - MM	0	0	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CPO no C-alpha	0	0	0

CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
EP no Ca	0	0	0
EP na Surperfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Energia Total de Contatos Não-usados	0,0625	0,0026	0
Energia de Ligação de Hidro – MWWS não usada	0,1227	0,0046	0,0041
Energia de Ligação de Hidro – MWWM não usada	0,0098	0,0088	0,039
Energia de Ligação de Hidro – MWS não usada	0,0293	0,1976	0,1281
Energia de Ligação de Hidro – MWM não usada	0,1337	0,3557	0,1423
Energia de Ligação de Hidro – MS não usada	0,8219	0,3385	0,0736
Energia de Ligação de Hidro – MM não usada	0,1305	0,0143	0
Energia de Contatos Hidrofóbicos não usados	0,9557	0,0016	0
phi	0	0	0
psi	0,4973	0,0176	0

### Histidina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0	0	0
Energia de Contatos Aromáticos	0	0	0
Energia de Ligação de Hidro - SWWS	0,3487	0,9062	1
Energia de Ligação de Hidro - MWWS	0,9489	0,0276	0,7331
Energia de Ligação de Hidro - MWWM	0,8771	0,3265	1
Energia de Ligação de Hidro - SWS	0,8001	0,2669	1
Energia de Ligação de Hidro - MWS	0,1228	1,00E-04	0,0397
Energia de Ligação de Hidro - MWM	0,9281	0,0075	0,6231
Energia de Ligação de Hidro - SS	0,0022	0	1,00E-04
Energia de Ligação de Hidro - MS	0,0013	0	0
Energia de Ligação de Hidro - MM	0	0,0031	0
Energia de contatos carregados repulsivos	0	0	0
Energia de contatos carregados atrativos	0	0	0
Energia de Contatos Hidrofóbicos	0	0	0

hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	1,00E-04	0
EP no LHA	0	0	0
EP na Surperficie	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,3295	0	0
Energia de Contatos Aromáticos não usados	0,003	0,5147	0,0146
Energia de Ligação de Hidro – SWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWS não usada	0,0342	0,0037	0,0222
Energia de Ligação de Hidro – MWWM não usada	0,0323	0	5,00E-04
Energia de Ligação de Hidro – SWS não usada	2,00E-04	0	0
Energia de Ligação de Hidro – MWS não usada	0,5715	0,2453	0,2204
Energia de Ligação de Hidro – MWM não usada	2,00E-04	0	0
Energia de Ligação de Hidro – SS não usada	0	5,00E-04	1,00E-04
Energia de Ligação de Hidro – MS não usada	0,0069	0	0

Energia de Ligação de Hidro – MM não usada	0,6111	0,9388	0,001
Energia de Contatos carregados repulsivos não usados	0,0282	0,6127	1,00E-04
Energia de Contatos carregados atrativos não usados	0,5881	0,0086	0
Energia de Contatos Hidrofóbicos não usados	0,0312	0	0
phi	0,0058	0,0033	1,00E-04
psi	0,0991	0,2486	0,196
chi_1	0,4947	0,6867	0,5869
chi_2	0,0392	0,0601	0,129

### Isoleucina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,0369	0	0
Energia de Ligação de Hidro - MWWS	0,5365	1,00E-04	0,589
Energia de Ligação de Hidro - MWWM	0,0247	0	0,1928
Energia de Ligação de Hidro - MWS	0,0721	0	0,1127
Energia de Ligação de Hidro - MWM	0	0	0
Energia de Ligação de Hidro - MS	0,4023	0	2,00E-04
Energia de Ligação de Hidro - MM	0	0,0011	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0

EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0,2863	9,00E-04	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0	0,1826	0
Energia de Ligação de Hidro – MWWS não usada	0,0265	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0
Energia de Ligação de Hidro – MWS não usada	0	0	0
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – MS não usada	0	0	0
Energia de Ligação de Hidro – MM não usada	0	0,0173	0
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	1,00E-04	0	0
psi	0	0	0
chi_1	0,9674	0,0429	0,0066
chi_2	0,1851	0,4966	0,3982

### Leucina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,1268	0	0
Energia de Ligação de Hidro - MWWS	0,1085	0	0,1072
Energia de Ligação de Hidro - MWWM	0,1026	0	0,0586
Energia de Ligação de Hidro - MWS	0,3435	0	0,0072
Energia de Ligação de Hidro - MWM	0	0	0
Energia de Ligação de Hidro - MS	0,5139	0	4,00E-04
Energia de Ligação de Hidro - MM	0	1,00E-04	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0

CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
Densidade no ca3	0,2266	0,2442	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0	0,7035	0
Energia de Ligação de Hidro – MWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0
Energia de Ligação de Hidro – MWS não usada	0,0024	0	0
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – MS não usada	0,6208	0,4191	0,0641
Energia de Ligação de Hidro – MM não usada	0	0	0
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	0	0	0
psi	0	0	0
chi_1	0,4382	0,2863	0,0169
chi_2	0,1073	0,0453	0,0092

Lisina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,351	0	0



Energia de Contatos Aromáticos	6,00E-04	0,309	0,4926
Energia de Ligação de Hidro - SWWS	0	0	0,7039
Energia de Ligação de Hidro - MWWS	0,003	0,0222	0,8578
Energia de Ligação de Hidro - MWWM	9,00E-04	0,2056	0,9997
Energia de Ligação de Hidro - SWS	0,0023	0,1308	1
Energia de Ligação de Hidro - MWS	0,0264	0,2137	0,9521
Energia de Ligação de Hidro - MWM	0,029	0,7319	1
Energia de Ligação de Hidro - SS	0,0823	0,8921	1
Energia de Ligação de Hidro - MS	0,1454	0	0
Energia de Ligação de Hidro - MM	0	0,4749	0
Energia de contatos carregados repulsivos	1,00E-04	0,2298	0,4382
Energia de contatos carregados atrativos	0	0	0
Energia de Contatos Hidrofóbicos	0,0338	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,6339	0,3616	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0,4801	0,9446	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0

Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0083	0,7679	0,0019
Energia de Contatos Aromáticos não usados	0,1979	0,0167	0,3046
Energia de Ligação de Hidro – SWWS não usada	0,0667	0,0029	0,0076
Energia de Ligação de Hidro – MWWS não usada	2,00E-04	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0,0011
Energia de Ligação de Hidro – SWS não usada	0,2203	0,0187	0,0438
Energia de Ligação de Hidro – MWS não usada	4,00E-04	0,001	0,0076
Energia de Ligação de Hidro – MWM não usada	0	0	0,0014
Energia de Ligação de Hidro – SS não usada	0,5401	0,0656	0,2887
Energia de Ligação de Hidro – MS não usada	0,1616	0,3565	2,00E-04
Energia de Ligação de Hidro – MM não usada	0,0112	0,8502	0,0023
Energia de Contatos carregados repulsivos não usados	0,0388	0,1349	0,0199
Energia de Contatos carregados atrativos não usados	0,0293	0	0
Energia de Contatos Hidrofóbicos não usados	0,1506	0,5285	8,00E-04
phi	0,002	0,0103	0,0838
psi	0,046	0,0283	0,0334
chi_1	0,8104	0,1265	0,0214
chi_2	0,7031	0,6398	0,041
chi_3	0,594	0,6791	0,0022
chi_4	0,3392	0,2973	1,00E-04

### Metionina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,3931	0	0
Energia de Ligação de Hidro - MWWS	0,1529	0,0028	0,8723
Energia de Ligação de Hidro - MWWM	0,0958	5,00E-04	0,7
Energia de Ligação de Hidro - MWS	0,3792	0,0141	0,8662
Energia de Ligação de Hidro - MWM	0,9427	0,0017	0,4707
Energia de Ligação de Hidro - MS	0,0539	0,0256	0,3259
Energia de Ligação de Hidro - MM	0	0,659	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0

CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,0077	0,0027	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0,0586	8,00E-04	0
EP no LHA	0,0458	0,0068	0
EP na Surperficie	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0643	0,9678	0,0014
Energia de Ligação de Hidro – MWWS não usada	0,5856	0,0128	0,0031
Energia de Ligação de Hidro – MWWM não usada	0,4993	0,0521	0,0151
Energia de Ligação de Hidro – MWS não usada	0,1812	0,0449	0,022
Energia de Ligação de Hidro – MWM não usada	0,0062	0,0786	0,0142
Energia de Ligação de Hidro – MS não usada	0,3397	0,1807	0,0375
Energia de Ligação de Hidro – MM não usada	0,0052	0,1537	0,0024
Energia de Contatos Hidrofóbicos não usados	7,00E-04	0	0
phi	0,0035	0,0079	0,024
psi	4,00E-04	6,00E-04	0,0017
chi_1	0,7032	0,0329	0,0165

chi_2	0,384	0,3249	0,5282
chi_3	0,6712	0,8678	0,8695

### Fenilalanina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0	0	0
Energia de Contatos Aromáticos	0	0	0
Energia de Ligação de Hidro - MWWS	0,9207	0,0026	0,8094
Energia de Ligação de Hidro - MWWM	0,0894	0,0163	0,9272
Energia de Ligação de Hidro - MWS	0,7552	2,00E-04	0,3109
Energia de Ligação de Hidro - MWM	0,8222	0	0,0762
Energia de Ligação de Hidro - MS	0,0288	0,0118	0,2096
Energia de Ligação de Hidro - MM	0	0,0232	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no C-beta	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,0455	0,0016	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0

Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0	0	0
Energia de Contatos Aromáticos não usados	0,011	8,00E-04	0
Energia de Ligação de Hidro – MWWS não usada	1,00E-04	0	0
Energia de Ligação de Hidro – MWWM não usada	0	0	0
Energia de Ligação de Hidro – MWS não usada	0	0	0
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – MS não usada	0	0	0
Energia de Ligação de Hidro – MM não usada	0	0,3683	0
Energia de Contatos Hidrofóbicos não usados	0,3346	0	0
phi	0	0	0
psi	0	0	0
chi_1	0,5841	0,429	0,1435
chi_2	0,7081	0,4457	0,1018

#### Prolina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,195	0	0
Energia de Ligação de Hidro - MWWS	0,7623	0,1626	1
Energia de Ligação de Hidro - MWWM	0,4153	8,00E-04	0,8985
Energia de Ligação de Hidro - MWS	0,01	0	0,4115
Energia de Ligação de Hidro - MWM	0,4394	7,00E-04	0,643
Energia de Ligação de Hidro - MS	0,0091	0	4,00E-04
Energia de Ligação de Hidro - MM	1,00E-04	0	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0

CLO no LHA	0	0	0
Densidade no ca3	0,0204	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0,2309	0,2758	0,0015
EP no LHA	4,00E-04	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	3,00E-04	0,622	0
Energia de Ligação de Hidro – MWWS não usada	0,0034	0	3,00E-04
Energia de Ligação de Hidro – MWWM não usada	1,00E-04	1,00E-04	9,00E-04
Energia de Ligação de Hidro – MWS não usada	0,0437	3,00E-04	0,001
Energia de Ligação de Hidro – MWM não usada	2,00E-04	5,00E-04	0,0017
Energia de Ligação de Hidro – MS não usada	0,0268	0,112	0,0046
Energia de Ligação de Hidro – MM não usada	0,0914	0,3363	0,0022
Energia de Contatos Hidrofóbicos não usados	0,659	0	0
phi	0,0819	3,00E-04	0
psi	0,3878	0,0288	0,0645

Serina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,1972	0	0
Energia de Ligação de Hidro - SWWS	0,8843	0,0185	0,9238
Energia de Ligação de Hidro - MWWS	0,0091	0	0,0037
Energia de Ligação de Hidro - MWWM	0,8628	0,0019	0,8312
Energia de Ligação de Hidro - SWS	0,8561	2,00E-04	0,2457
Energia de Ligação de Hidro - MWS	0,0039	0	0
Energia de Ligação de Hidro - MWM	0,7091	0	0,0624

Energia de Ligação de Hidro - SS	0	0	0
Energia de Ligação de Hidro - MS	0,5475	0	0
Energia de Ligação de Hidro - MM	0	0	0
Energia de contatos carregados repulsivos	0,5676	0,6947	1
Energia de contatos carregados atrativos	0,683	0,038	1
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha4	0	0	0
CED no lha5	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
Densidade no ca3	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0,0033	0,0271	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	2,00E-04	0,2924	0
Energia de Ligação de Hidro – SWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWS não usada	0	0,0015	1,00E-04
Energia de Ligação de Hidro – MWWM não usada	0	0	1,00E-04

Energia de Ligação de Hidro – SWS não usada	0	0	3,00E-04
Energia de Ligação de Hidro – MWS não usada	4,00E-04	0,0118	3,00E-04
Energia de Ligação de Hidro – MWM não usada	7,00E-04	0,001	9,00E-04
Energia de Ligação de Hidro – SS não usada	0,0202	0,215	3,00E-04
Energia de Ligação de Hidro – MS não usada	0,0173	0,3258	1,00E-04
Energia de Ligação de Hidro – MM não usada	0,0278	0,5922	1,00E-04
Energia de Contatos carregados repulsivos não usados	0,1838	0,2419	1
Energia de Contatos carregados atrativos não usados	0,117	0,2441	1
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	0,7673	0,8154	0,0276
psi	0,4778	0,0592	0,0175
chi_1	0	0	0

#### Treonina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,0093	0	0
Energia de Ligação de Hidro - SWWS	0,6993	5,00E-04	0,4646
Energia de Ligação de Hidro - MWWS	0,0052	0	0
Energia de Ligação de Hidro - MWWM	0,3003	0,1292	1
Energia de Ligação de Hidro - SWS	0,6824	0	0,0443
Energia de Ligação de Hidro - MWS	0,3154	0	0
Energia de Ligação de Hidro - MWM	0,1672	0	0,162
Energia de Ligação de Hidro - SS	0,0121	0	0
Energia de Ligação de Hidro - MS	0,0026	0	0
Energia de Ligação de Hidro - MM	0	0,0258	0
Energia de contatos carregados repulsivos	0,5589	0,6128	1
Energia de contatos carregados atrativos	0,658	0,6742	1
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0



CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,0292	0	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,0321	0,0251	0
Energia de Ligação de Hidro – SWWS não usada	0,7101	0,1976	0,9275
Energia de Ligação de Hidro – MWWS não usada	0,021	0,1031	0,0053
Energia de Ligação de Hidro – MWWM não usada	3,00E-04	0,0028	0,0013
Energia de Ligação de Hidro – SWS não usada	0,8661	0,1189	0,473
Energia de Ligação de Hidro – MWS não usada	3,00E-04	7,00E-04	1,00E-04
Energia de Ligação de Hidro – MWM não usada	0,0204	0,0215	0,0104
Energia de Ligação de Hidro – SS não usada	0,4584	0,8756	0,0135
Energia de Ligação de Hidro – MS não usada	0,1101	0,2603	2,00E-04
Energia de Ligação de Hidro – MM não usada	0,0449	0,4265	0,0047
Energia de Contatos carregados repulsivos não usados	0,7713	0,9385	1
Energia de Contatos carregados atrativos não usados	0,6588	0,9331	1
Energia de Contatos Hidrofóbicos não usados	0	0	0

phi	0,0245	0,0021	0
psi	1,00E-04	0	0
chi_1	0,0102	0,0022	0,0068

### Triptofano:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,061	0	0
Energia de Contatos Aromáticos	1,00E-04	0	0
Energia de Ligação de Hidro - SWWS	0,0107	0,0123	0,9578
Energia de Ligação de Hidro - MWWS	0,3283	0,0724	0,9246
Energia de Ligação de Hidro - MWWM	0,1682	0,5006	1
Energia de Ligação de Hidro - SWS	0,5068	0,0925	0,996
Energia de Ligação de Hidro - MWS	0,9266	0,0038	0,223
Energia de Ligação de Hidro - MWM	0,6978	0,1974	0,9988
Energia de Ligação de Hidro - SS	0,0981	0	1,00E-04
Energia de Ligação de Hidro - MS	0,0559	0	0
Energia de Ligação de Hidro - MM	0	0,0968	0
Energia de contatos carregados repulsivos	0,2872	0,5101	1
Energia de contatos carregados atrativos	0,9291	0,6708	1
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no C-beta	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0,8039	0,7	0
Densidade no ca4	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0,0017	0	0
Densidade no lha5	0	0	0

Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Surperficie	0	0	0
Esponja no ca3	9,00E-04	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0,312	0,004	0
Energia de Contatos Aromáticos não usados	0,7393	0,0444	1,00E-04
Energia de Ligação de Hidro – SWWS não usada	0,0433	0,061	0,0706
Energia de Ligação de Hidro – MWWS não usada	0,0492	0,0451	0,0667
Energia de Ligação de Hidro – MWWM não usada	0,4403	0,0417	0,1545
Energia de Ligação de Hidro – SWS não usada	0,4367	0,216	0,6803
Energia de Ligação de Hidro – MWS não usada	0,2585	0,3645	0,2827
Energia de Ligação de Hidro – MWM não usada	0,1726	0,1273	0,2701
Energia de Ligação de Hidro – SS não usada	0,1956	0,0993	0,1163
Energia de Ligação de Hidro – MS não usada	0,5159	0,0555	2,00E-04
Energia de Ligação de Hidro – MM não usada	0,357	0,3399	0,061
Energia de Contatos carregados repulsivos não usados	0,6664	0,7067	1
Energia de Contatos carregados atrativos não usados	0,7332	0,7099	1
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	0,3873	0,9235	0,6656
psi	0,0046	0,1159	2,00E-04
chi_1	0,499	0,1939	0,1037
chi_2	0,1035	0,3593	0,0244

#### Tirosina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,0049	0	0
Energia de Contatos Aromáticos	0	0	0
Energia de Ligação de Hidro - SWWS	0,306	5,00E-04	0,3843
Energia de Ligação de Hidro - MWWS	3,00E-04	0	0

Energia de Ligação de Hidro - MWWM	0,0229	0,9515	1
Energia de Ligação de Hidro - SWS	0,0068	0	0,002
Energia de Ligação de Hidro - MWS	0	0	0
Energia de Ligação de Hidro - MWM	0,086	0,0154	0,7432
Energia de Ligação de Hidro - SS	0	0	0
Energia de Ligação de Hidro - MS	0	0	0
Energia de Ligação de Hidro - MM	0	3,00E-04	0
Energia de contatos carregados repulsivos	0,2479	0,8299	1
Energia de contatos carregados atrativos	0,2349	0,4675	1
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha4	0	0	0
CED no lha7	0	0	0
CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	4,00E-04	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Surperficie	0	0	0
Esponja no ca3	0	0	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0	0	0
Energia de Contatos Aromáticos não usados	0,0742	0,4385	0

Energia de Ligação de Hidro – SWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWS não usada	0,0067	0	0
Energia de Ligação de Hidro – MWWM não usada	8,00E-04	0	0
Energia de Ligação de Hidro – SWS não usada	0,0629	3,00E-04	0
Energia de Ligação de Hidro – MWS não usada	0	0,0039	0
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – SS não usada	0,1663	0,0815	0
Energia de Ligação de Hidro – MS não usada	0,0138	0,2297	0
Energia de Ligação de Hidro – MM não usada	9,00E-04	0,8137	0
Energia de Contatos carregados repulsivos não usados	0,4551	0,2579	1
Energia de Contatos carregados atrativos não usados	0,633	0,2589	1
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	0	0	0
psi	0	2,00E-04	0
chi_1	0,2707	0,259	0,072
chi_2	0,9969	0,5217	0,0417

### Valina:

Descritor	Welch	Wilcoxon-Mann-Whitney	Kolmogorov-Smirnov
Energia Total de Contatos	0,7598	0	0
Energia de Ligação de Hidro - MWWS	0,1038	0	0,2565
Energia de Ligação de Hidro - MWWM	6,00E-04	0	0,1162
Energia de Ligação de Hidro - MWS	0,0815	0	0,0594
Energia de Ligação de Hidro - MWM	1,00E-04	0	0,003
Energia de Ligação de Hidro - MS	0,9739	0	0,0036
Energia de Ligação de Hidro - MM	0	0,0096	0
Energia de Contatos Hidrofóbicos	0	0	0
hydrophobicity	0	0	0
CED no ca3	0	0	0
CED no ca4	0	0	0
CED no ca5	0	0	0
CED no ca7	0	0	0
CED no lha3	0	0	0
CED no lha7	0	0	0

CPO no C-alpha	0	0	0
CPO no C-beta	0	0	0
CPO no LHA	0	0	0
CLO no C-alpha	0	0	0
CLO no LHA	0	0	0
Densidade no ca3	0	0	0
Densidade no ca7	0	0	0
Densidade no lha3	0	0	0
Densidade no lha4	0	0	0
Densidade no lha5	0	0	0
Densidade no lha7	0	0	0
EP no Ca	0	0	0
EP no LHA	0	0	0
EP na Superfície	0	0	0
Esponja no ca3	0,3076	0,0032	0
Esponja no ca7	0	0	0
Esponja no lha3	0	0	0
Esponja no lha4	0	0	0
Esponja no lha7	0	0	0
Energia Total de Contatos Não-usados	0	0,4305	0
Energia de Ligação de Hidro – MWWS não usada	0	0	0
Energia de Ligação de Hidro – MWWM não usada	1,00E-04	4,00E-04	0,0031
Energia de Ligação de Hidro – MWS não usada	0,0124	0	0
Energia de Ligação de Hidro – MWM não usada	0	0	0
Energia de Ligação de Hidro – MS não usada	0	0	0
Energia de Ligação de Hidro – MM não usada	0,0045	0,0503	0
Energia de Contatos Hidrofóbicos não usados	0	0	0
phi	0	0	0
psi	0	0	0
chi_1	0,0038	0,0727	0,0051

## 6.2 Apêndice 2 – Resultado dos testes estatísticos de MANOVA

Resultado dos testes estatísticos de MANOVA para cada tipo de aminoácido, como mostrado na saída do pacote HSAUR2.0 do software R.

### Arginina:

```
> summary(arg_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr    1 0.32548   410.64    64 54464 < 2.2e-16 ***
Residuals 54527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(arg_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr    1 0.67452   410.64    64 54464 < 2.2e-16 ***
Residuals 54527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(arg_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr    1      0.48254   410.64    64 54464 < 2.2e-16 ***
Residuals 54527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(arg_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr    1 0.48254   410.64    64 54464 < 2.2e-16 ***
Residuals 54527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

### Asparagina:

```
> summary(asn_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr    1 0.34108    339    64 41914 < 2.2e-16 ***
Residuals 41977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asn_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr    1 0.65892    339    64 41914 < 2.2e-16 ***
Residuals 41977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asn_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr    1      0.51763    339    64 41914 < 2.2e-16 ***
Residuals 41977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asn_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr    1 0.51763    339    64 41914 < 2.2e-16 ***
Residuals 41977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Ácido aspártico:

```
> summary(asp_manova, test = "Pillai")
      Df Pillai approx F num Df den Df    Pr(>F)
ifr      1 0.55977  1408.7     54 59827 < 2.2e-16 ***
Residuals 59880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asp_manova, test = "Wilks")
      Df Wilks approx F num Df den Df    Pr(>F)
ifr      1 0.44023  1408.7     54 59827 < 2.2e-16 ***
Residuals 59880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asp_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
ifr      1          1.2715  1408.7     54 59827 < 2.2e-16 ***
Residuals 59880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(asp_manova, test = "Roy")
      Df Roy approx F num Df den Df    Pr(>F)
ifr      1 1.2715  1408.7     54 59827 < 2.2e-16 ***
Residuals 59880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Cisteína:

```
> summary(cys_manova, test = "Pillai")
      Df Pillai approx F num Df den Df    Pr(>F)
ifr      1 0.43336  171.06     47 10512 < 2.2e-16 ***
Residuals 10558
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(cys_manova, test = "Wilks")
      Df Wilks approx F num Df den Df    Pr(>F)
ifr      1 0.56664  171.06     47 10512 < 2.2e-16 ***
Residuals 10558
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(cys_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
ifr      1          0.7648  171.06     47 10512 < 2.2e-16 ***
Residuals 10558
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(cys_manova, test = "Roy")
      Df Roy approx F num Df den Df    Pr(>F)
ifr      1 0.7648  171.06     47 10512 < 2.2e-16 ***
Residuals 10558
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.



## Glutamina:

```
> summary(gln_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.32018   291.26     62 38342 < 2.2e-16 ***
Residuals 38403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gln_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.67982   291.26     62 38342 < 2.2e-16 ***
Residuals 38403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gln_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1           0.47098   291.26     62 38342 < 2.2e-16 ***
Residuals 38403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gln_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.47098   291.26     62 38342 < 2.2e-16 ***
Residuals 38403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Ácido Glutâmico:

```
> summary(glu_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.30574   547.56     59 73358 < 2.2e-16 ***
Residuals 73416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(glu_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.69426   547.56     59 73358 < 2.2e-16 ***
Residuals 73416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(glu_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1           0.44039   547.56     59 73358 < 2.2e-16 ***
Residuals 73416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(glu_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.44039   547.56     59 73358 < 2.2e-16 ***
Residuals 73416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Glicina:

```
> summary(gly_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.11151   252.72    32 64434 < 2.2e-16 ***
Residuals 64465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gly_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.88849   252.72    32 64434 < 2.2e-16 ***
Residuals 64465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gly_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1           0.12551   252.72    32 64434 < 2.2e-16 ***
Residuals 64465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(gly_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.12551   252.72    32 64434 < 2.2e-16 ***
Residuals 64465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Histidina:

```
> summary(his_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.37949   261.23    57 24347 < 2.2e-16 ***
Residuals 24403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(his_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.62051   261.23    57 24347 < 2.2e-16 ***
Residuals 24403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(his_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1           0.61158   261.23    57 24347 < 2.2e-16 ***
Residuals 24403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(his_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.61158   261.23    57 24347 < 2.2e-16 ***
Residuals 24403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Isoleucina:

```
> summary(ile_manova, test = "Pillai")
      Df Pillai approx F num Df den Df Pr(>F)
ifr      1 0.45445  881.43    46 48674 < 2.2e-16 ***
Residuals 48719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ile_manova, test = "Wilks")
      Df Wilks approx F num Df den Df Pr(>F)
ifr      1 0.54555  881.43    46 48674 < 2.2e-16 ***
Residuals 48719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ile_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
ifr      1          0.83301  881.43    46 48674 < 2.2e-16 ***
Residuals 48719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ile_manova, test = "Roy")
      Df Roy approx F num Df den Df Pr(>F)
ifr      1 0.83301  881.43    46 48674 < 2.2e-16 ***
Residuals 48719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Leucina:

```
> summary(leu_manova, test = "Pillai")
      Df Pillai approx F num Df den Df Pr(>F)
ifr      1 0.44597 1470.1    45 82184 < 2.2e-16 ***
Residuals 82228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(leu_manova, test = "Wilks")
      Df Wilks approx F num Df den Df Pr(>F)
ifr      1 0.55403 1470.1    45 82184 < 2.2e-16 ***
Residuals 82228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(leu_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
ifr      1          0.80496 1470.1    45 82184 < 2.2e-16 ***
Residuals 82228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(leu_manova, test = "Roy")
      Df Roy approx F num Df den Df Pr(>F)
ifr      1 0.80496 1470.1    45 82184 < 2.2e-16 ***
Residuals 82228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Lisina:

```
> summary(lys_manova, test = "Pillai")
      Df Pillai approx F num Df den Df Pr(>F)
ifr    1 0.24986  332.75    60 59939 < 2.2e-16 ***
Residuals 59998
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lys_manova, test = "Wilks")
      Df Wilks approx F num Df den Df Pr(>F)
ifr    1 0.75014  332.75    60 59939 < 2.2e-16 ***
Residuals 59998
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lys_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
ifr    1          0.33309  332.75    60 59939 < 2.2e-16 ***
Residuals 59998
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lys_manova, test = "Roy")
      Df Roy approx F num Df den Df Pr(>F)
ifr    1 0.33309  332.75    60 59939 < 2.2e-16 ***
Residuals 59998
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Metionina:

```
> summary(met_manova, test = "Pillai")
      Df Pillai approx F num Df den Df Pr(>F)
ifr    1 0.41858  233.13    47 15220 < 2.2e-16 ***
Residuals 15266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(met_manova, test = "Wilks")
      Df Wilks approx F num Df den Df Pr(>F)
ifr    1 0.58142  233.13    47 15220 < 2.2e-16 ***
Residuals 15266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(met_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
ifr    1          0.71992  233.13    47 15220 < 2.2e-16 ***
Residuals 15266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(met_manova, test = "Roy")
      Df Roy approx F num Df den Df Pr(>F)
ifr    1 0.71992  233.13    47 15220 < 2.2e-16 ***
Residuals 15266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Fenilalanina:

```
> summary(phe_manova, test = "Pillai")
      Df Pillai approx F num Df den Df    Pr(>F)
ifr      1 0.46759   667.19    48 36465 < 2.2e-16 ***
Residuals 36512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(phe_manova, test = "Wilks")
      Df Wilks approx F num Df den Df    Pr(>F)
ifr      1 0.53241   667.19    48 36465 < 2.2e-16 ***
Residuals 36512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(phe_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
ifr      1           0.87824   667.19    48 36465 < 2.2e-16 ***
Residuals 36512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(phe_manova, test = "Roy")
      Df Roy approx F num Df den Df    Pr(>F)
ifr      1 0.87824   667.19    48 36465 < 2.2e-16 ***
Residuals 36512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Prolina:

```
> summary(pro_manova, test = "Pillai")
      Df Pillai approx F num Df den Df    Pr(>F)
ifr      1 0.40834   715.07    44 45589 < 2.2e-16 ***
Residuals 45632
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(pro_manova, test = "Wilks")
      Df Wilks approx F num Df den Df    Pr(>F)
ifr      1 0.59166   715.07    44 45589 < 2.2e-16 ***
Residuals 45632
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(pro_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
ifr      1           0.69015   715.07    44 45589 < 2.2e-16 ***
Residuals 45632
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(pro_manova, test = "Roy")
      Df Roy approx F num Df den Df    Pr(>F)
ifr      1 0.69015   715.07    44 45589 < 2.2e-16 ***
Residuals 45632
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Serina:

```
> summary(ser_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.36764   582.23     56 56081 < 2.2e-16 ***
Residuals 56136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ser_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.63236   582.23     56 56081 < 2.2e-16 ***
Residuals 56136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ser_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1              0.58139   582.23     56 56081 < 2.2e-16 ***
Residuals 56136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ser_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.58139   582.23     56 56081 < 2.2e-16 ***
Residuals 56136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Treonina:

```
> summary(thr_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr      1 0.41685   660.42     56 51738 < 2.2e-16 ***
Residuals 51793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(thr_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr      1 0.58315   660.42     56 51738 < 2.2e-16 ***
Residuals 51793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(thr_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr      1              0.71483   660.42     56 51738 < 2.2e-16 ***
Residuals 51793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(thr_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr      1 0.71483   660.42     56 51738 < 2.2e-16 ***
Residuals 51793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Triptofano:

```
> summary(trp_manova, test = "Pillai")
      Df Pillai approx F num Df den Df      Pr(>F)
ifr      1 0.6252   351.92     60 12658 < 2.2e-16 ***
Residuals 12717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(trp_manova, test = "Wilks")
      Df Wilks approx F num Df den Df      Pr(>F)
ifr      1 0.3748   351.92     60 12658 < 2.2e-16 ***
Residuals 12717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(trp_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df      Pr(>F)
ifr      1             1.6681   351.92     60 12658 < 2.2e-16 ***
Residuals 12717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(trp_manova, test = "Roy")
      Df      Roy approx F num Df den Df      Pr(>F)
ifr      1 1.6681   351.92     60 12658 < 2.2e-16 ***
Residuals 12717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Tirosina:

```
> summary(tyr_manova, test = "Pillai")
      Df Pillai approx F num Df den Df      Pr(>F)
ifr      1 0.40049   399.23     57 34065 < 2.2e-16 ***
Residuals 34121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(tyr_manova, test = "Wilks")
      Df Wilks approx F num Df den Df      Pr(>F)
ifr      1 0.59951   399.23     57 34065 < 2.2e-16 ***
Residuals 34121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(tyr_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df      Pr(>F)
ifr      1             0.66802   399.23     57 34065 < 2.2e-16 ***
Residuals 34121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(tyr_manova, test = "Roy")
      Df      Roy approx F num Df den Df      Pr(>F)
ifr      1 0.66802   399.23     57 34065 < 2.2e-16 ***
Residuals 34121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

## Valina:

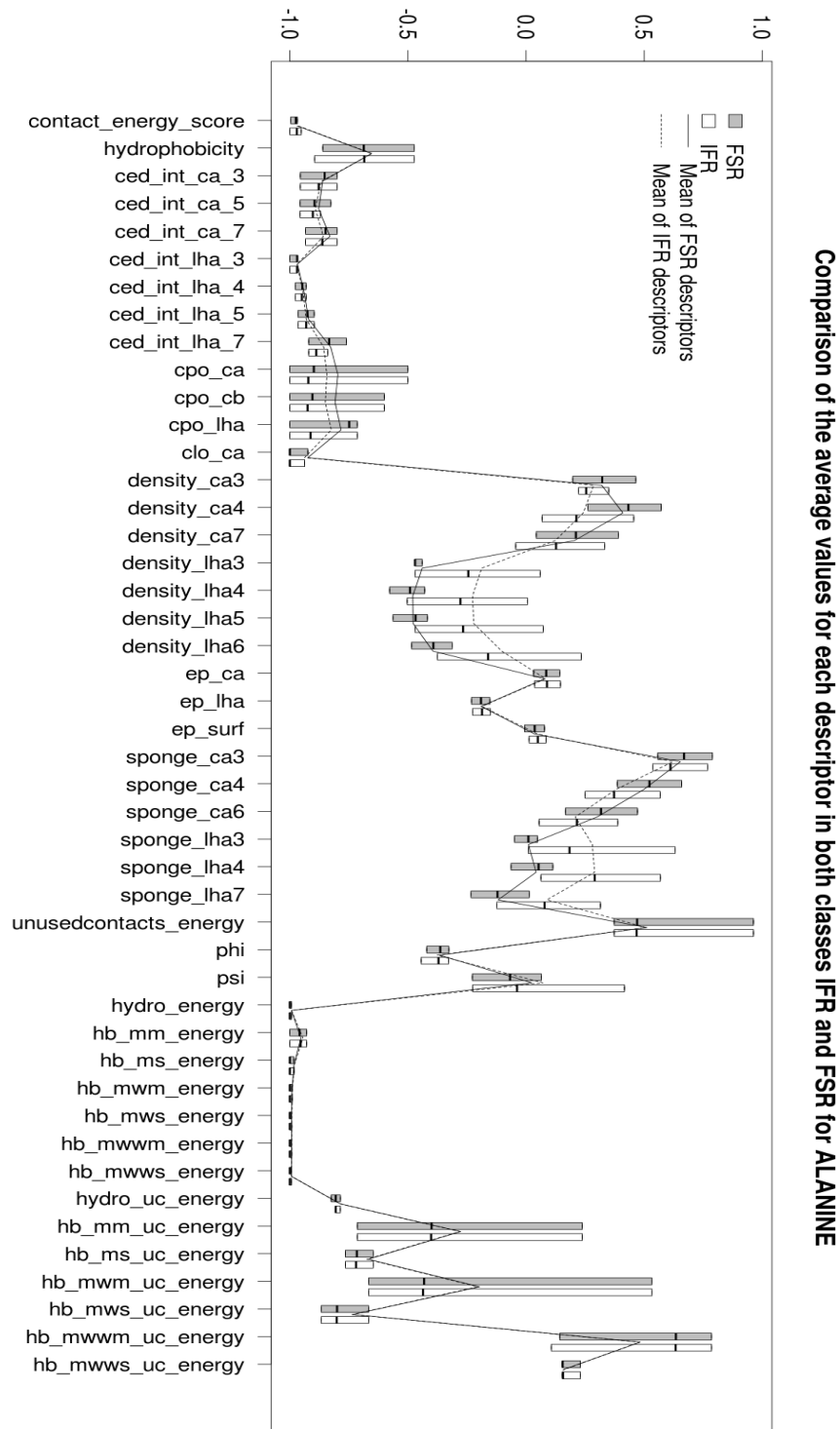
```
> summary(val_manova, test = "Pillai")
      Df Pillai approx F num Df den Df   Pr(>F)
ifr    1 0.44514  1036.4    45 58135 < 2.2e-16 ***
Residuals 58179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(val_manova, test = "Wilks")
      Df Wilks approx F num Df den Df   Pr(>F)
ifr    1 0.55486  1036.4    45 58135 < 2.2e-16 ***
Residuals 58179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(val_manova, test = "Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
ifr    1          0.80225  1036.4    45 58135 < 2.2e-16 ***
Residuals 58179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(val_manova, test = "Roy")
      Df Roy approx F num Df den Df   Pr(>F)
ifr    1 0.80225  1036.4    45 58135 < 2.2e-16 ***
Residuals 58179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comando do pacote HSAUR 2.0 do software R.

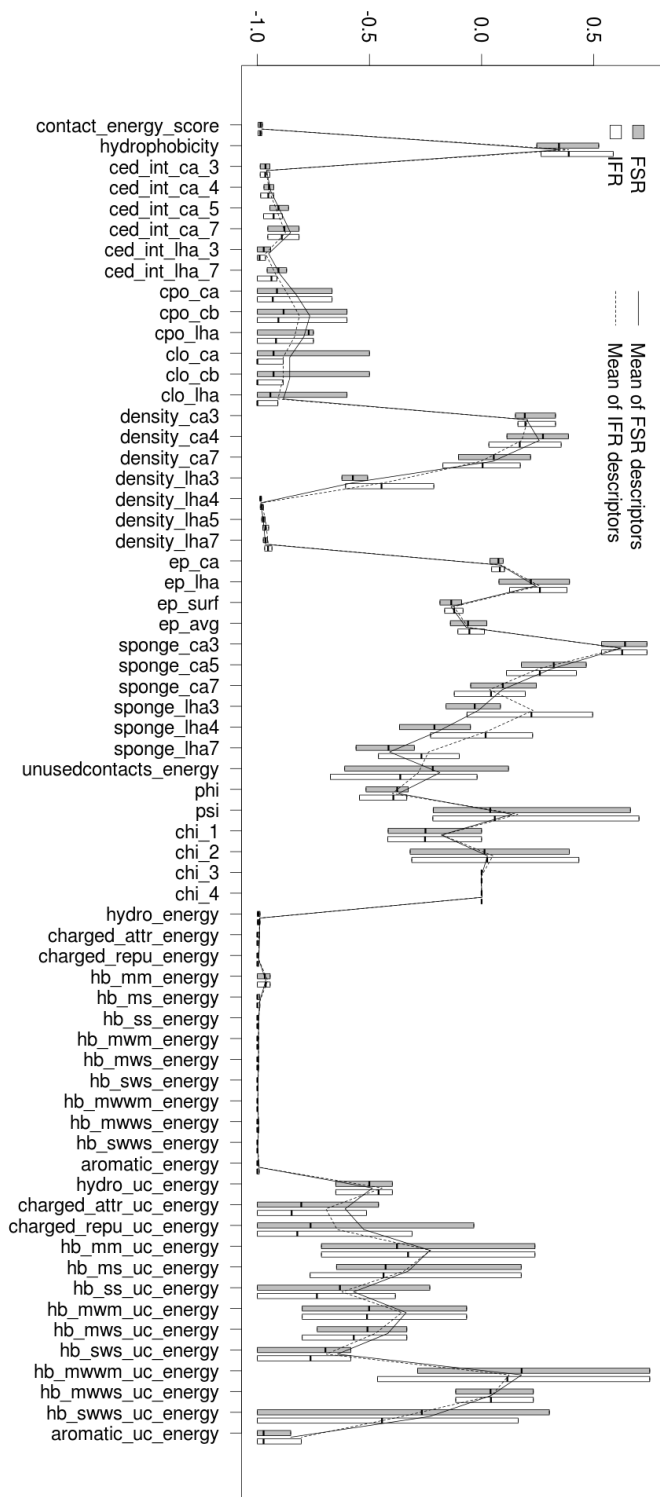


### 6.3 Apêndice 3 – Gráficos do tipo boxplot para todos os descritores e organizada por tipo de aminoácido.

Alanina:

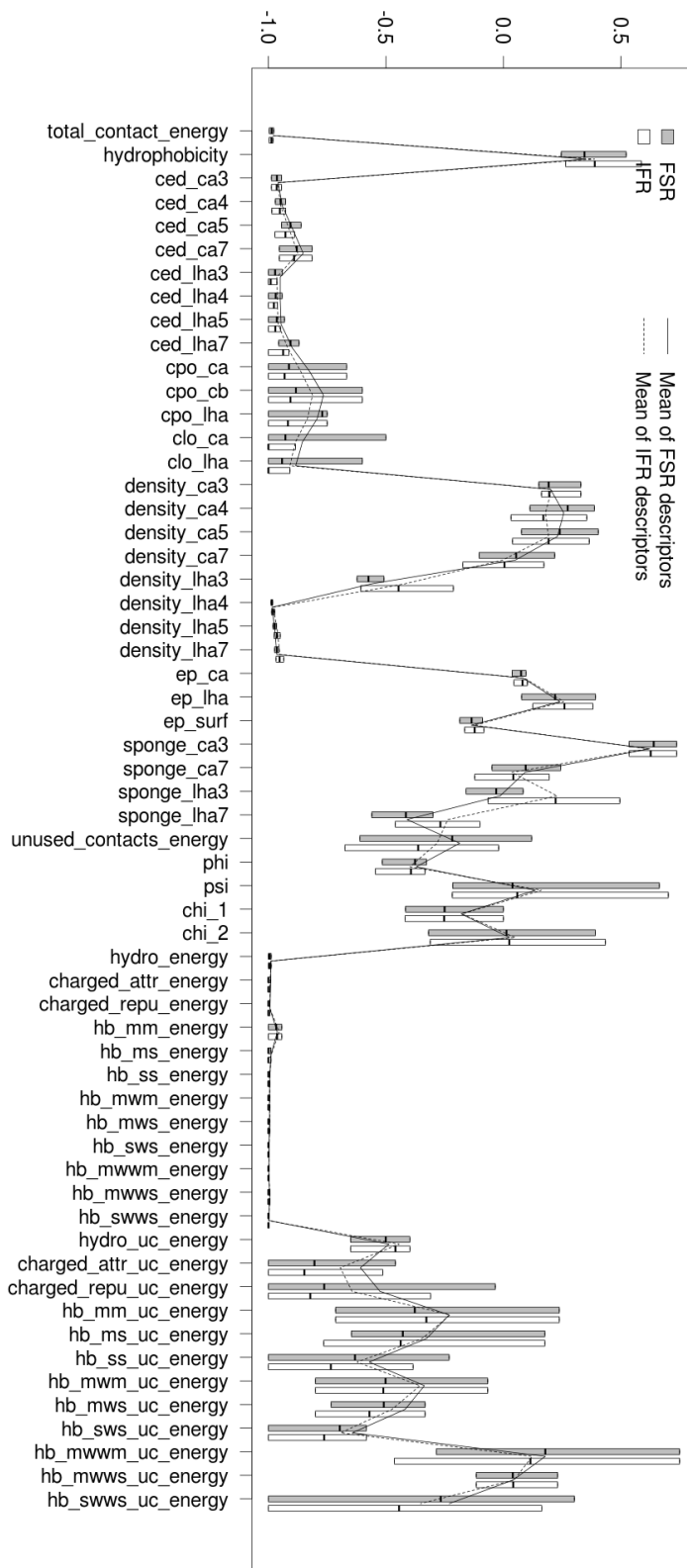


Arginina:

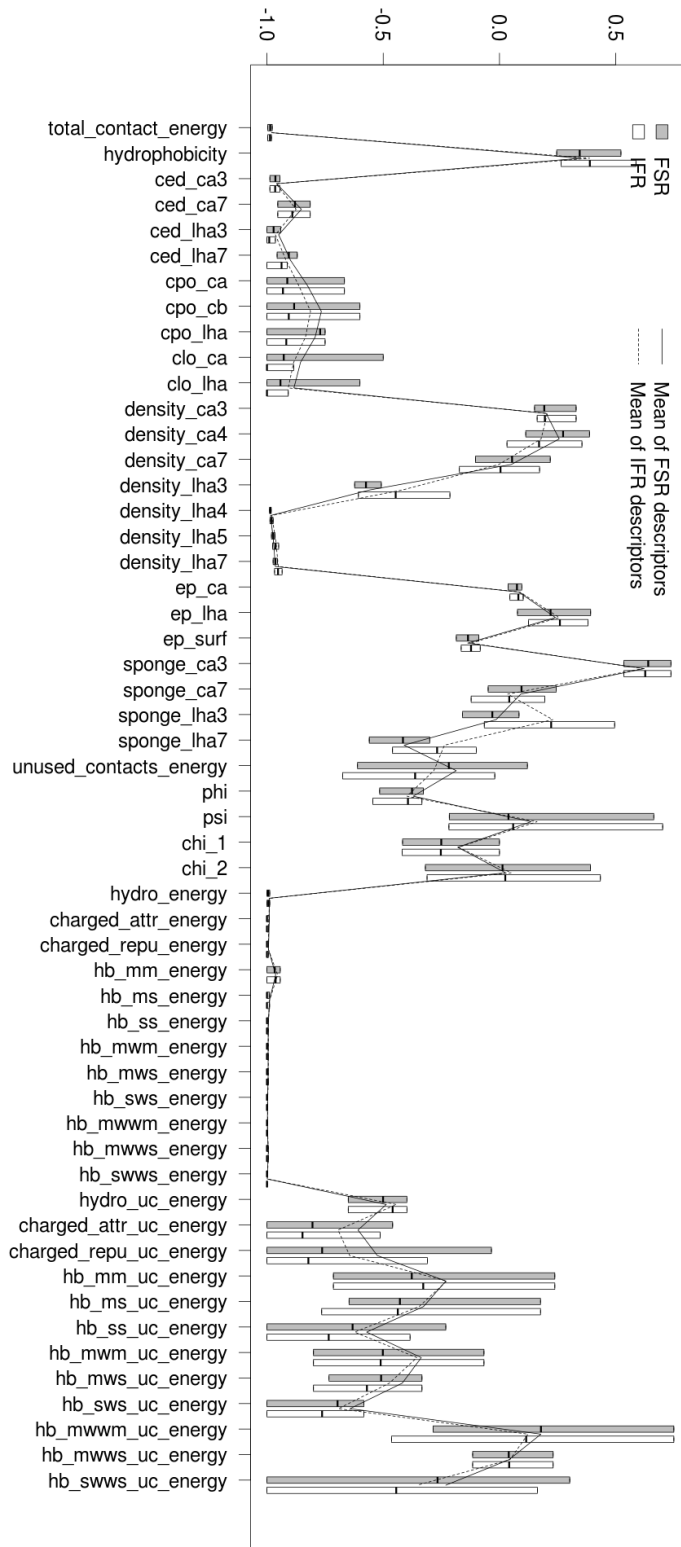


Comparison of the average values for each descriptor in both classes IFR and FSR for ARGININE

Asparagina:

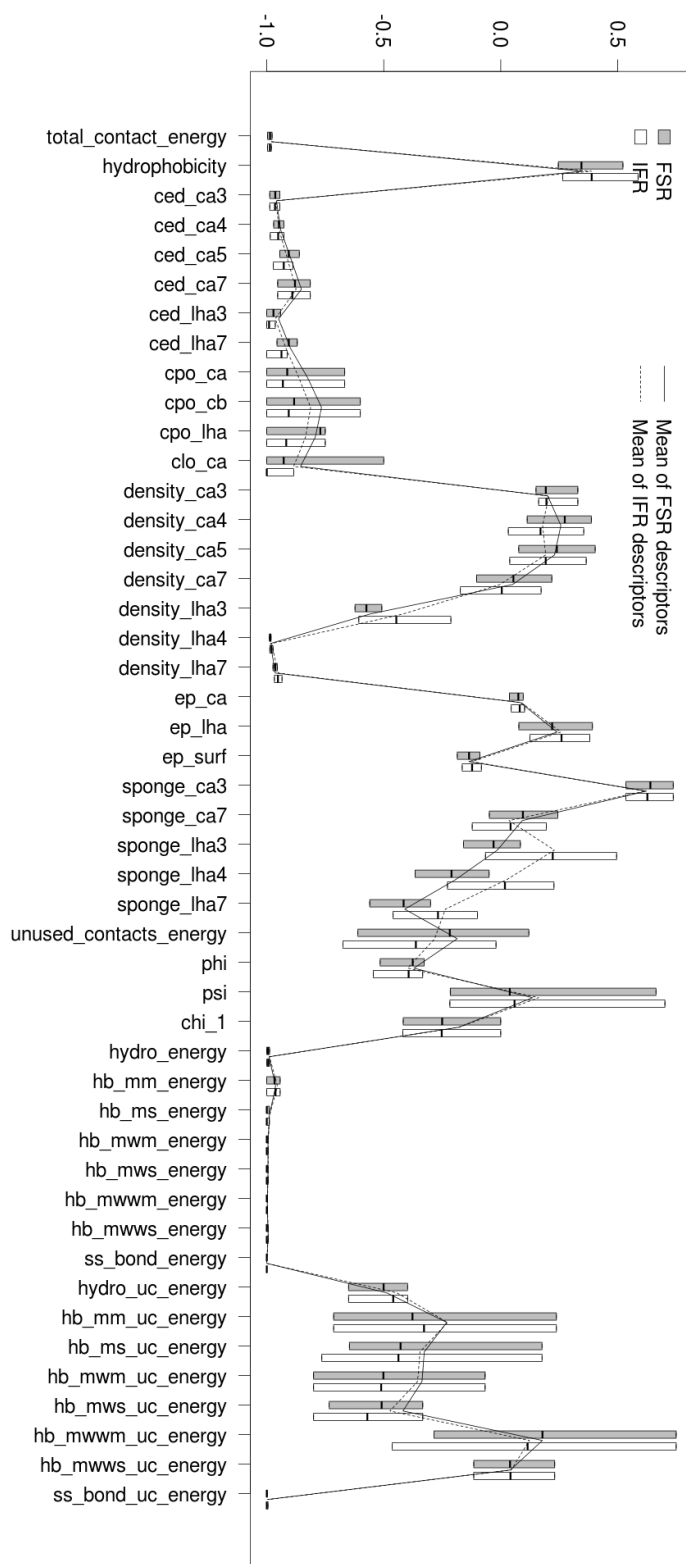


Ácido aspártico:

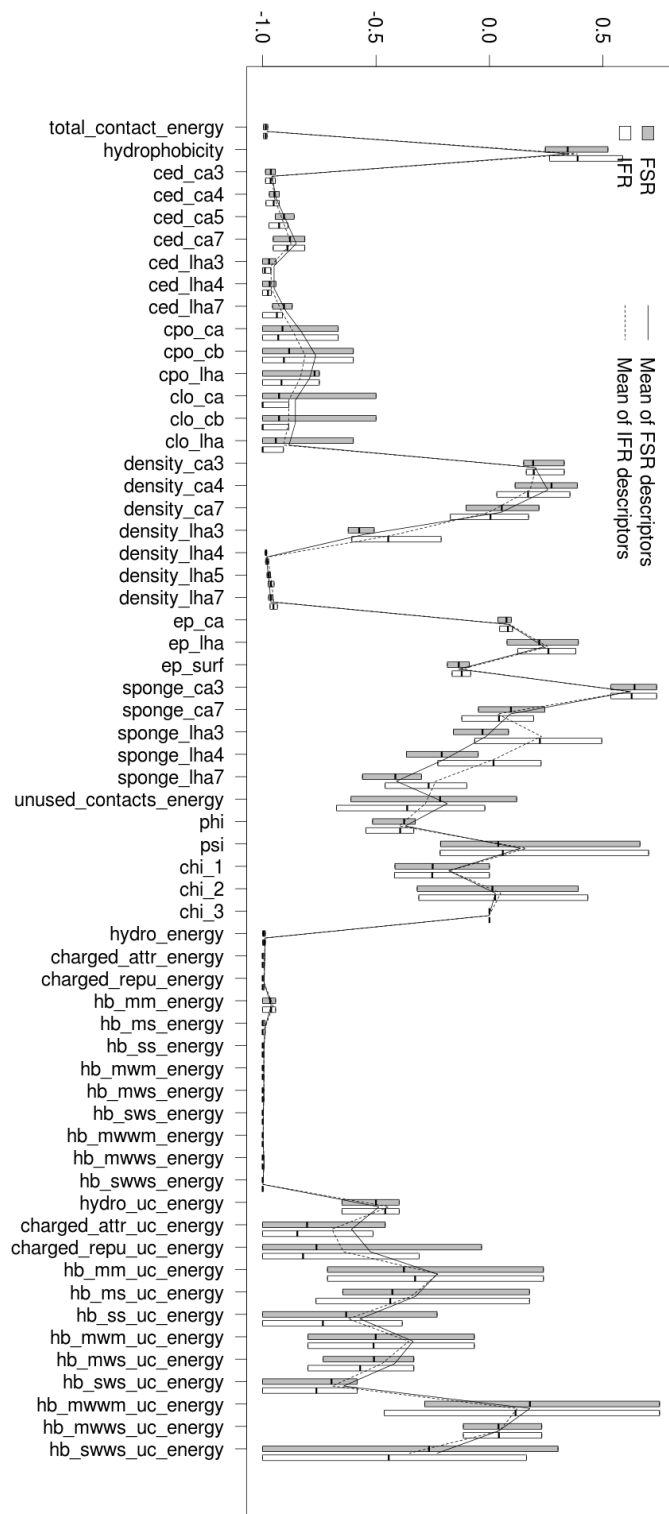


Comparison of the average values for each descriptor in both classes IFR and FSR for ASPARTIC ACID

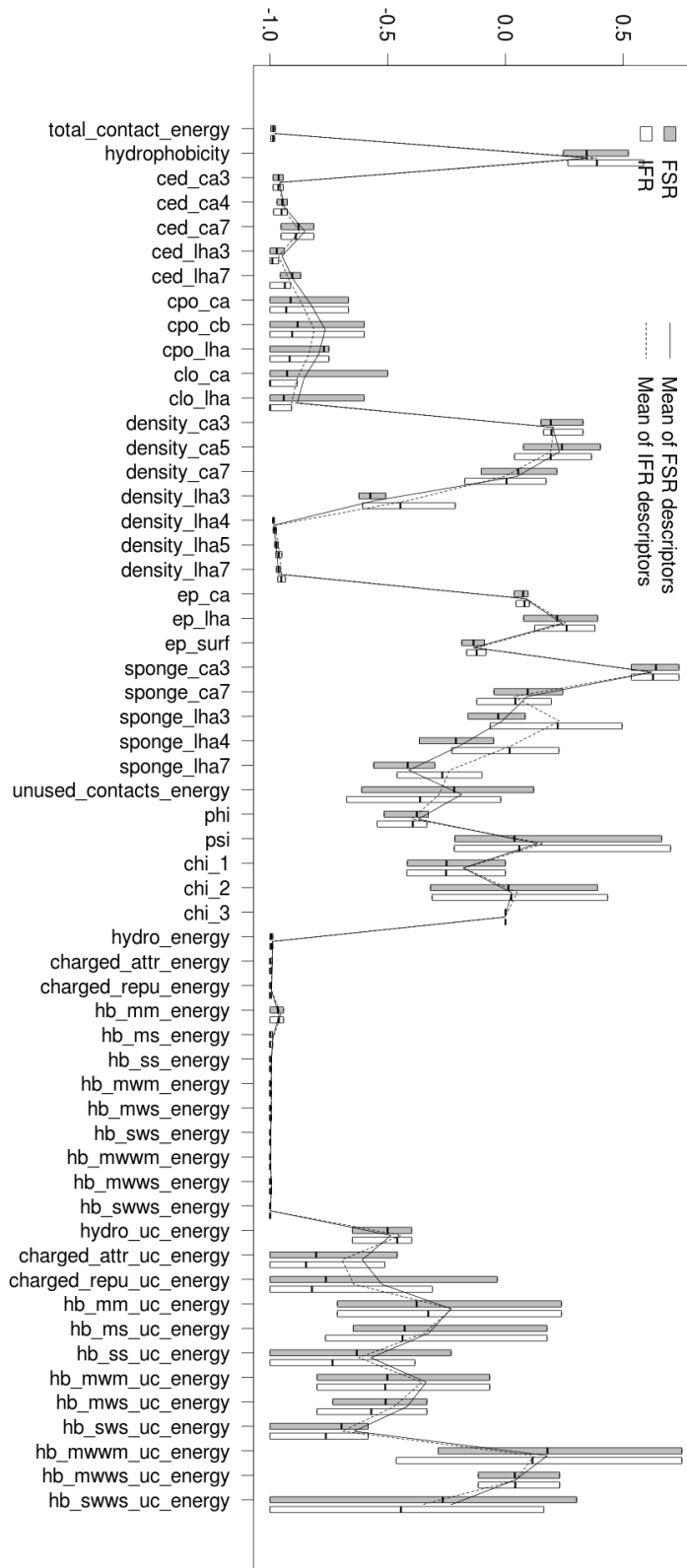
Cisteína:



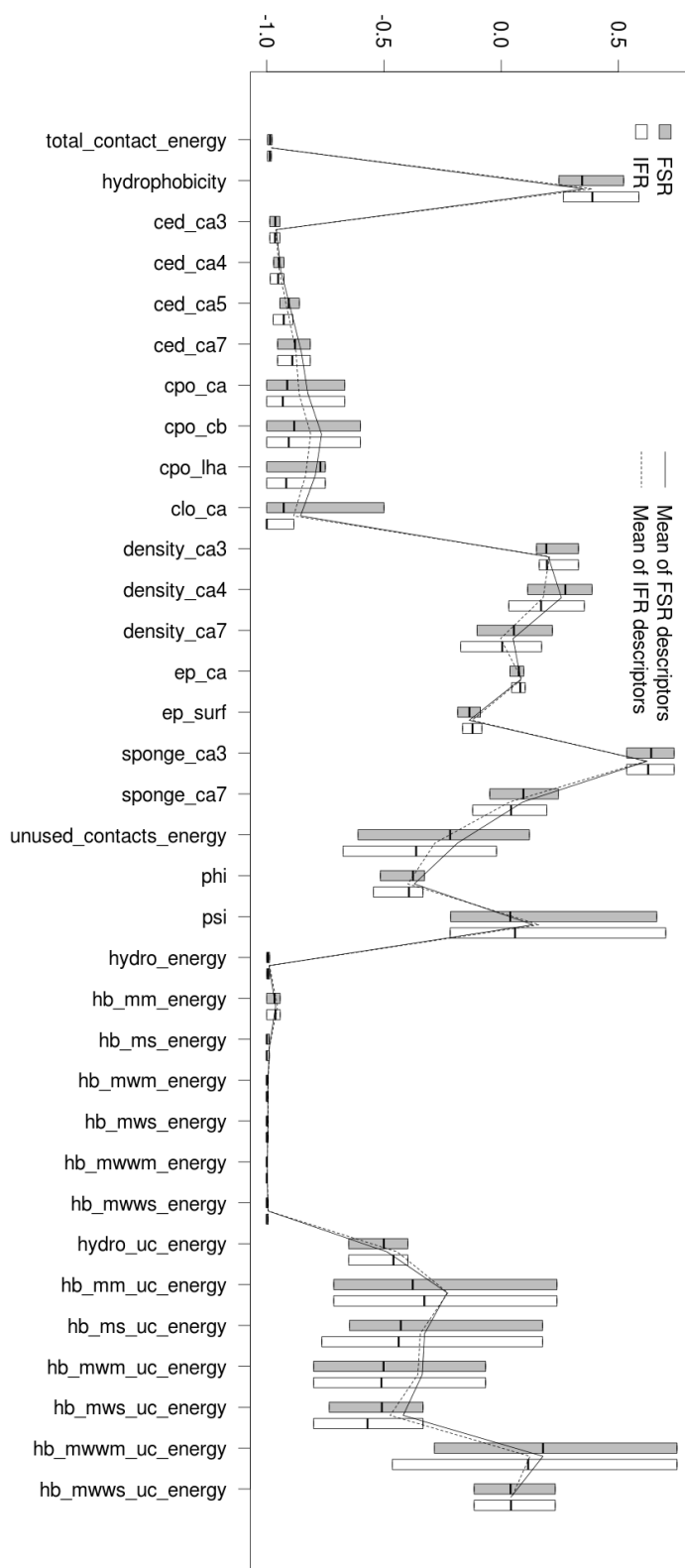
Glutamina:



Ácido Glutâmico:

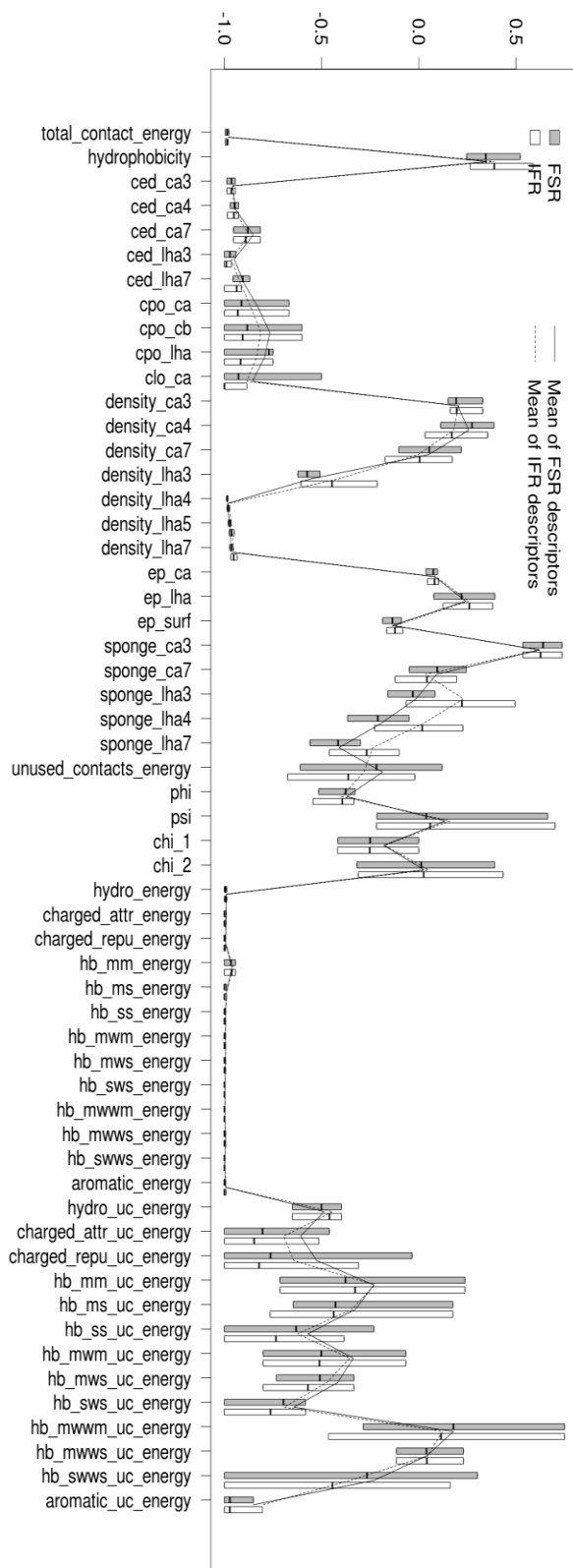


Glicina:



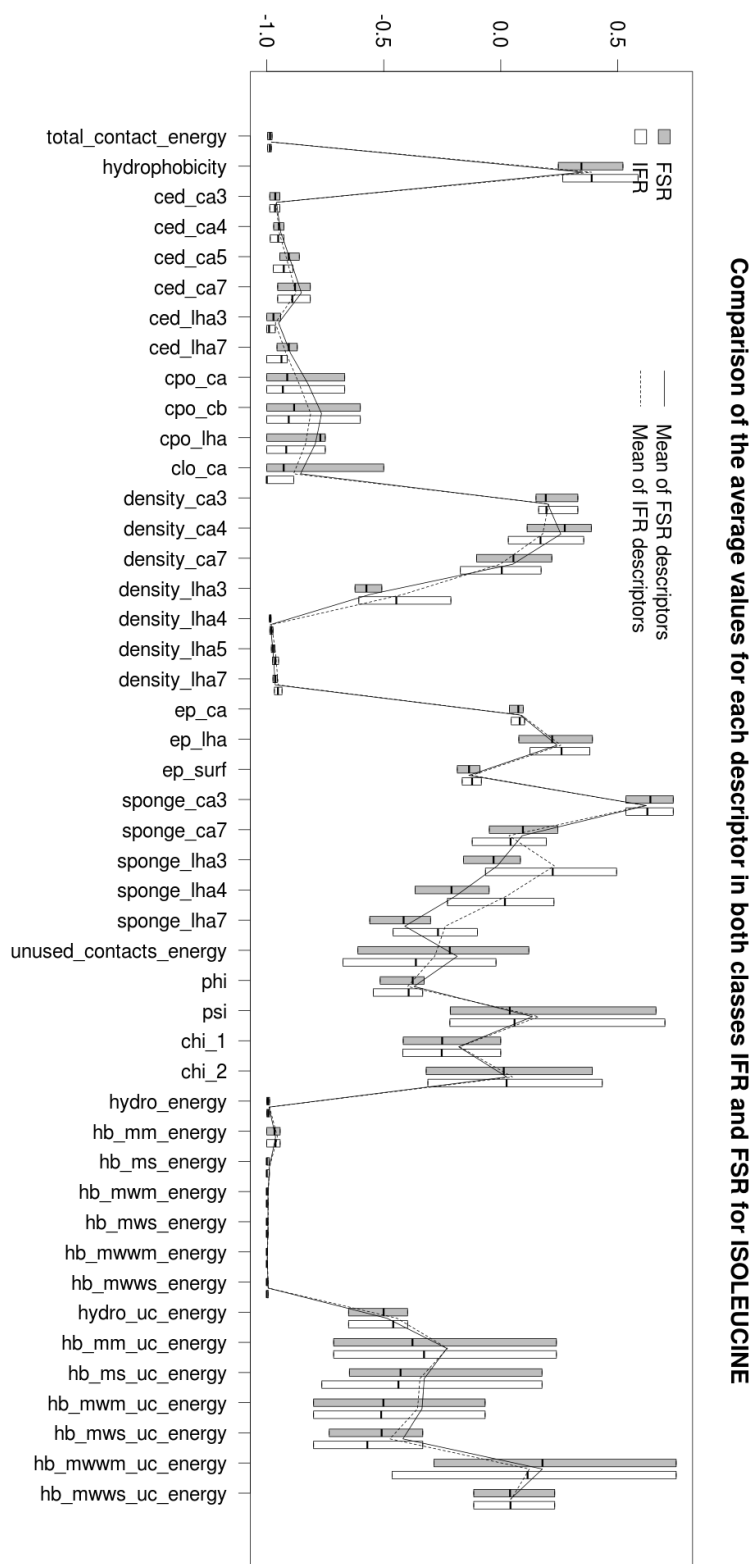


# Histidina:

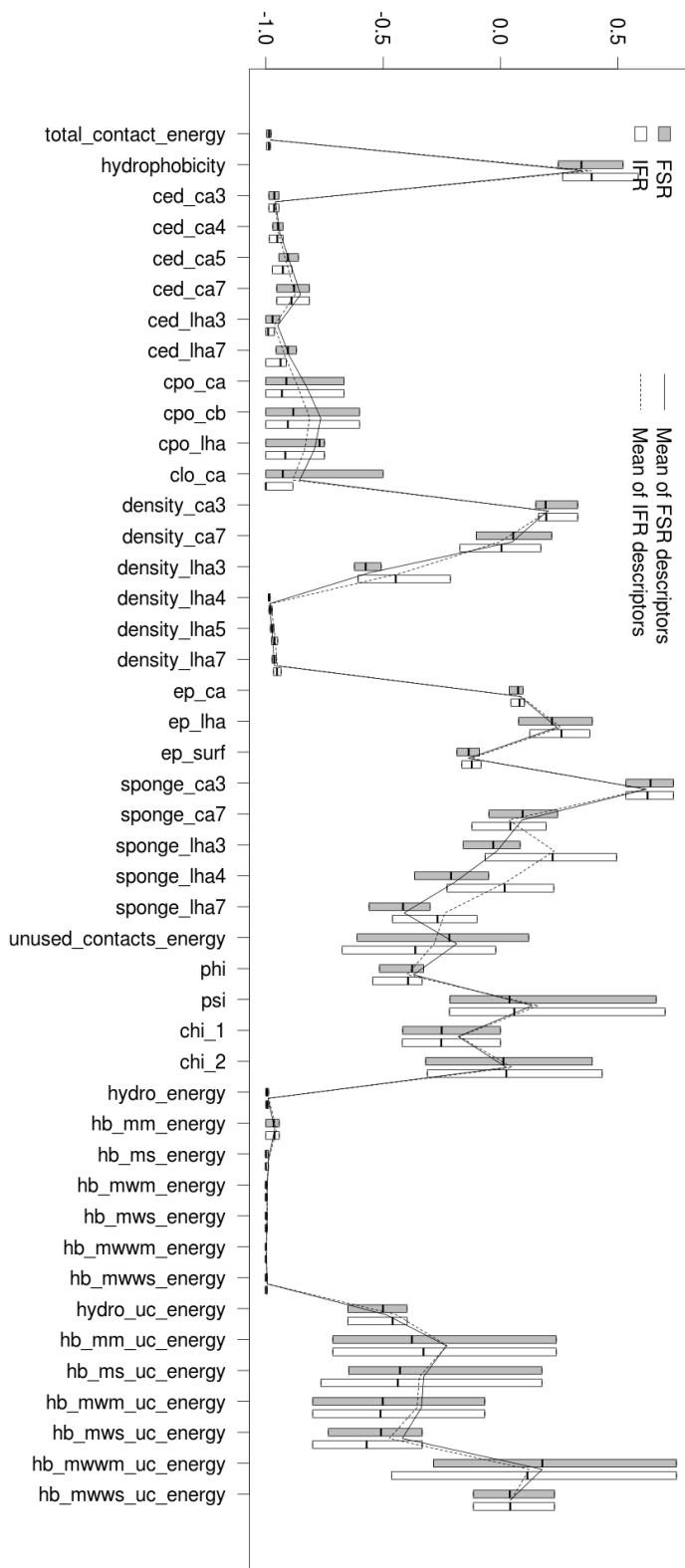


Comparison of the average values for each descriptor in both classes IFR and FSR for HISTIDINE

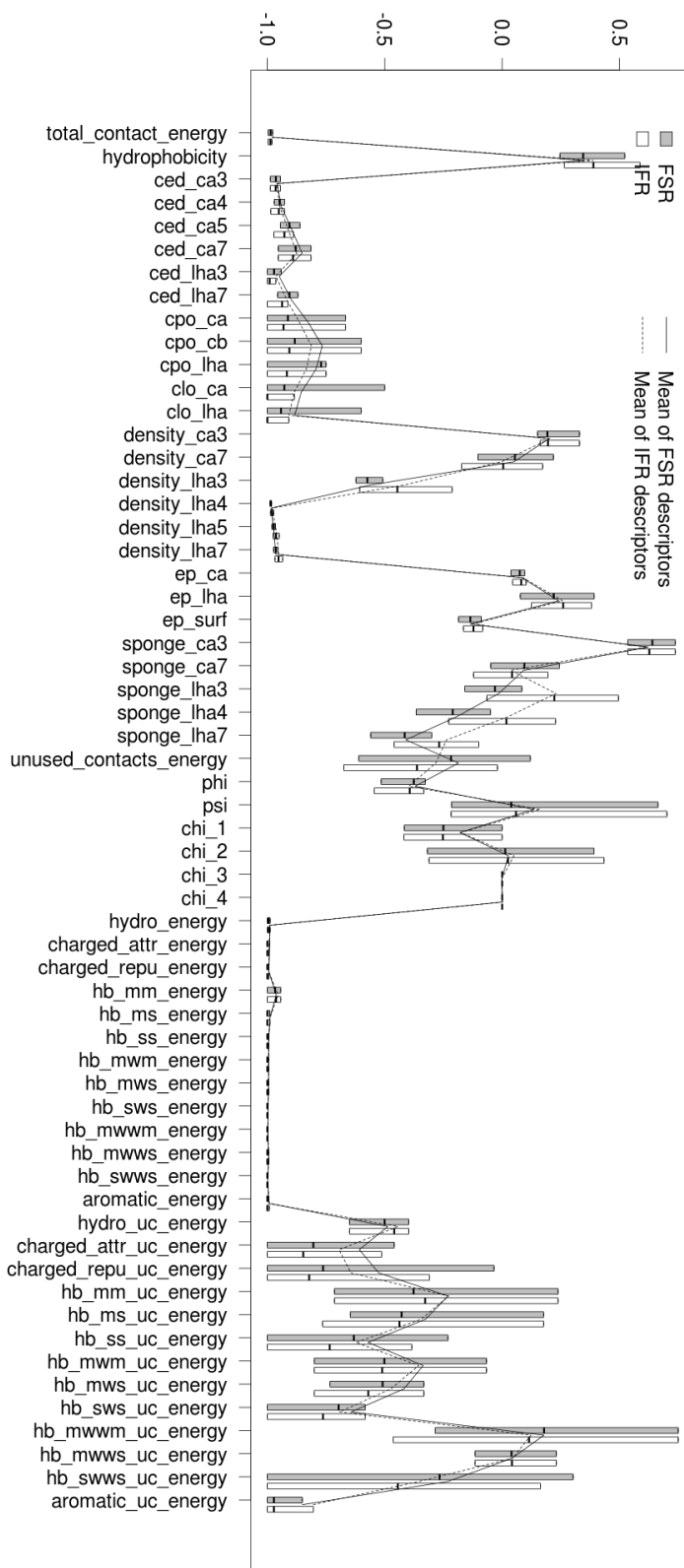
Isoleucina:



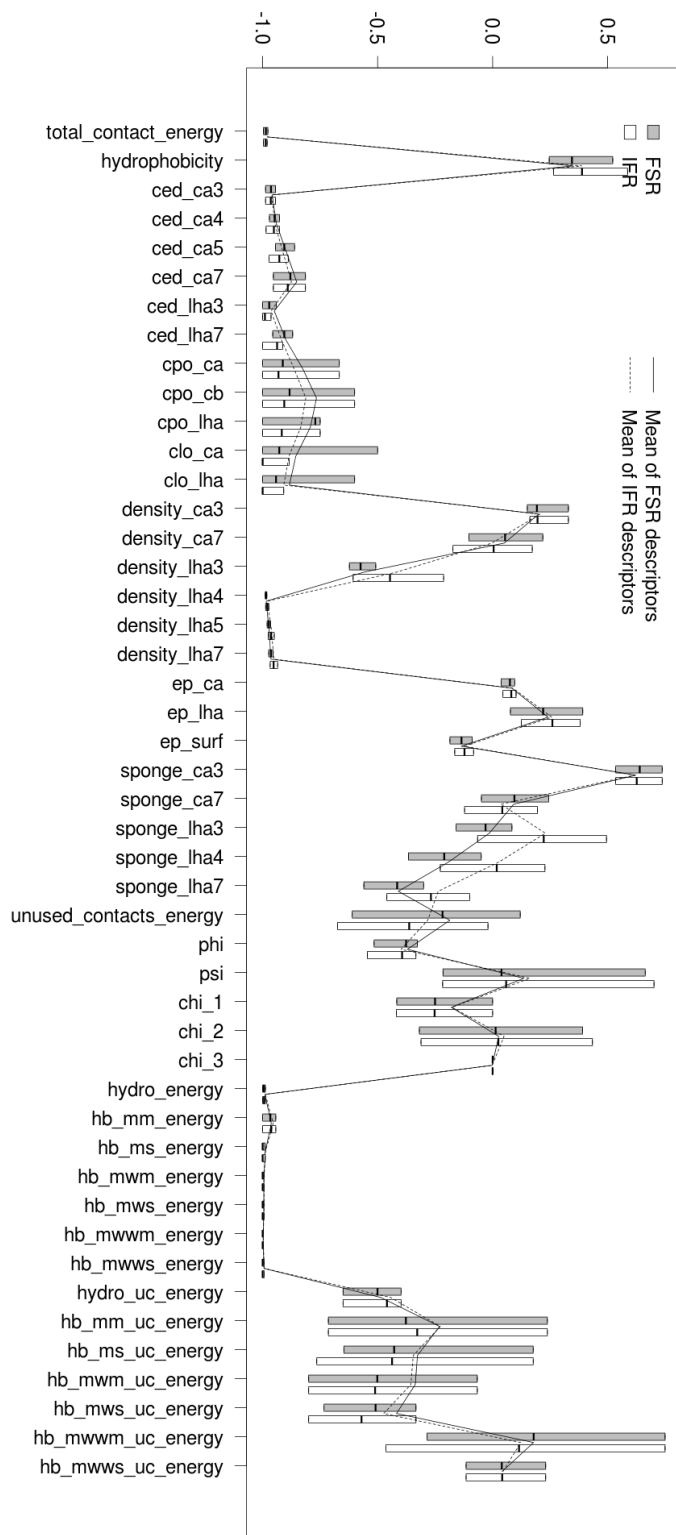
Leucina:



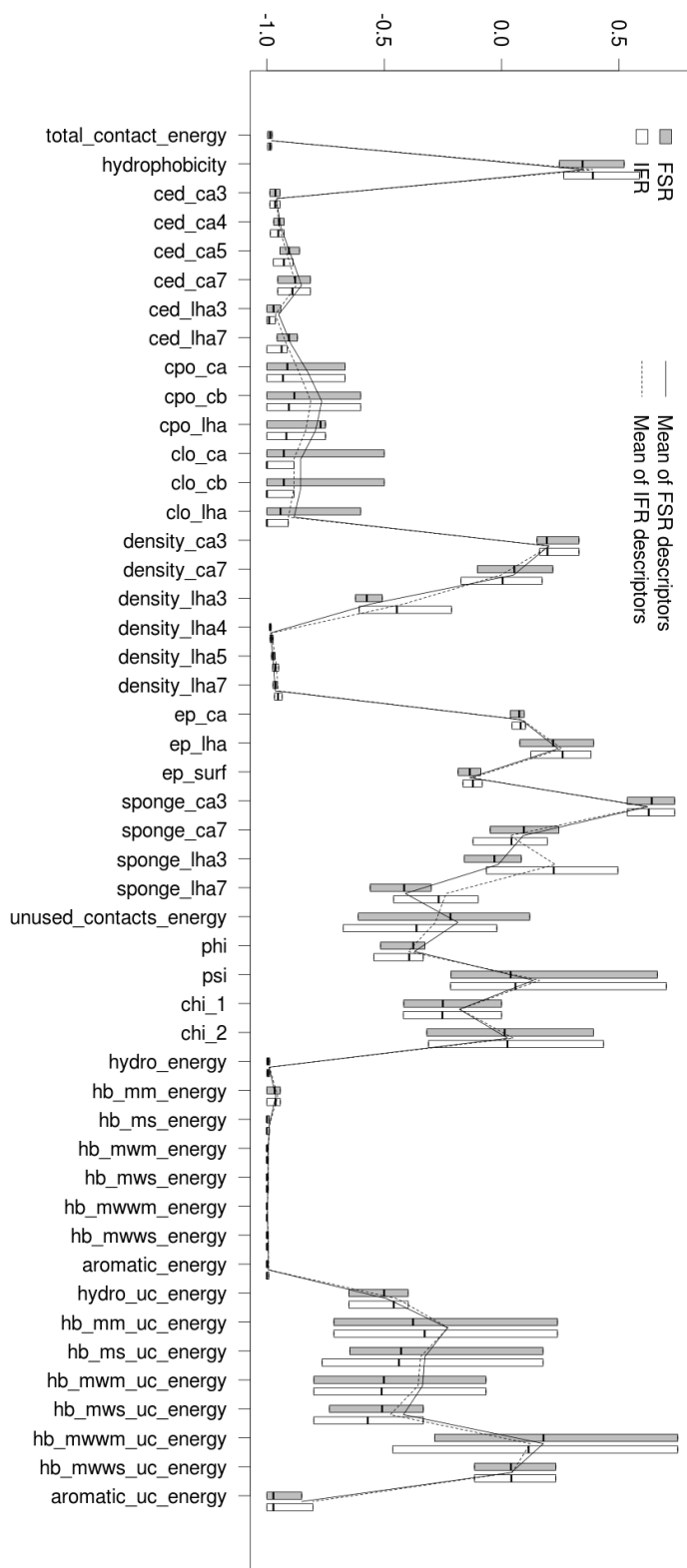
Lisina:



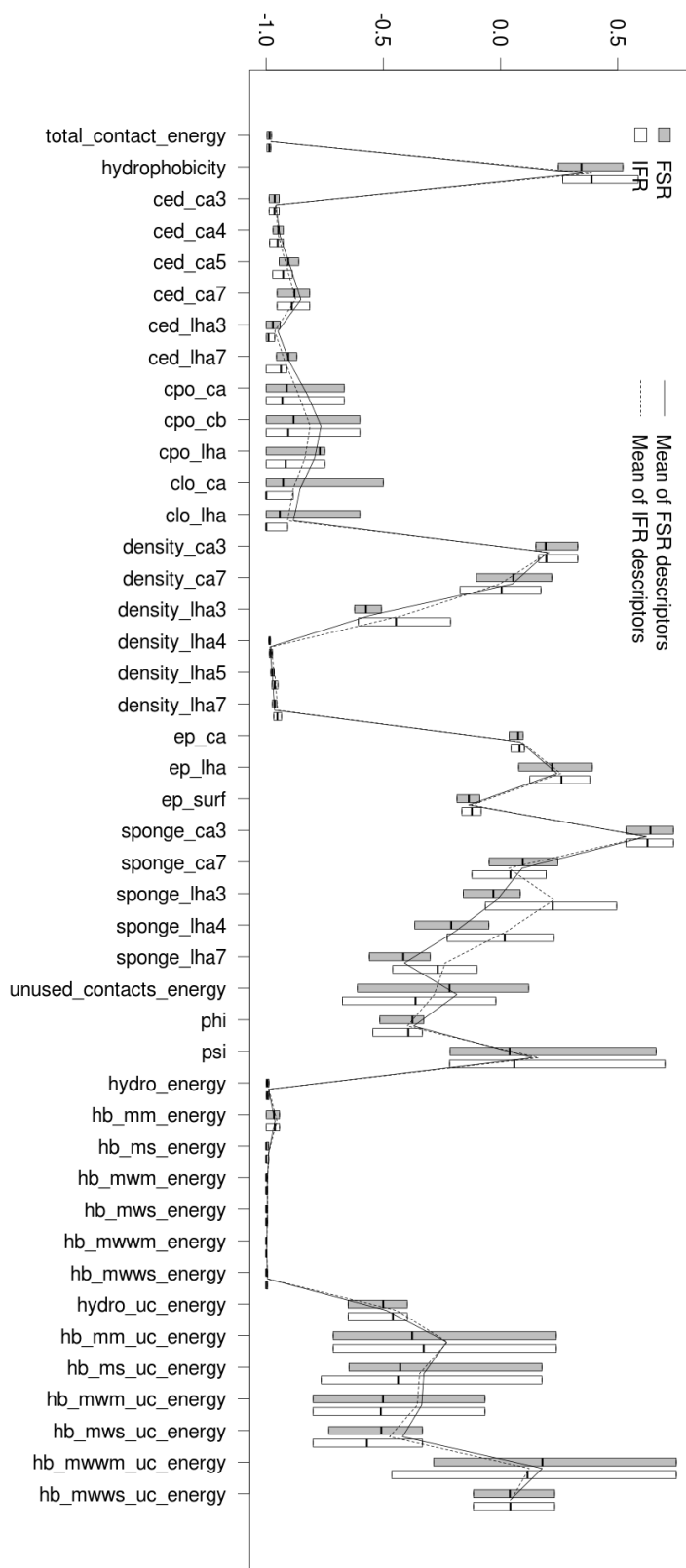
Metionina:



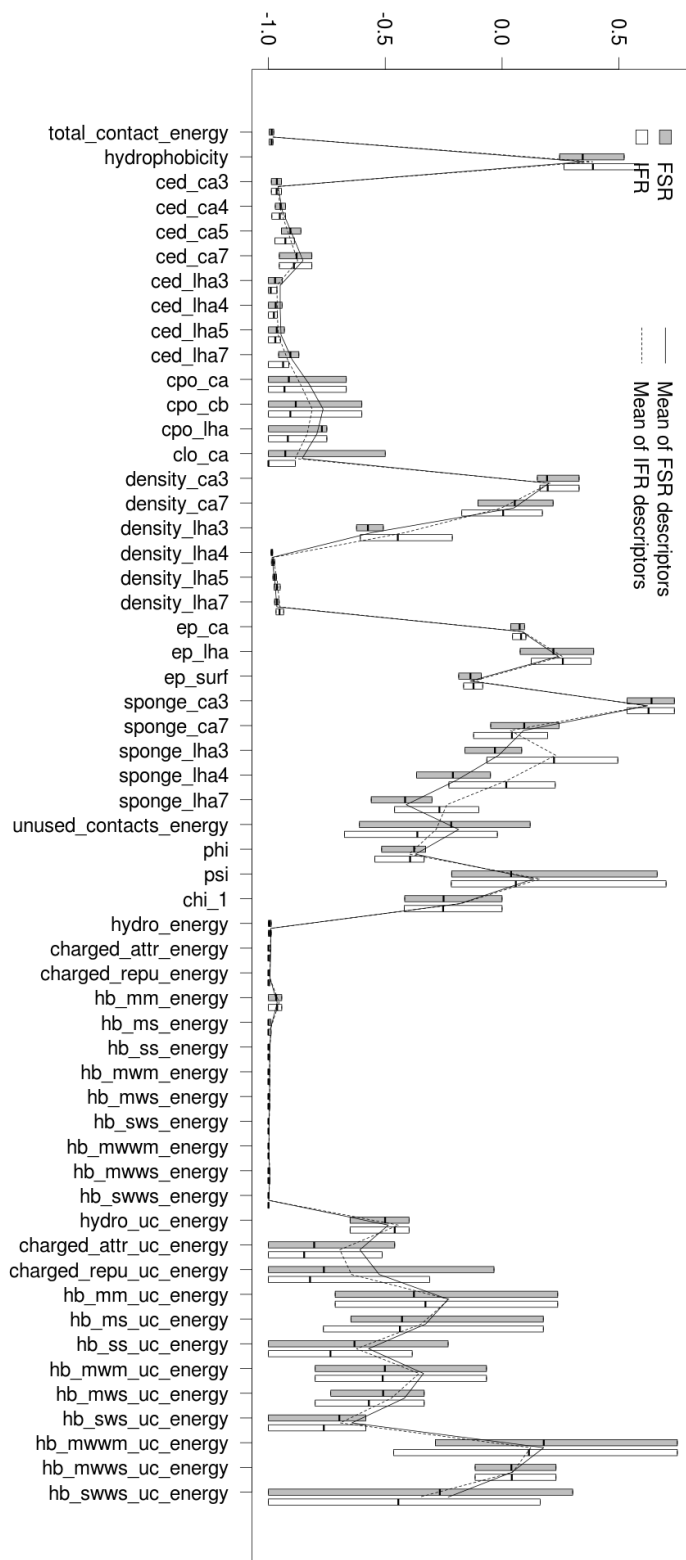
Fenilalanina:



Prolina:

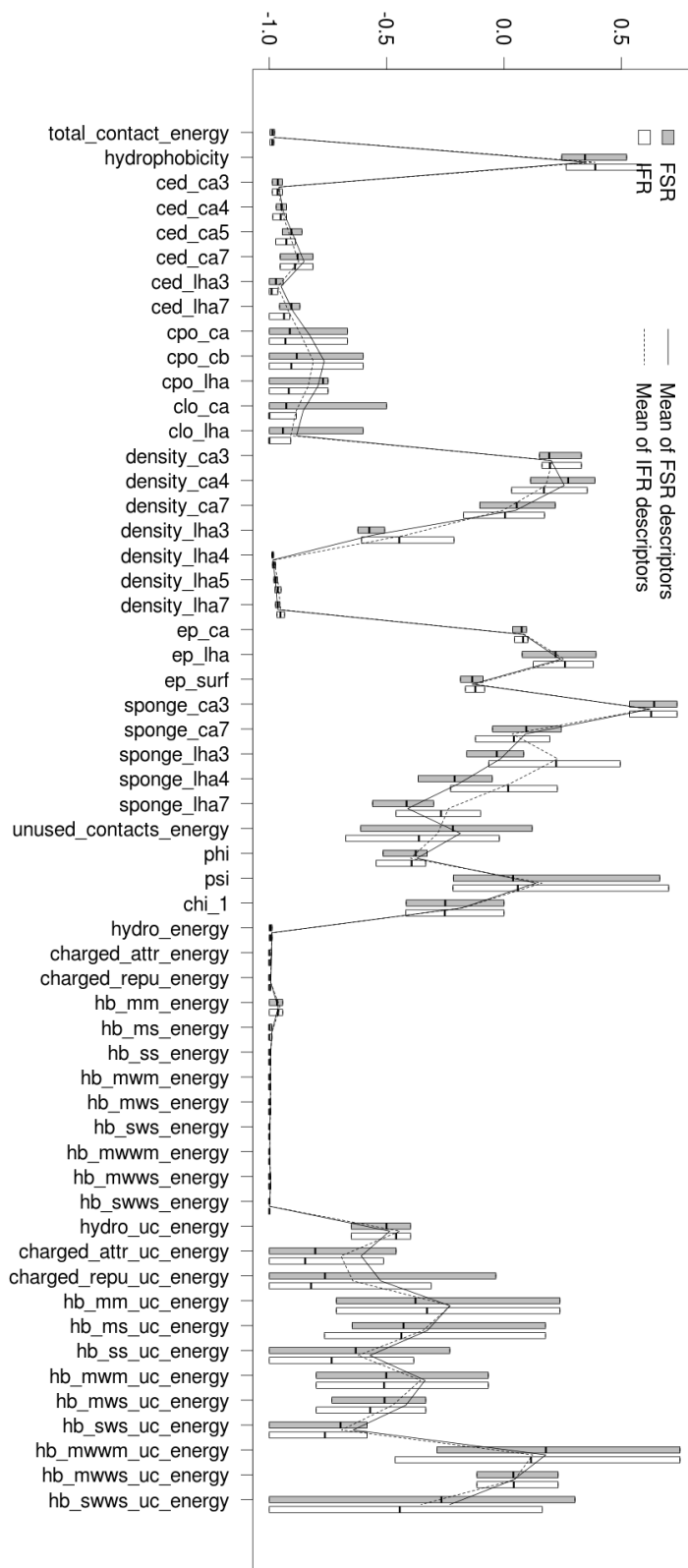


Serina:

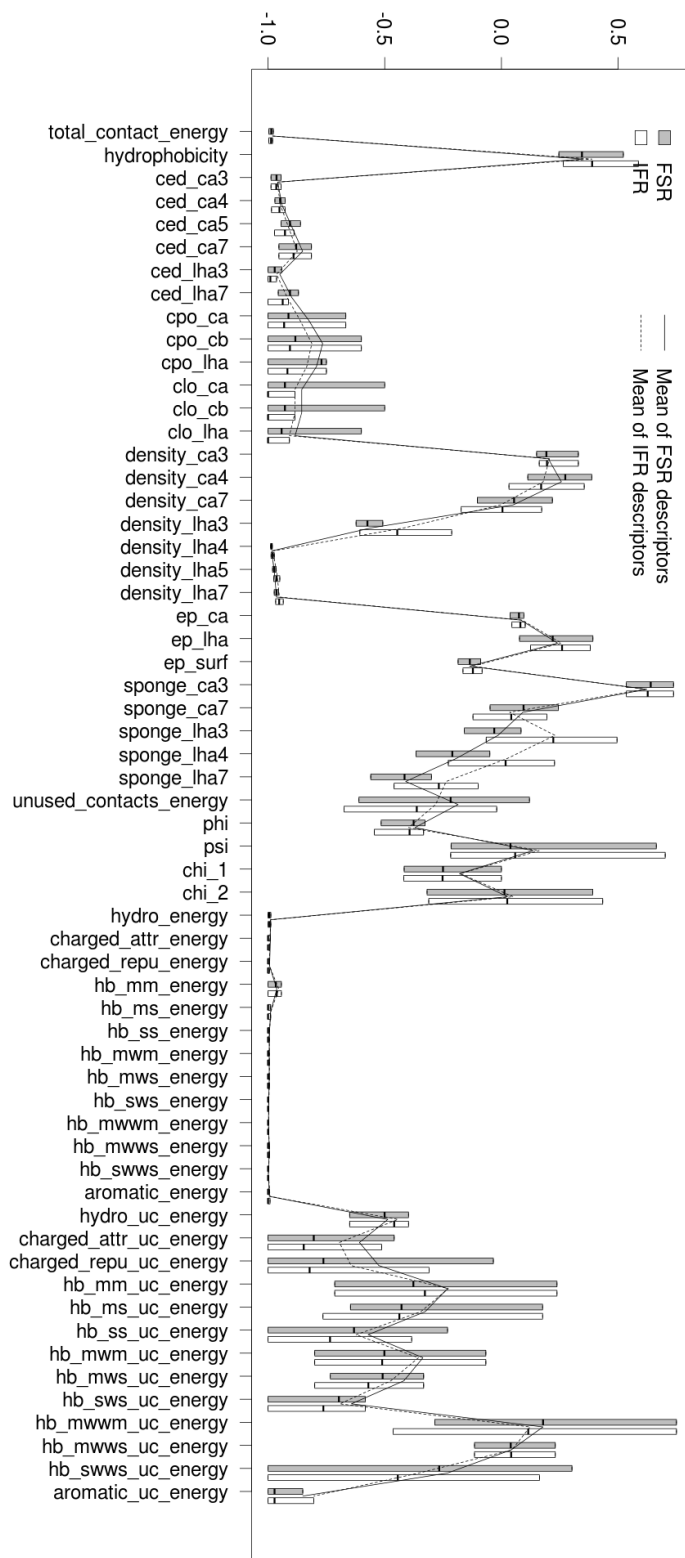




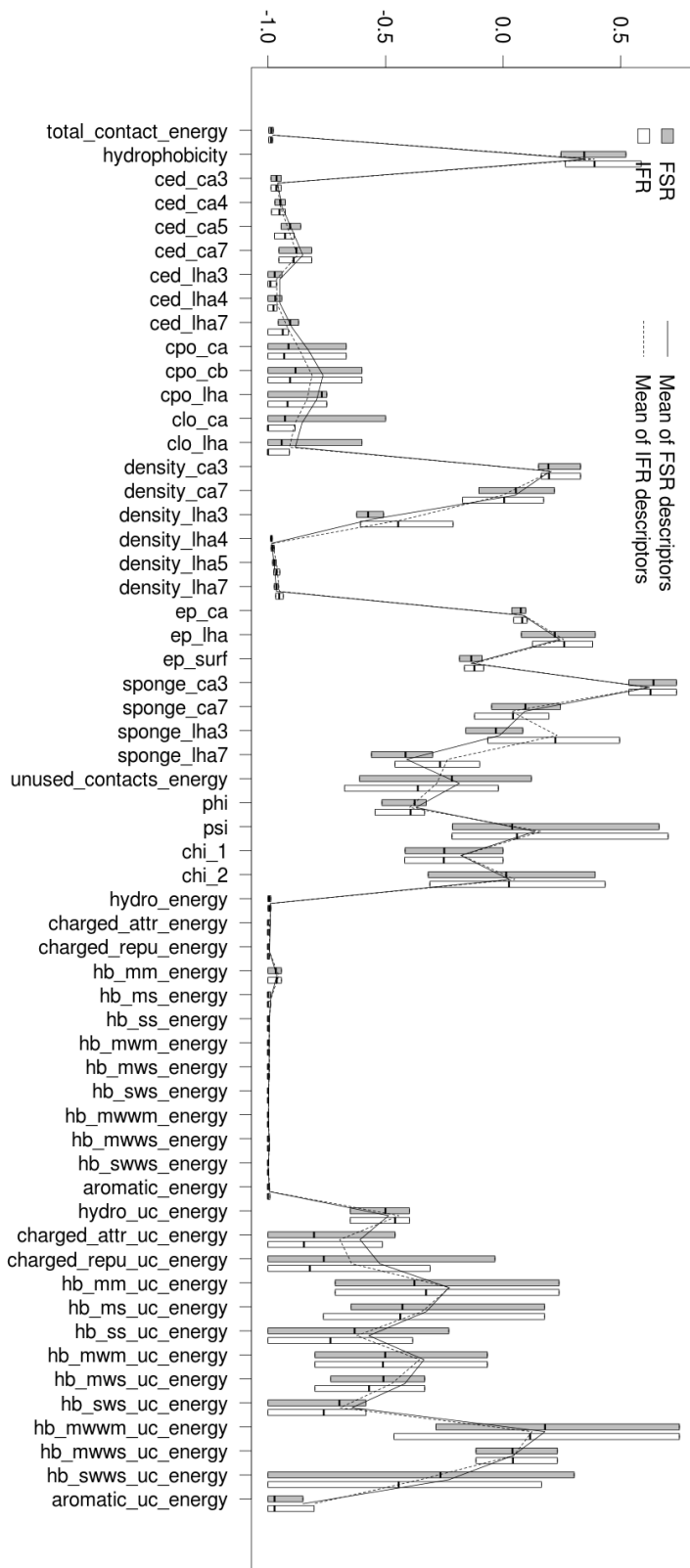
Treonina:



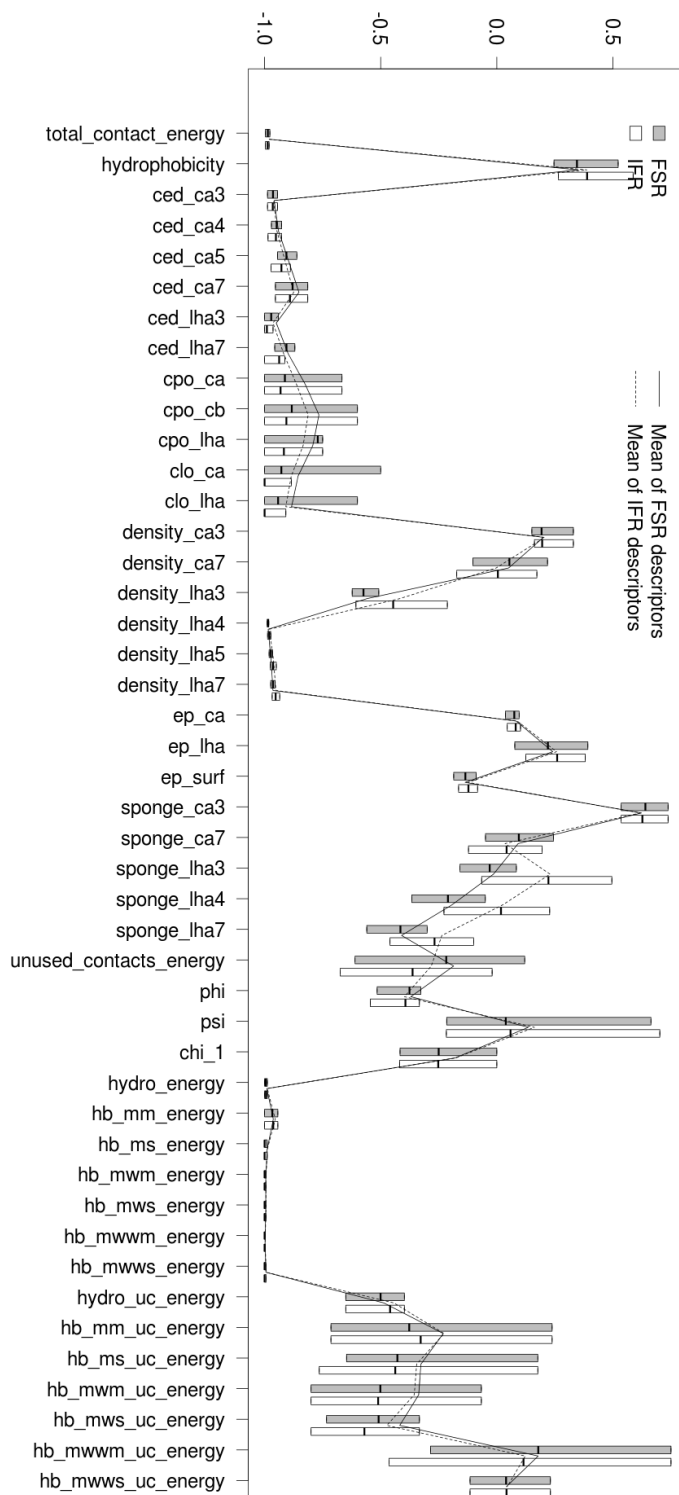
Triptofano:



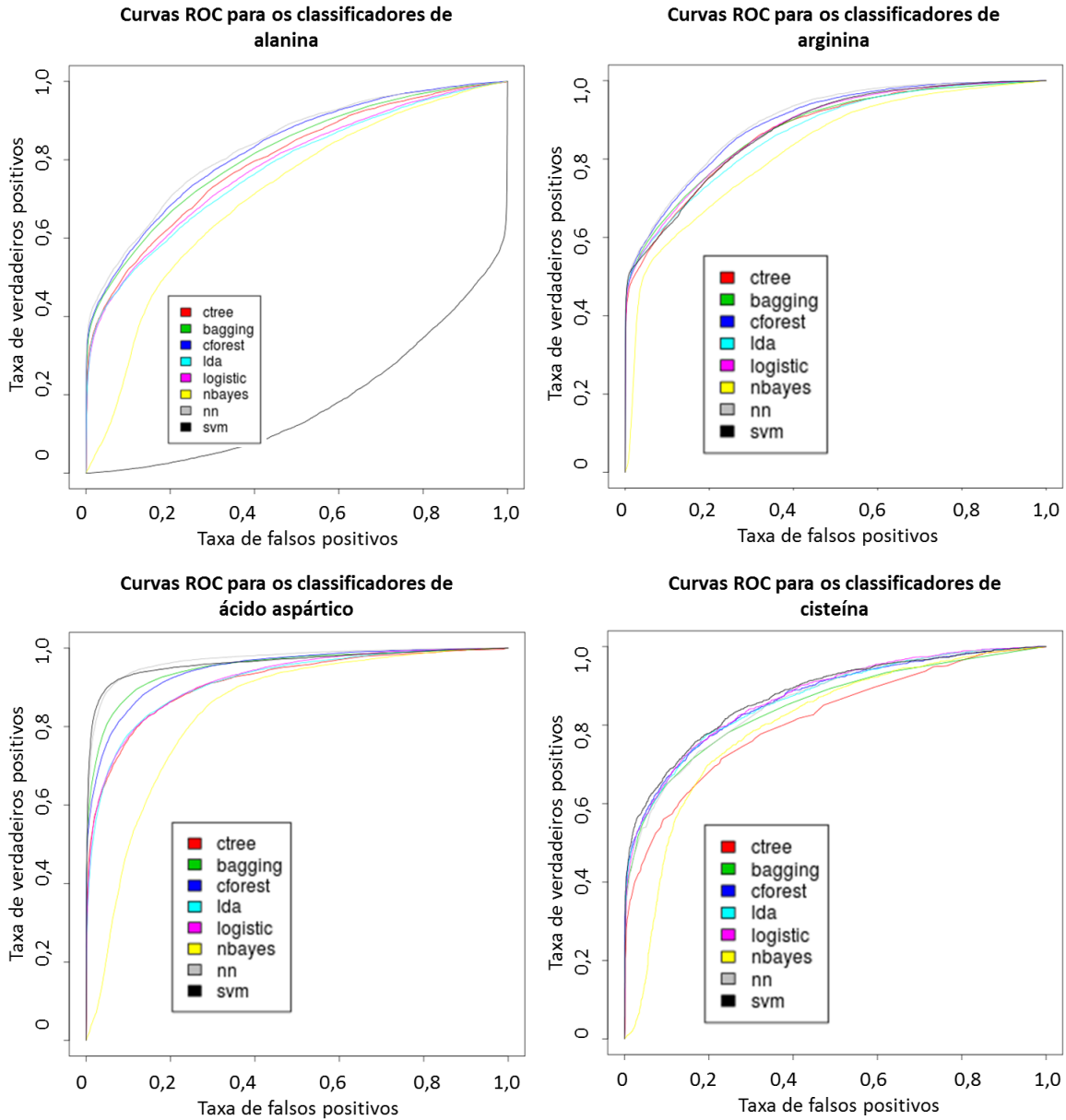
Tirosina:



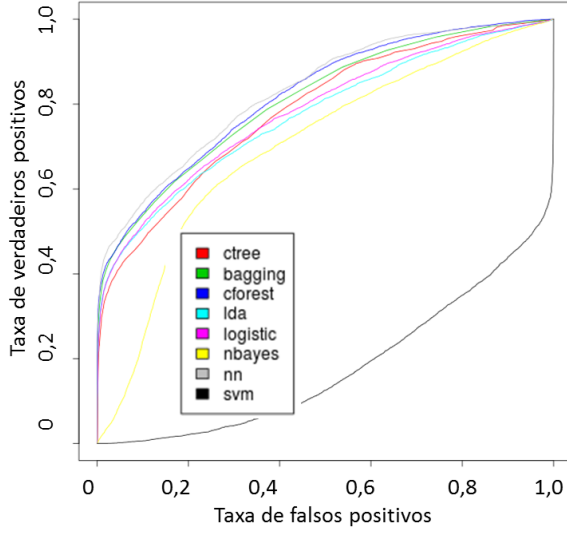
Valina:



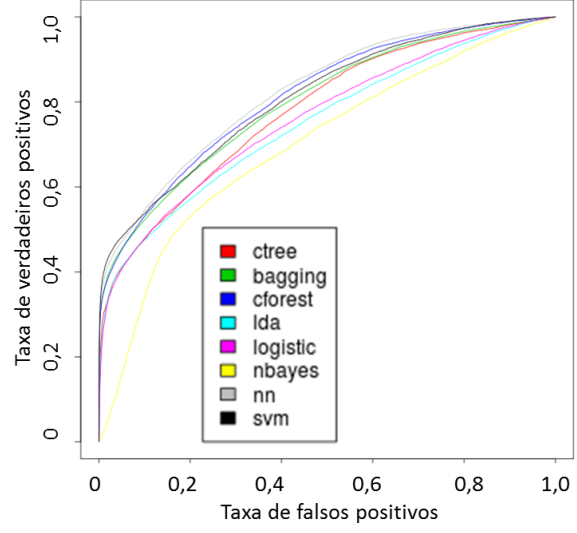
## 6.4 Apêndice 4 – Comparação das curvas ROC para cada um dos oito tipos de modelos classificadores para cada aminoácido.



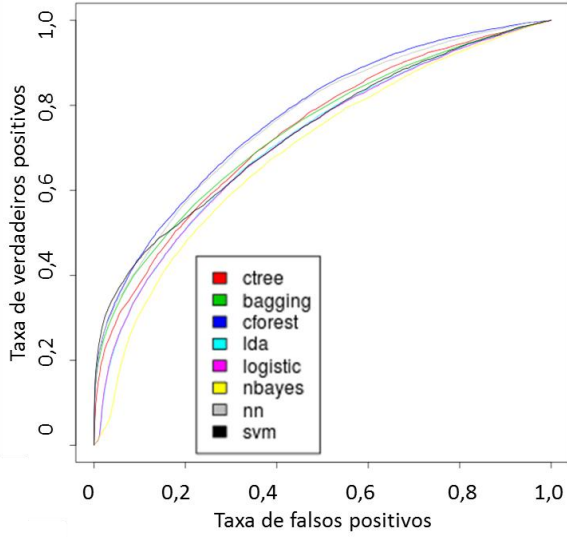
Curvas ROC para os classificadores de glutamina



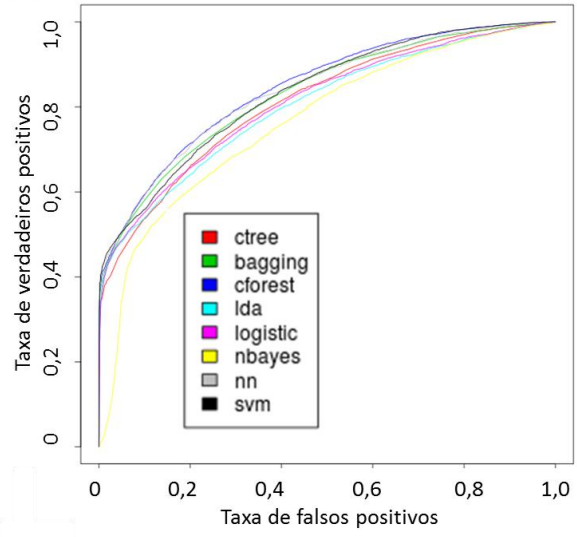
Curvas ROC para os classificadores de ácido glutâmico



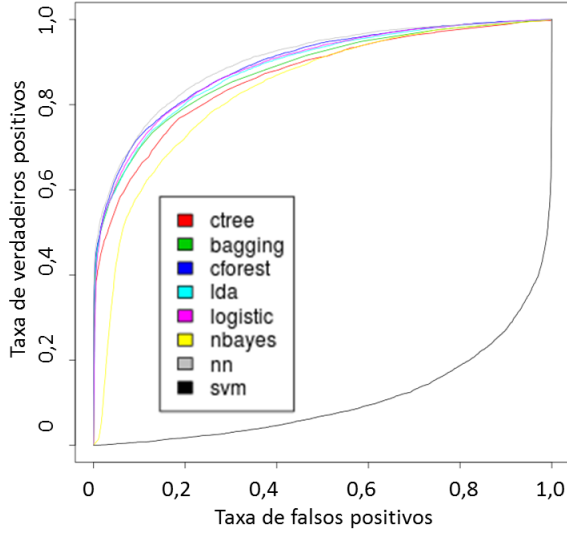
Curvas ROC para os classificadores de glicina



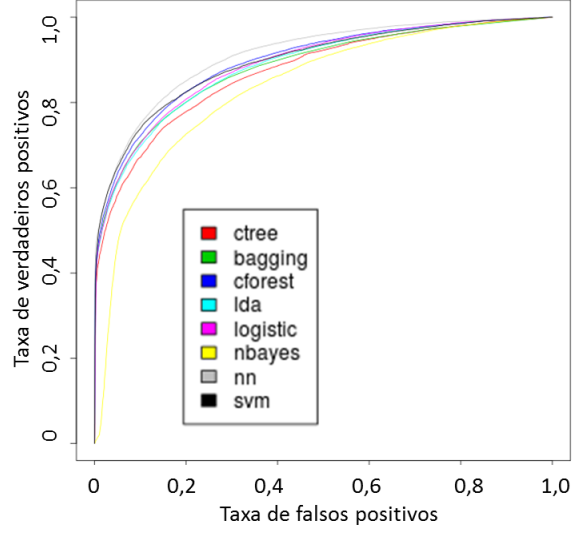
Curvas ROC para os classificadores de histidina



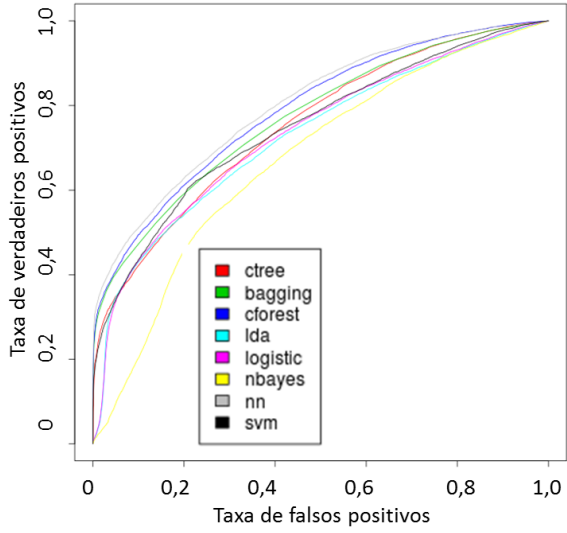
Curvas ROC para os classificadores de isoleucina



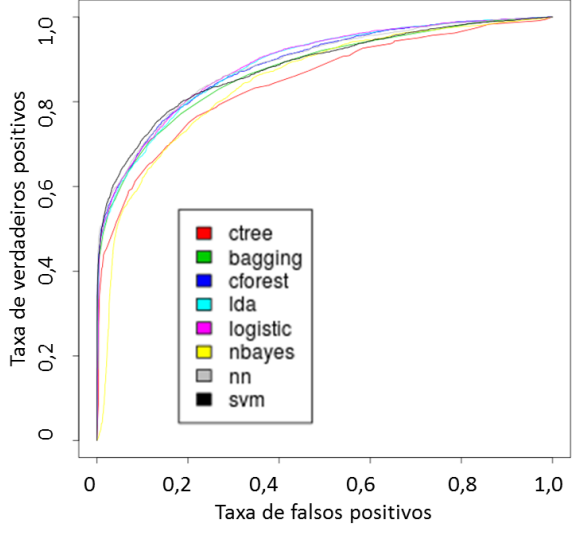
Curvas ROC para os classificadores de leucina



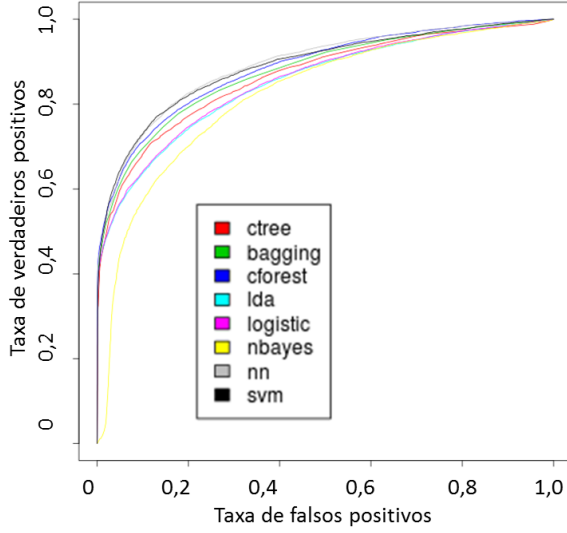
Curvas ROC para os classificadores de lisina



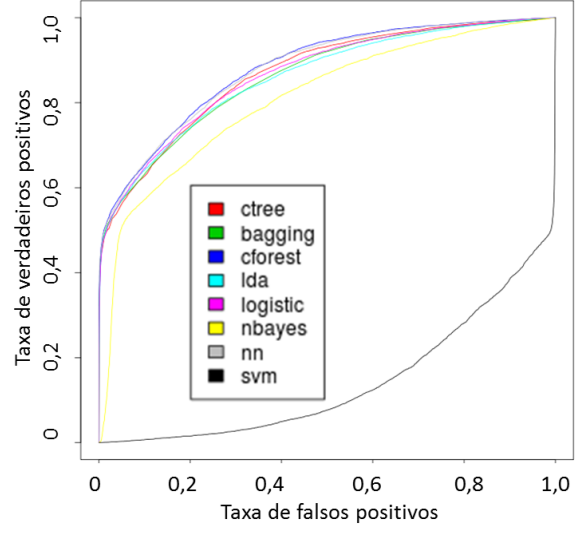
Curvas ROC para os classificadores de metionina



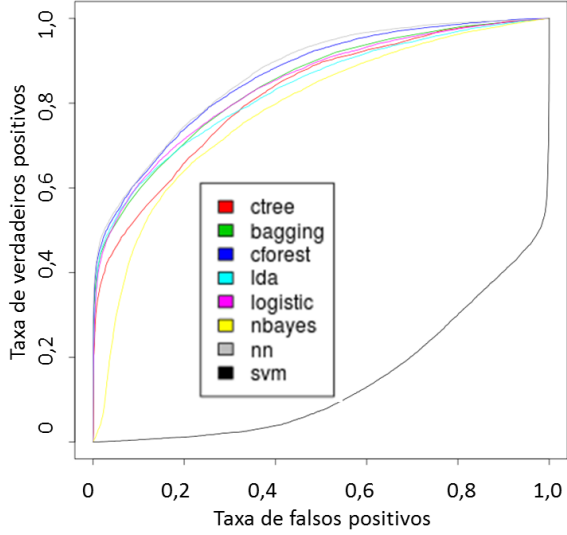
Curvas ROC para os classificadores de fenilalanina



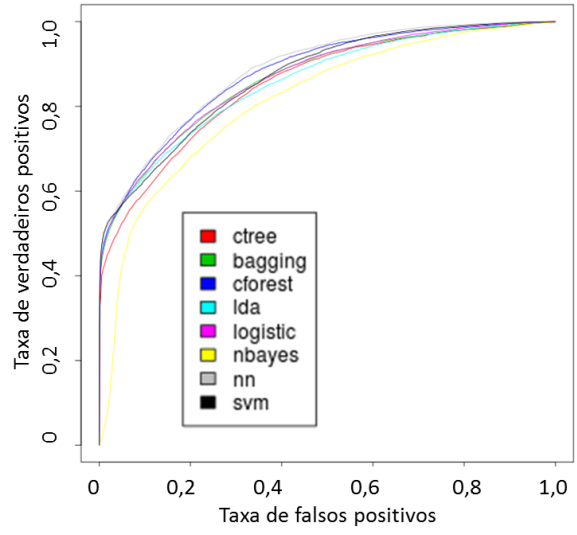
Curvas ROC para os classificadores de prolina



Curvas ROC para os classificadores de serina

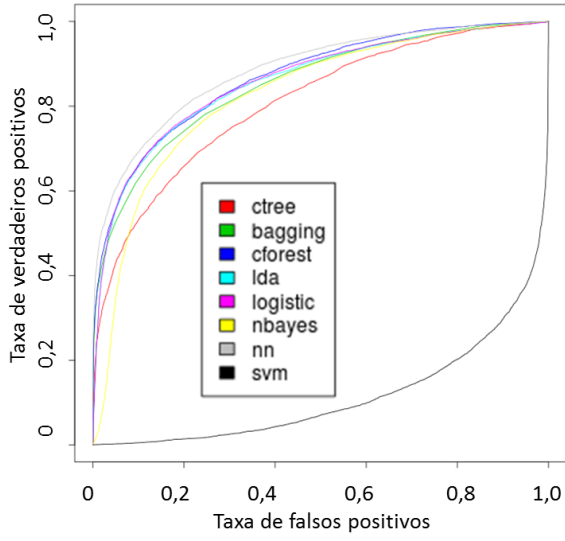


Curvas ROC para os classificadores de treonina

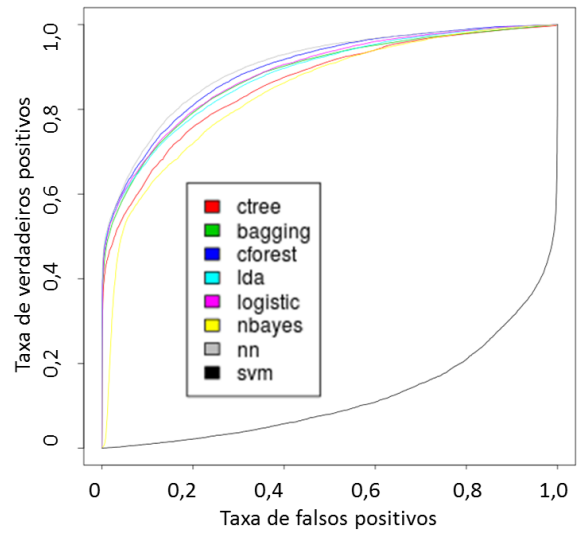




Curvas ROC para os classificadores de tirosina

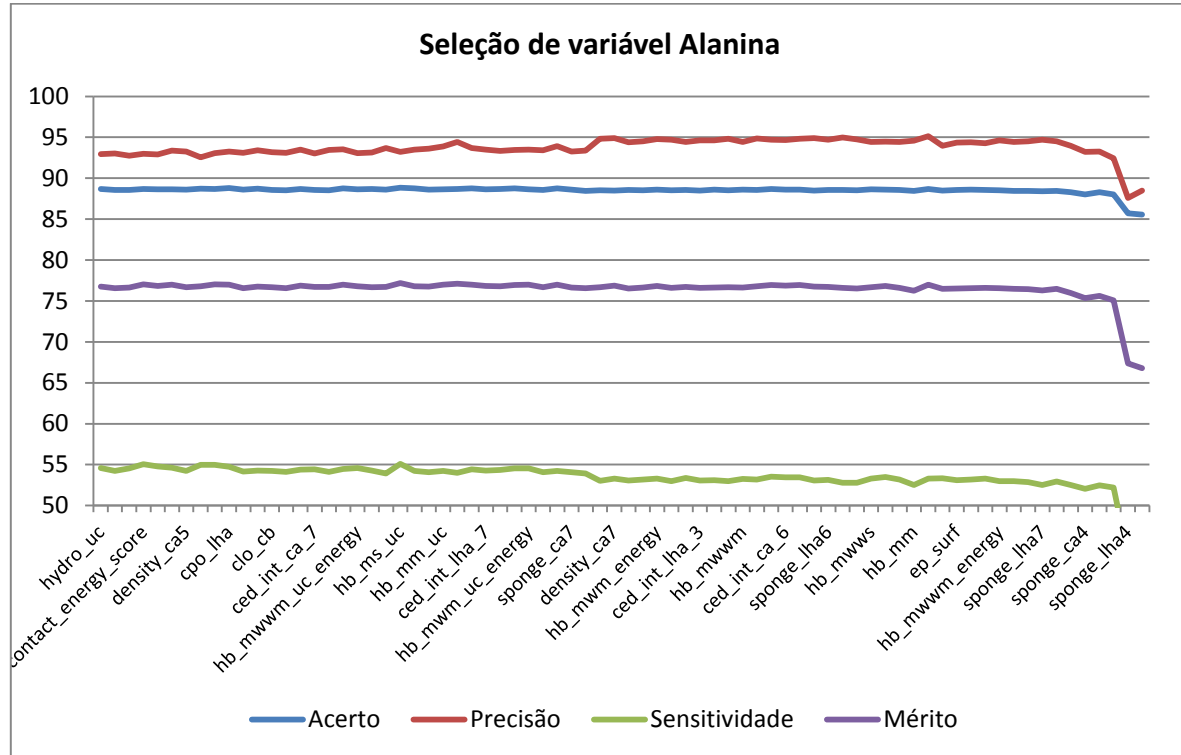


Curvas ROC para os classificadores de valina

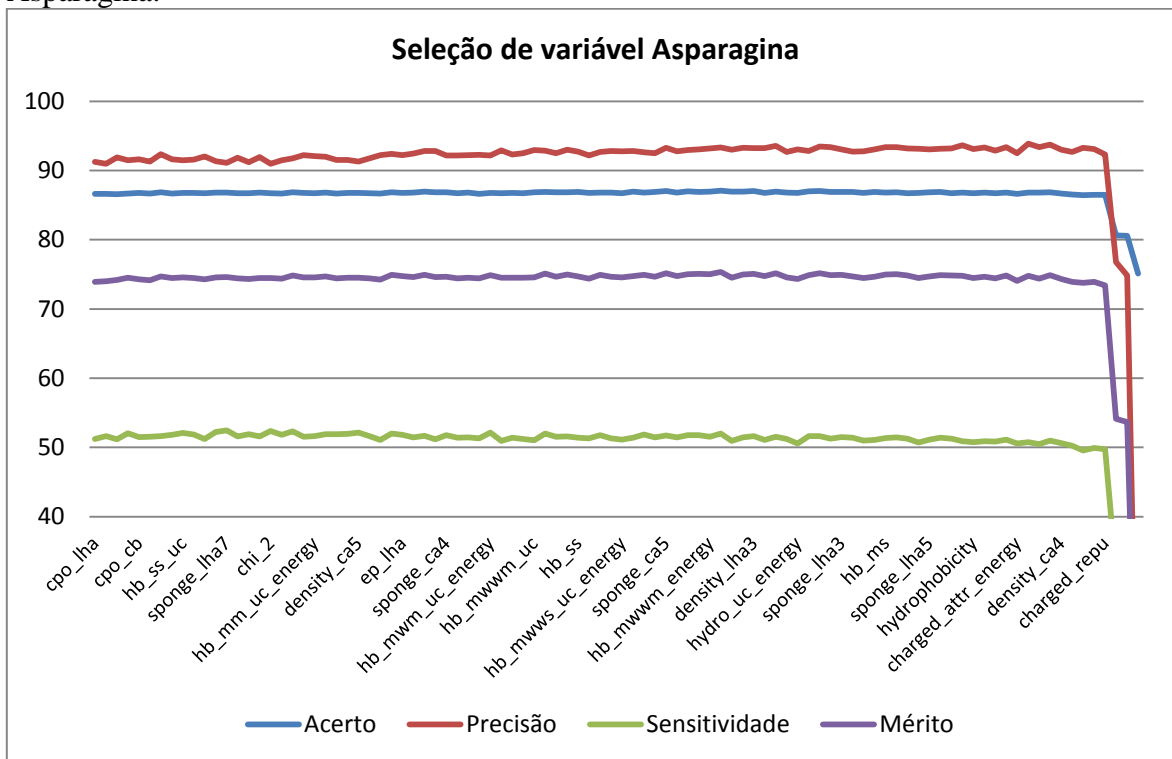


## 6.5 Apêndice 5 – Processo de seleção de variáveis para o classificador por redes neurais.

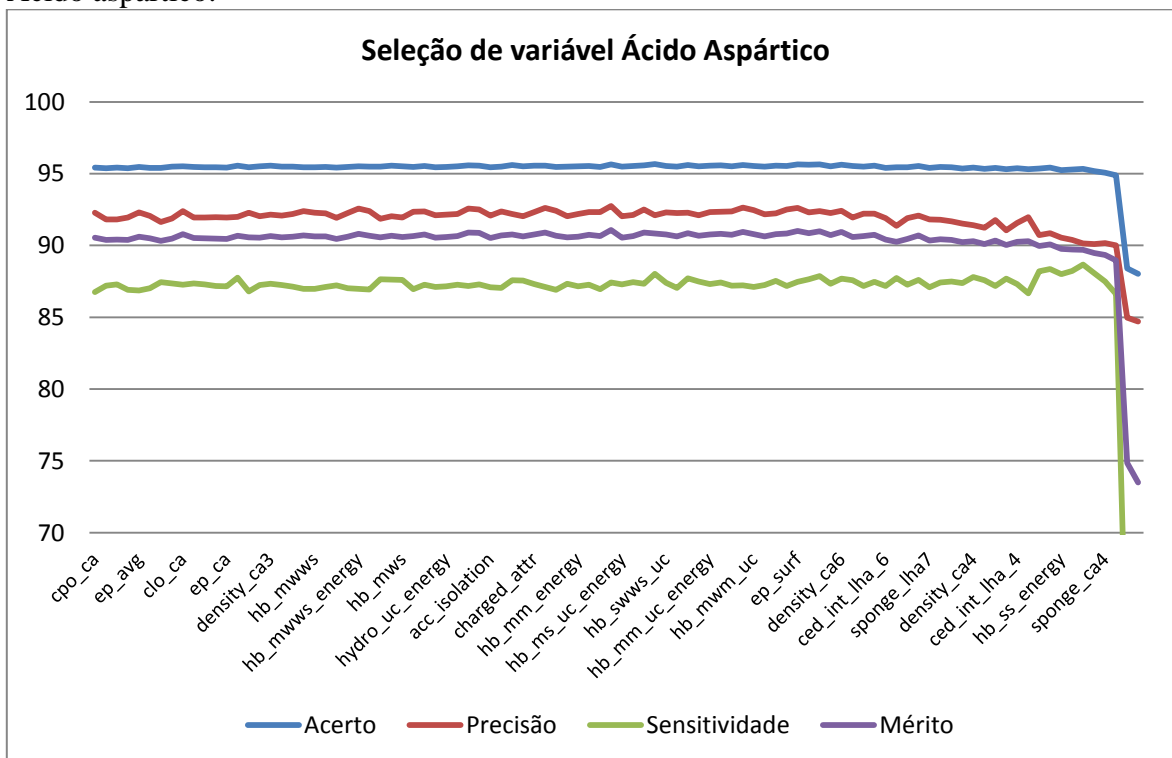
Alanina:



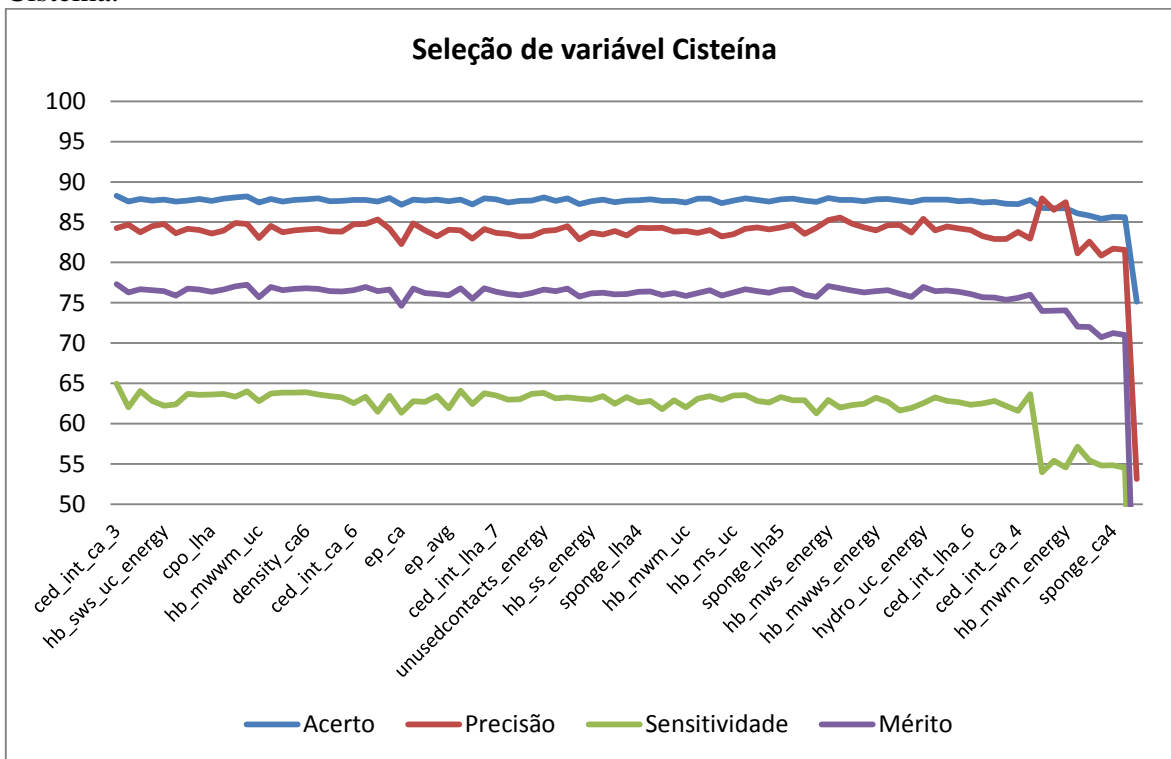
Asparagina:



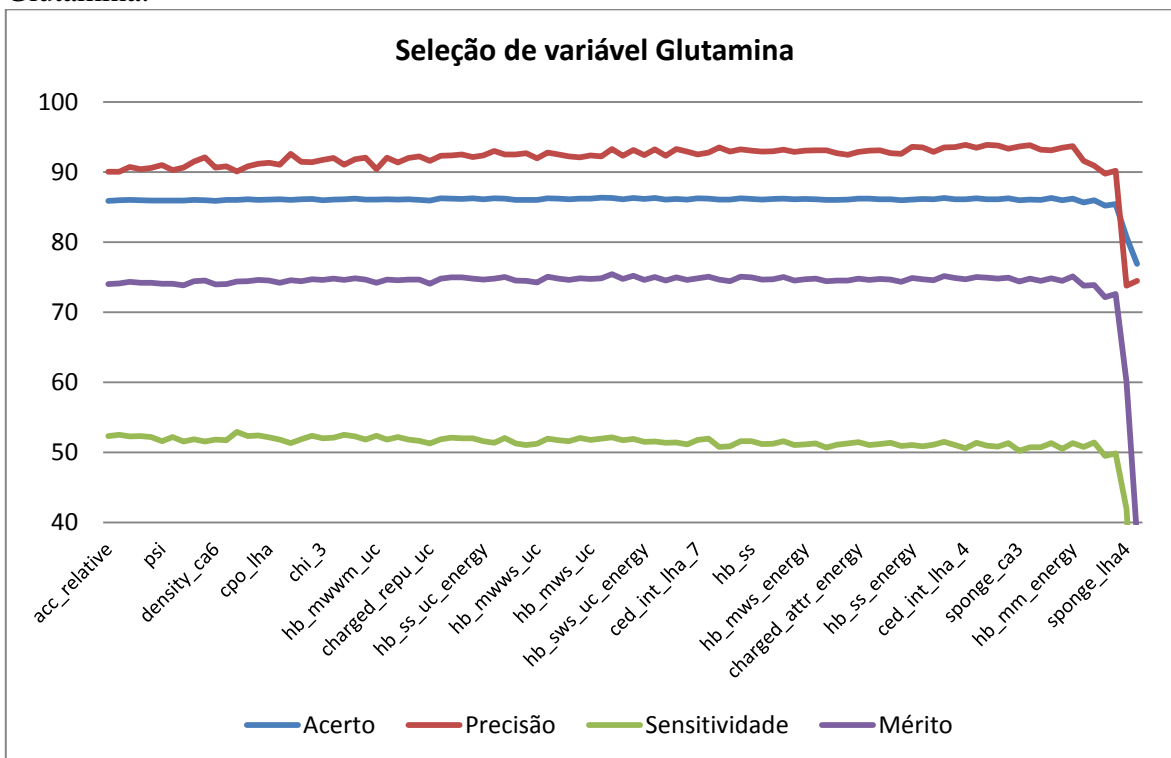
Ácido aspártico:



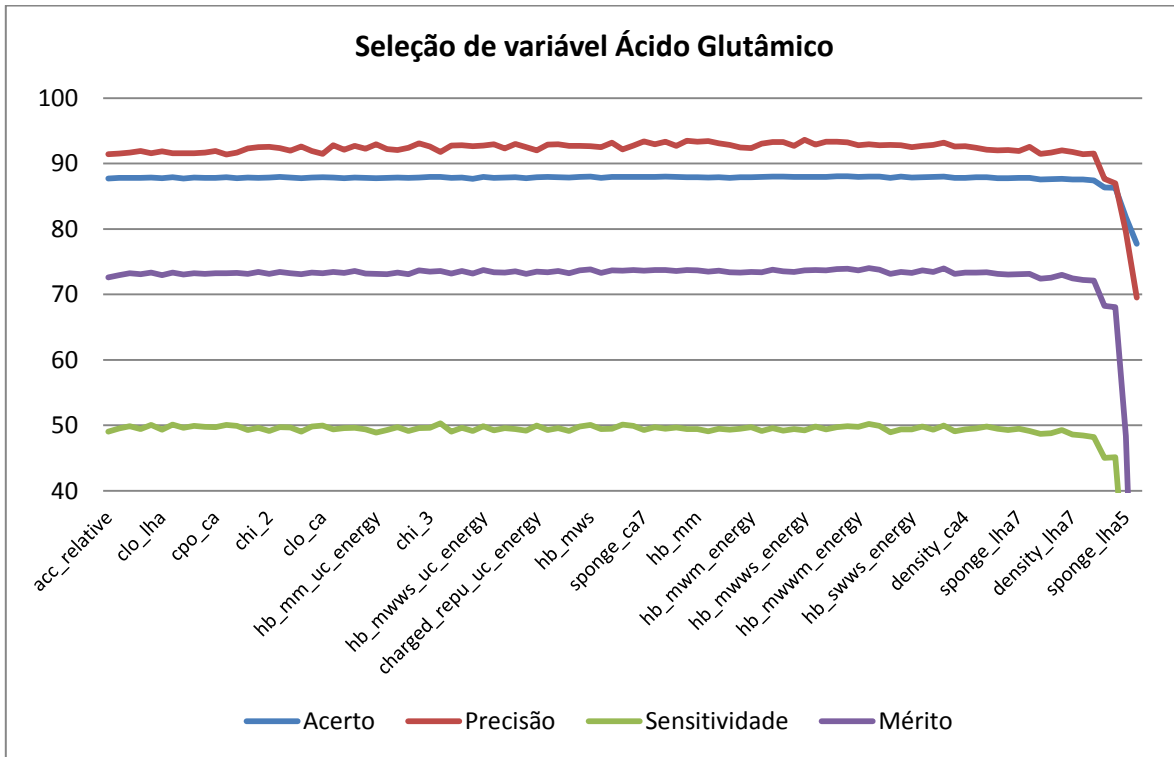
Cisteína:



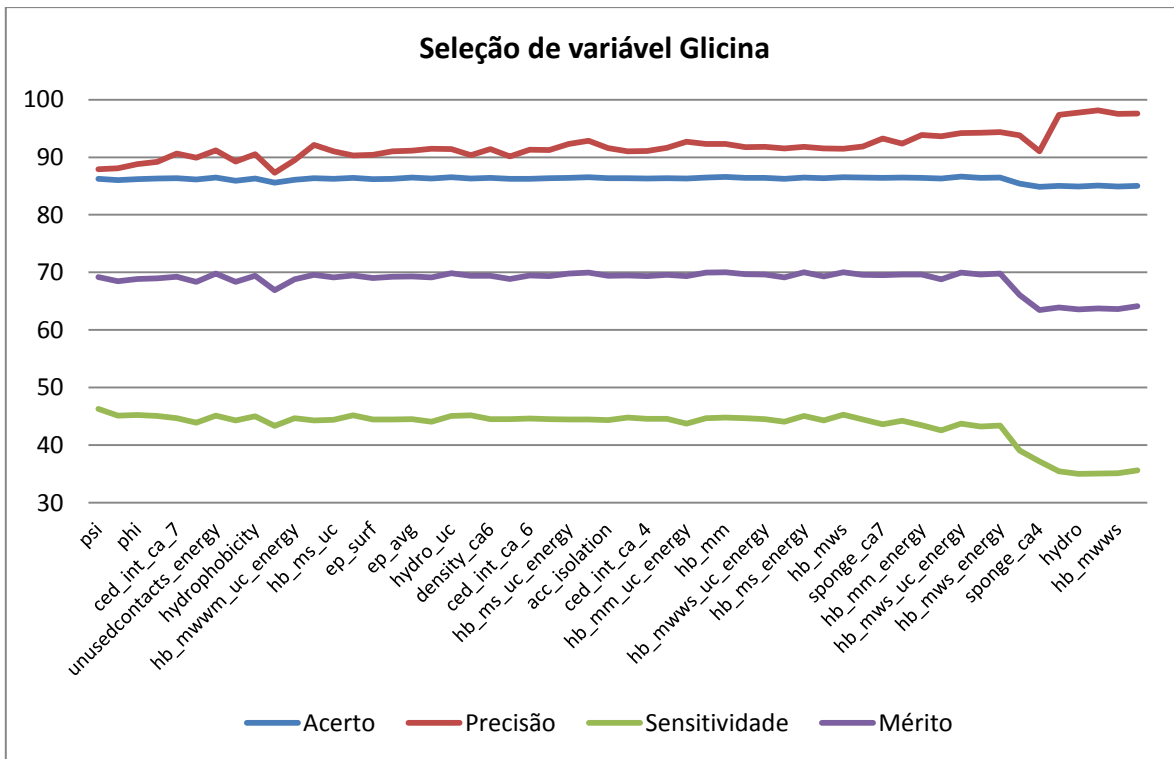
Glutamina:



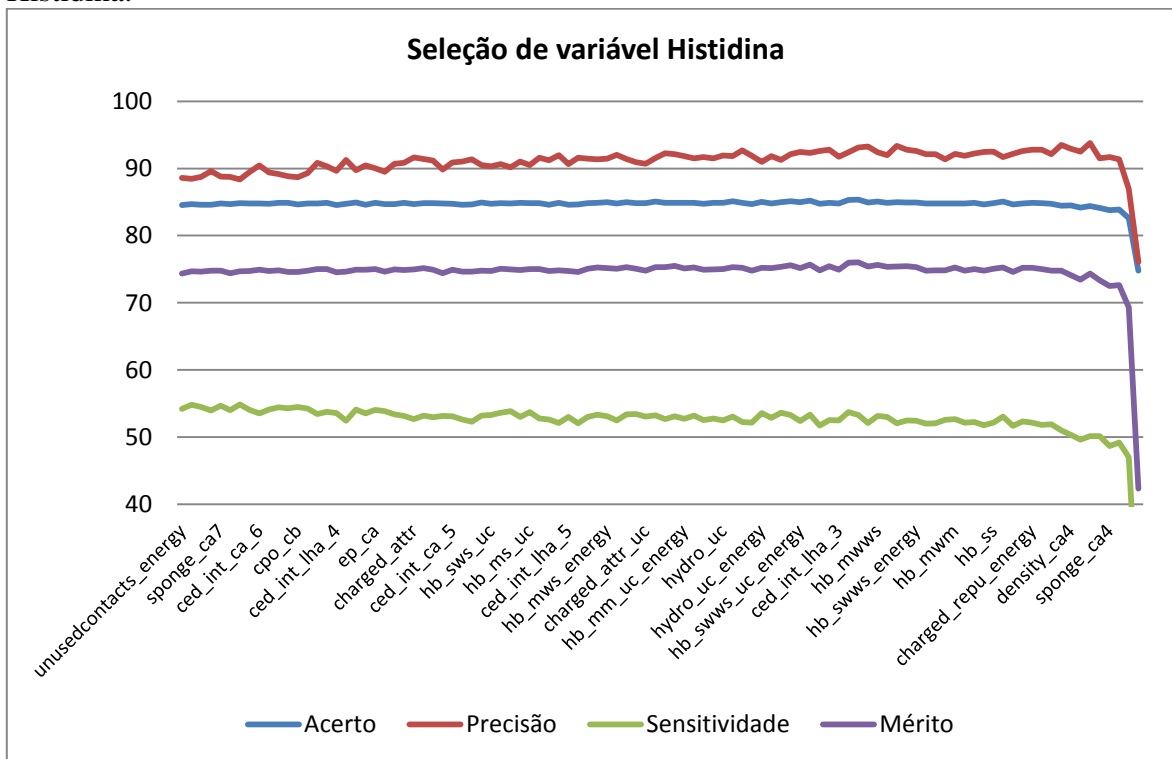
Ácido Glutâmico:



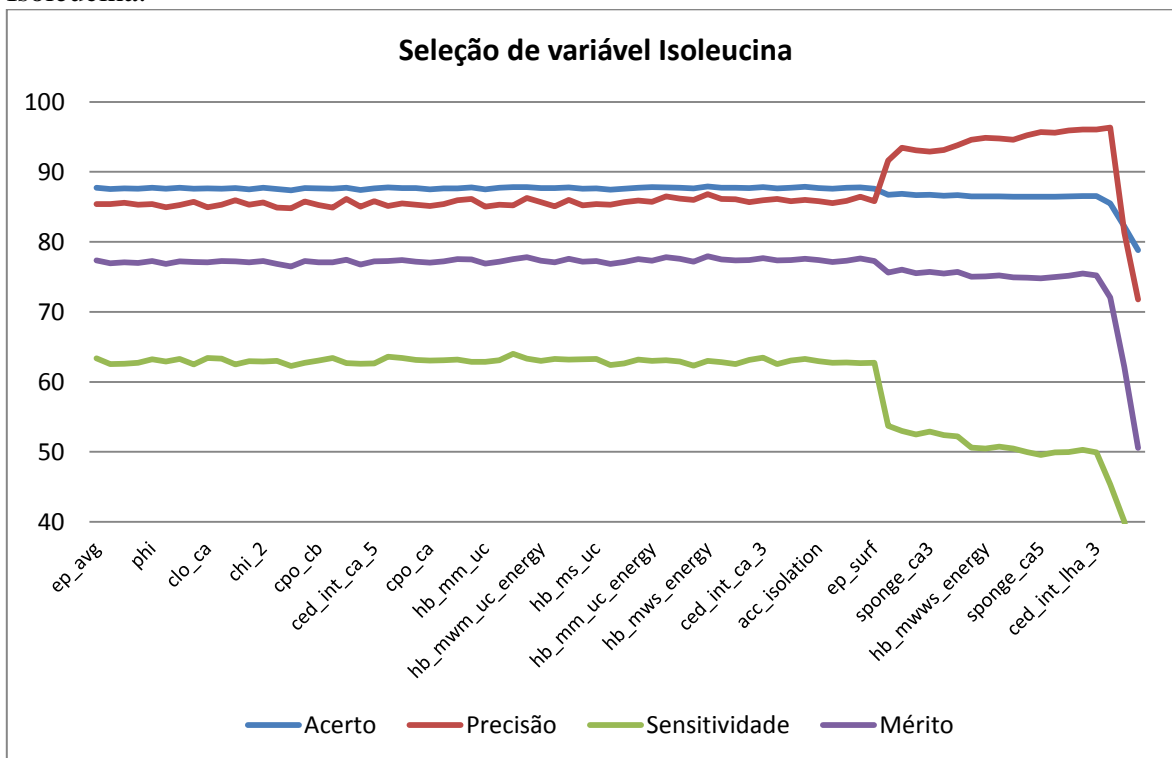
Glicina:



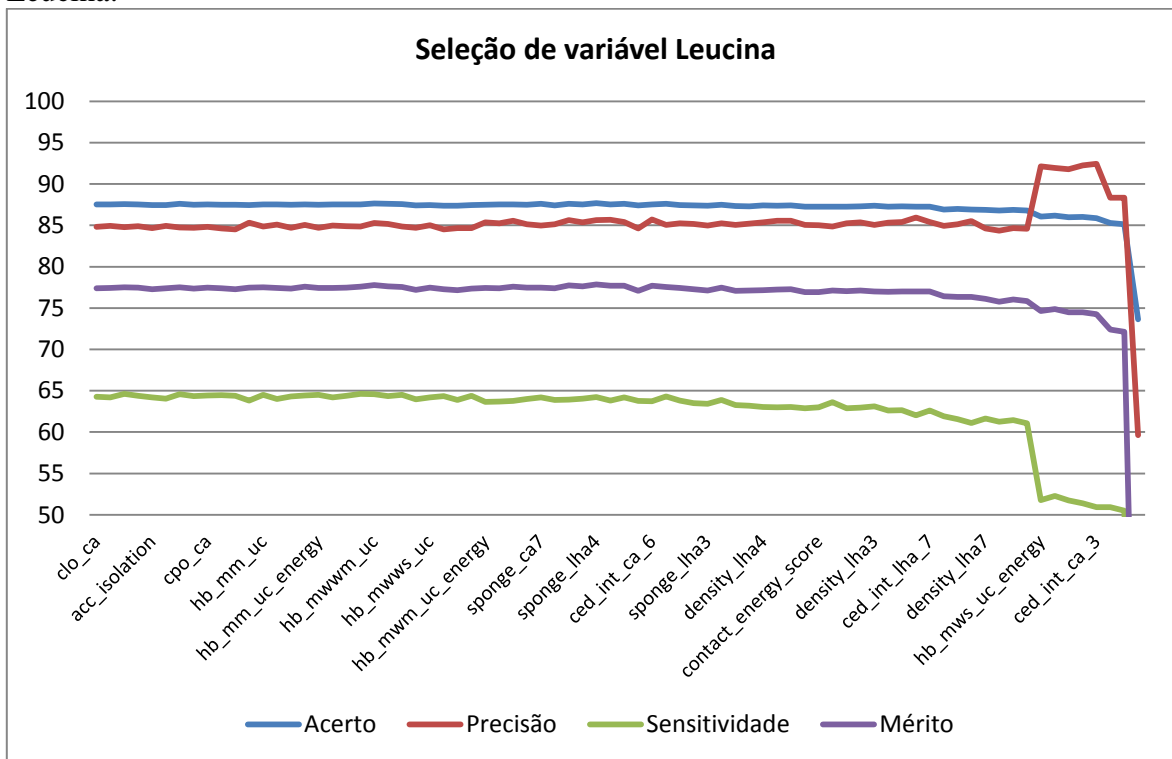
Histidina:



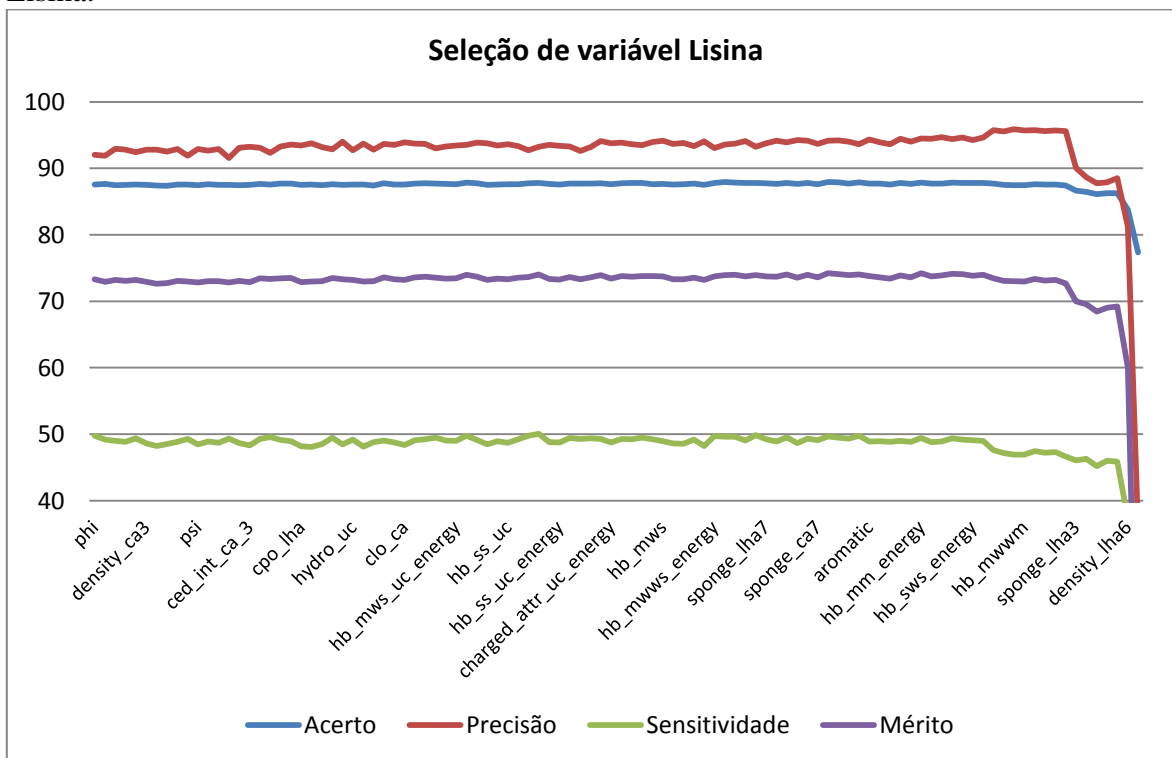
Isoleucina:



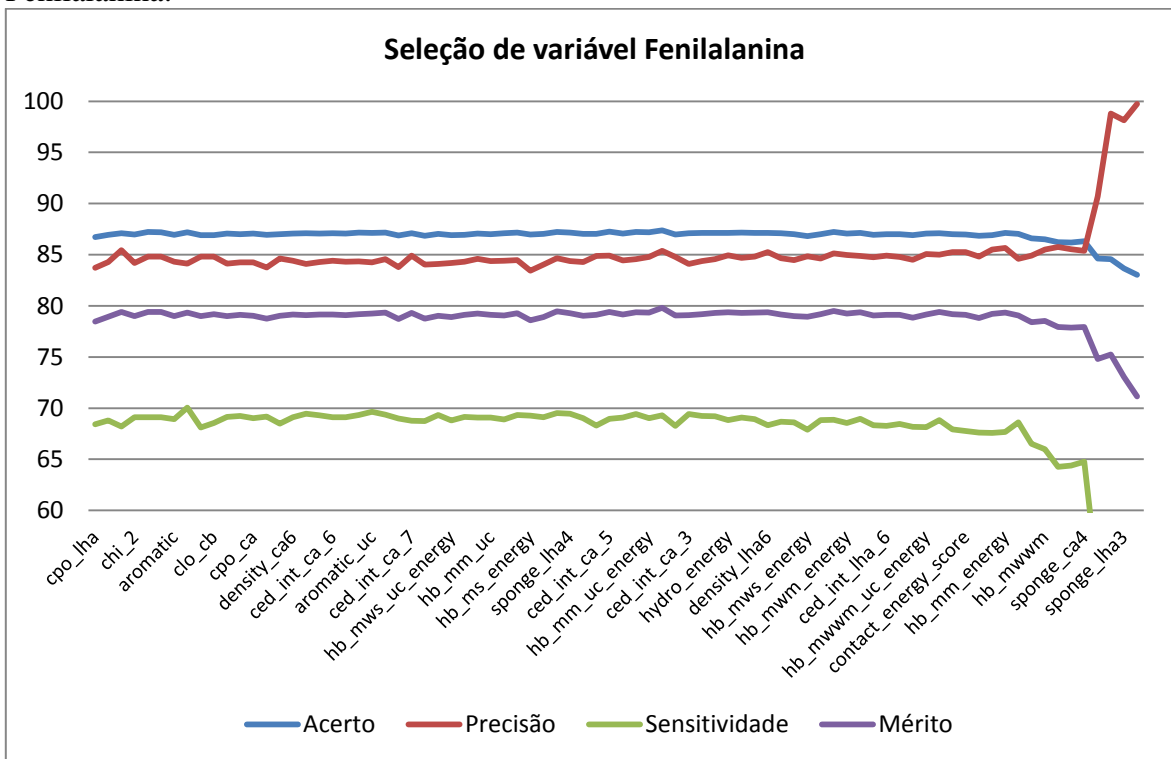
Leucina:



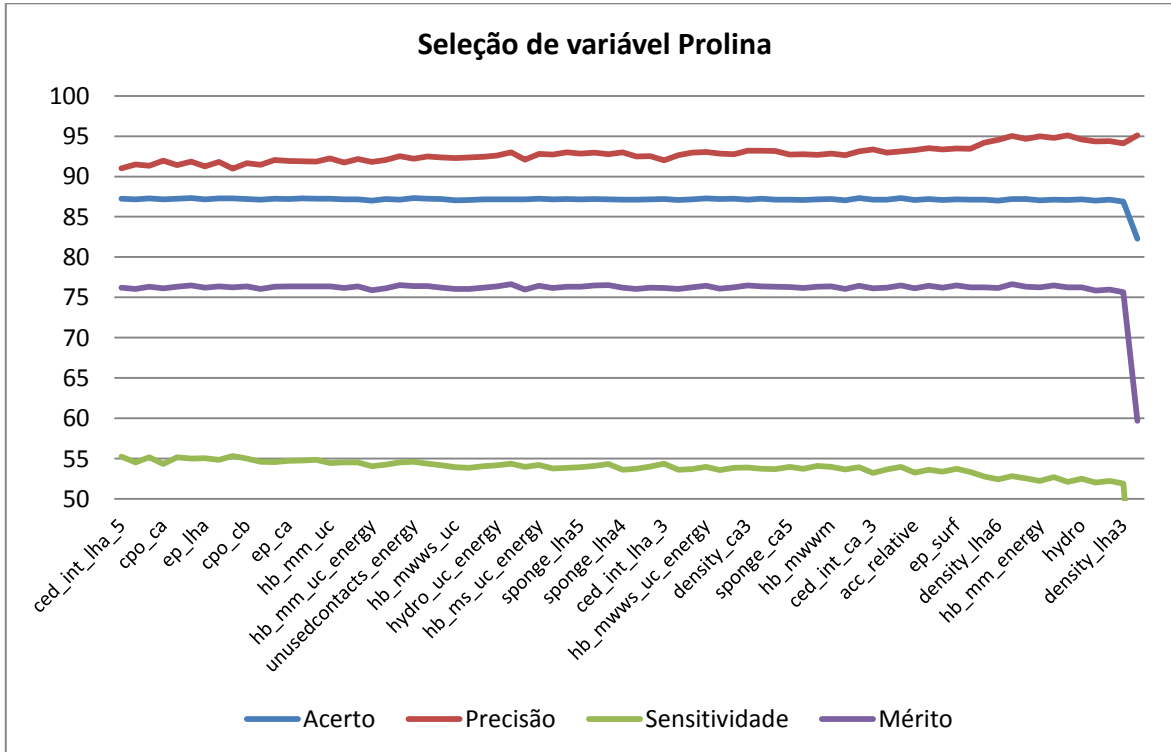
Lisina:



Fenilalanina:

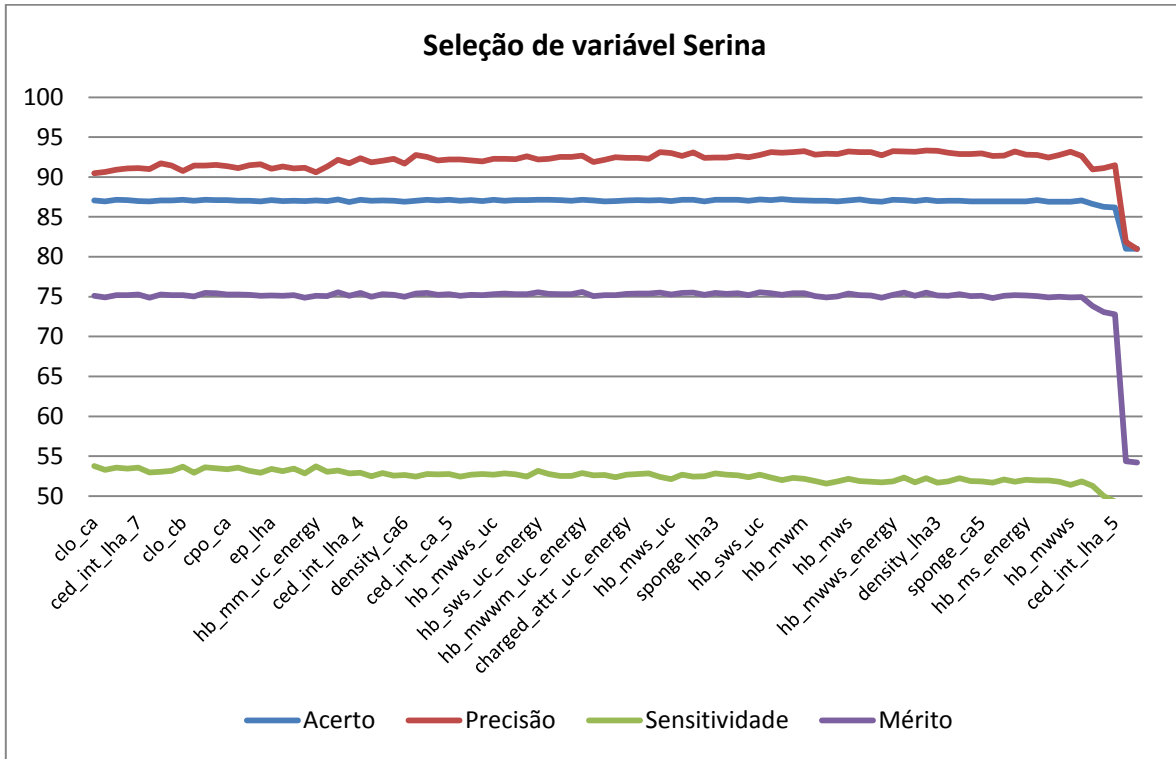


Prolina:

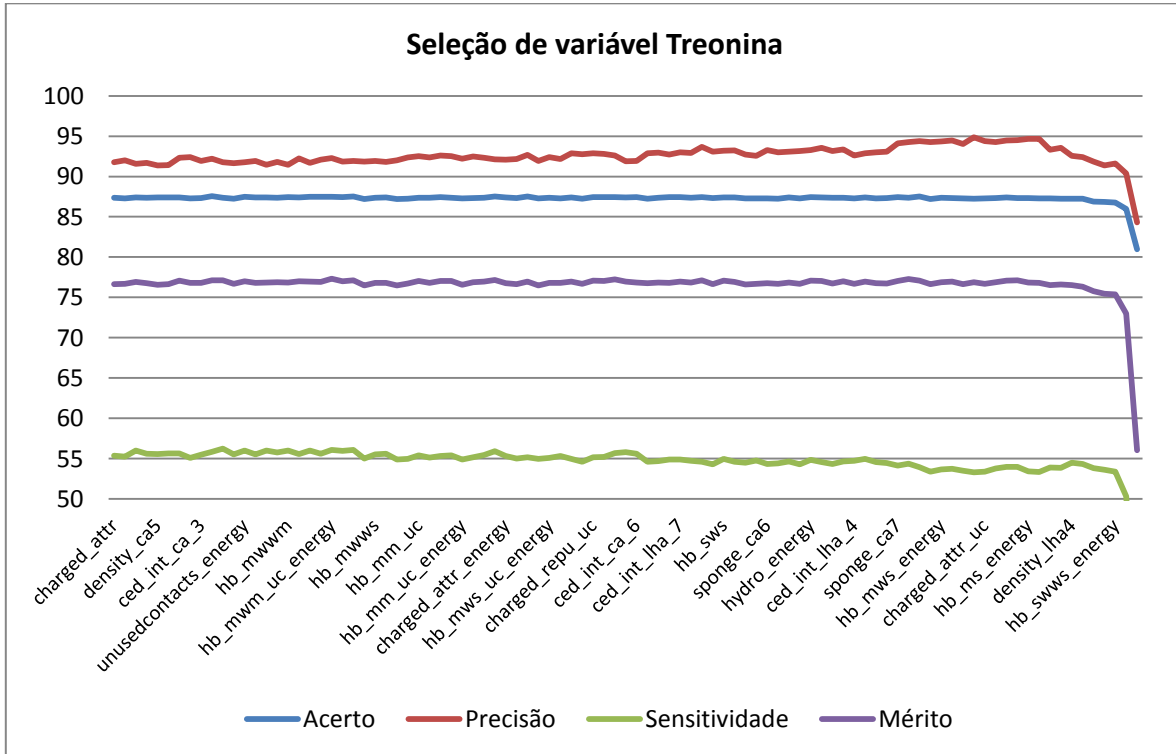




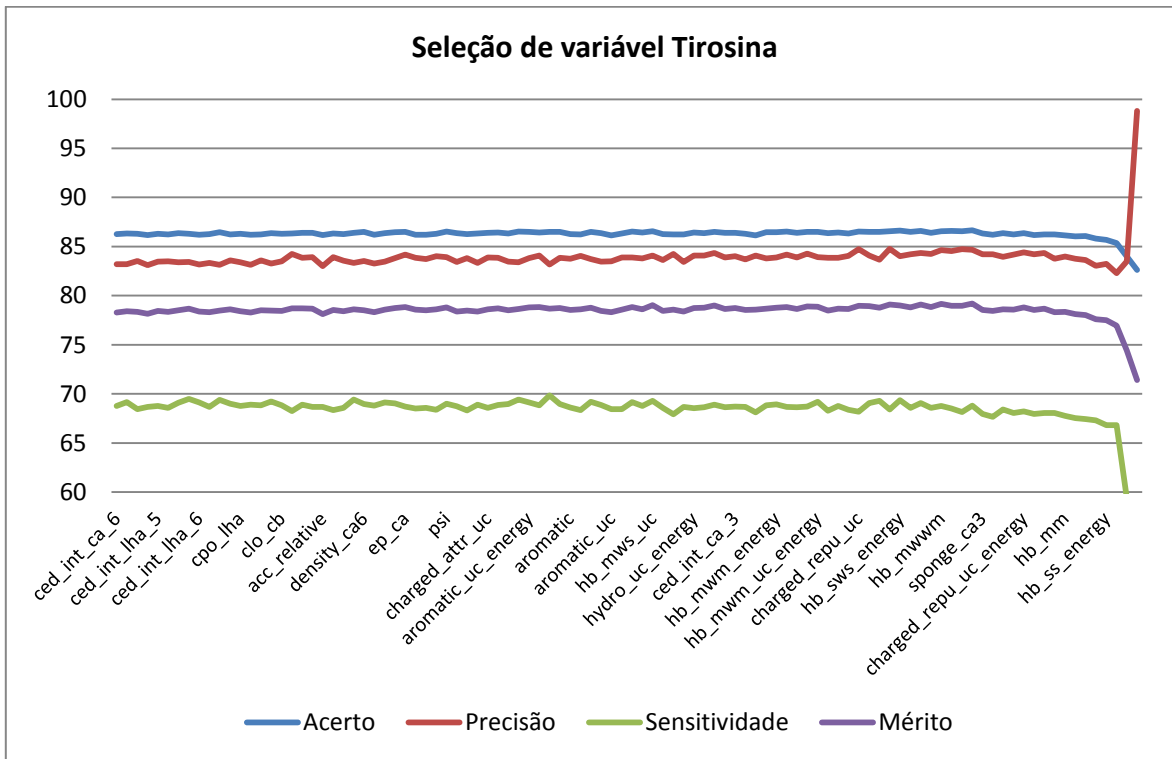
Serina:



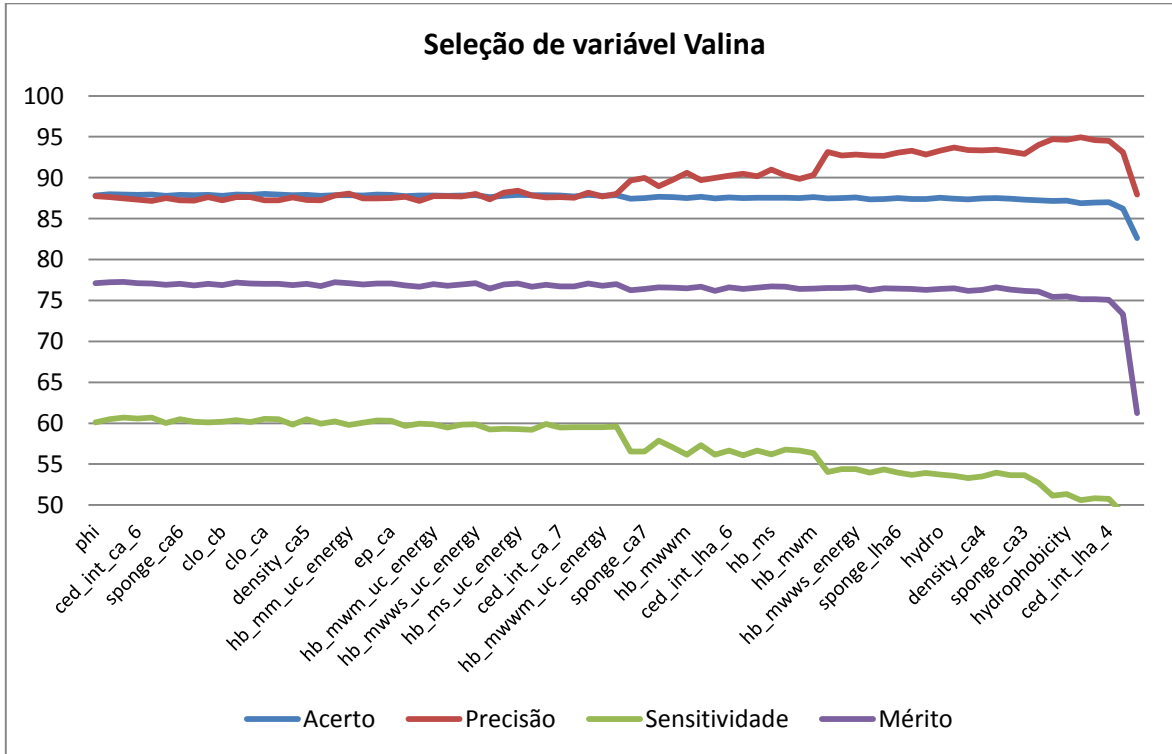
Treonina:



Tirosina:



Valina:



## **6.6 Apêndice 6 – Lista com a ordem dos descritores removidos no processo de seleção de variáveis**

Lista com a ordem dos descritores removidos no processo de seleção de variáveis para cada tipo de aminoácido e para o classificador inespecífico na síntese dos classificadores por rede neural, disponível no endereço:

<http://www.cbi.cnptia.embrapa.br/SMS/predictions/index.html>

**6.7 Apêndice 7 – Lista com os códigos PDB e cadeia de cada entrada utilizada como conjunto teste na metodologia de avaliação por holdout.**

PDB	chain	TP	TN	FP	FN	Acurácia	Precisão	Sensitividade	MCC
1A0D	A	153	210	9	15	0,9380	0,9107	0,9107	0,8737
1A25	A	20	86	1	5	0,9464	0,8000	0,8000	0,8412
1A2D	A	18	104	0	3	0,9760	0,8571	0,8571	0,9127
1A2Z	A	53	114	2	6	0,9543	0,8983	0,8983	0,8972
1A4E	A	182	241	10	15	0,9442	0,9239	0,9239	0,8867
1A4S	A	113	296	2	15	0,9601	0,8828	0,8828	0,9048
1ABR	A	44	164	5	8	0,9412	0,8462	0,8462	0,8338
1AF6	A	88	291	1	9	0,9743	0,9072	0,9072	0,9309
1AOG	A	77	345	5	6	0,9746	0,9277	0,9277	0,9177
1AQU	A	19	229	0	3	0,9880	0,8636	0,8636	0,9233
1AT3	A	28	146	3	6	0,9508	0,8235	0,8235	0,8330
1AX4	A	93	275	2	7	0,9761	0,9300	0,9300	0,9383
1AYM	1	131	117	6	16	0,9185	0,8912	0,8912	0,8391
1B3Q	A	41	293	0	3	0,9911	0,9318	0,9318	0,9604
1B4U	B	76	147	5	4	0,9612	0,9500	0,9500	0,9144
1B67	A	38	23	2	4	0,9104	0,9048	0,9048	0,8132
1B78	A	22	136	1	3	0,9753	0,8800	0,8800	0,9033
1BAI	A	43	71	3	1	0,9661	0,9773	0,9773	0,9288
1BBH	A	20	96	3	2	0,9587	0,9091	0,9091	0,8638
1BDQ	A	39	46	1	5	0,9341	0,8864	0,8864	0,8710
1BH9	B	37	46	2	0	0,9765	1,0000	1,0000	0,9535
1BND	A	38	56	3	10	0,8785	0,7917	0,7917	0,7579
1BO1	A	31	254	2	4	0,9794	0,8857	0,8857	0,9006
1BPL	B	66	170	3	6	0,9633	0,9167	0,9167	0,9108
1BTK	A	3	126	3	20	0,8487	0,1304	0,1304	0,1972
1BVZ	A	59	426	6	9	0,9700	0,8676	0,8676	0,8702
1BWW	A	21	75	0	2	0,9796	0,9130	0,9130	0,9430
1BYF	A	23	82	1	2	0,9722	0,9200	0,9200	0,9211
1C4T	A	66	127	4	7	0,9461	0,9041	0,9041	0,8821
1CAX	B	96	63	4	10	0,9191	0,9057	0,9057	0,8344
1CBI	A	16	109	1	2	0,9766	0,8889	0,8889	0,9012
1CCW	A	28	87	1	3	0,9664	0,9032	0,9032	0,9118
1CCW	B	73	286	4	9	0,9651	0,8902	0,8902	0,8968
1CH4	A	44	87	3	2	0,9632	0,9565	0,9565	0,9184
1CJX	A	38	264	4	6	0,9679	0,8636	0,8636	0,8655
1CM7	A	49	241	2	7	0,9699	0,8750	0,8750	0,8991
1CNZ	A	58	237	3	5	0,9736	0,9206	0,9206	0,9191
1CZY	A	47	91	2	9	0,9262	0,8393	0,8393	0,8431
1DDR	A	14	128	0	4	0,9726	0,7778	0,7778	0,8685
1DGR	N	63	9	1	3	0,9474	0,9545	0,9545	0,7922
1DHK	B	44	131	2	3	0,9722	0,9362	0,9362	0,9276
1DJ7	B	19	48	0	3	0,9571	0,8636	0,8636	0,9016
1DJR	D	62	29	1	5	0,9381	0,9254	0,9254	0,8641
1DK0	A	27	122	0	8	0,9490	0,7714	0,7714	0,8509
1DM5	A	69	197	3	9	0,9568	0,8846	0,8846	0,8918
1DPG	A	55	353	8	9	0,9600	0,8594	0,8594	0,8427
1DPS	A	77	47	2	11	0,9051	0,8750	0,8750	0,8092

1DPT	A	16	89	0	4	0,9633	0,8000	0,8000	0,8750
1DTD	A	24	213	7	6	0,9480	0,8000	0,8000	0,7574
1DUG	A	4	164	1	36	0,8195	0,1000	0,1000	0,2413
1E3I	A	36	256	5	6	0,9637	0,8571	0,8571	0,8465
1E5D	A	45	271	8	8	0,9518	0,8491	0,8491	0,8204
1E6C	A	35	107	4	9	0,9161	0,7955	0,7955	0,7891
1E6V	C	103	98	4	12	0,9263	0,8957	0,8957	0,8550
1E6W	A	81	119	4	13	0,9217	0,8617	0,8617	0,8417
1E6Y	B	175	160	7	11	0,9490	0,9409	0,9409	0,8981
1ECM	A	50	36	1	3	0,9556	0,9434	0,9434	0,9099
1ECP	A	73	121	3	4	0,9652	0,9481	0,9481	0,9262
1EE0	A	54	241	2	3	0,9833	0,9474	0,9474	0,9455
1EEQ	A	17	80	1	3	0,9604	0,8500	0,8500	0,8723
1EFU	A	69	236	7	10	0,9472	0,8734	0,8734	0,8559
1EFU	B	78	160	3	9	0,9520	0,8966	0,8966	0,8938
1EFV	A	87	173	4	8	0,9559	0,9158	0,9158	0,9025
1EFV	B	75	130	2	12	0,9361	0,8621	0,8621	0,8680
1EHK	B	76	63	0	7	0,9521	0,9157	0,9157	0,9078
1EJB	A	66	71	2	6	0,9448	0,9167	0,9167	0,8910
1EJO	H	47	137	2	9	0,9436	0,8393	0,8393	0,8605
1EK6	A	22	266	0	5	0,9829	0,8148	0,8148	0,8943
1EL6	A	103	78	5	9	0,9282	0,9196	0,9196	0,8548
1ELU	A	76	233	4	8	0,9626	0,9048	0,9048	0,9022
1EM9	A	8	121	0	2	0,9847	0,8000	0,8000	0,8871
1EMU	A	18	93	1	3	0,9652	0,8571	0,8571	0,8806
1EO6	A	19	85	2	1	0,9720	0,9500	0,9500	0,9099
1EPT	B	61	14	2	5	0,9146	0,9242	0,9242	0,7507
1ES7	B	25	47	2	4	0,9231	0,8621	0,8621	0,8343
1ESM	A	41	228	1	8	0,9676	0,8367	0,8367	0,8856
1ETO	A	51	26	5	3	0,9059	0,9444	0,9444	0,7951
1ETP	A	14	164	0	3	0,9834	0,8235	0,8235	0,8993
1EVH	A	11	88	0	2	0,9802	0,8462	0,8462	0,9096
1EXZ	A	27	91	1	4	0,9593	0,8710	0,8710	0,8906
1EYM	A	21	72	1	2	0,9688	0,9130	0,9130	0,9133
1EYV	A	26	80	3	6	0,9217	0,8125	0,8125	0,8012
1EZ4	A	87	159	3	10	0,9498	0,8969	0,8969	0,8928
1F1X	A	60	197	4	8	0,9554	0,8824	0,8824	0,8803
1F2T	A	46	78	1	7	0,9394	0,8679	0,8679	0,8756
1F2T	B	55	69	3	10	0,9051	0,8462	0,8462	0,8130
1F45	B	24	98	2	1	0,9760	0,9600	0,9600	0,9264
1F56	B	7	76	1	2	0,9651	0,7778	0,7778	0,8060
1F58	H	49	124	0	11	0,9402	0,8167	0,8167	0,8661
1F60	A	43	322	3	10	0,9656	0,8113	0,8113	0,8518
1F75	A	33	151	1	6	0,9634	0,8462	0,8462	0,8848
1F77	A	18	161	0	5	0,9728	0,7826	0,7826	0,8712
1F8M	A	185	174	9	16	0,9349	0,9204	0,9204	0,8703
1F9A	A	48	88	1	4	0,9645	0,9231	0,9231	0,9239
1FBQ	A	9	72	2	1	0,9643	0,9000	0,9000	0,8380
1FBV	C	18	115	5	4	0,9366	0,8182	0,8182	0,7626
1FM7	A	24	155	3	5	0,9572	0,8276	0,8276	0,8328
1FTH	A	27	73	2	6	0,9259	0,8182	0,8182	0,8227
1FUX	A	5	116	2	17	0,8643	0,2273	0,2273	0,3512
1FVP	A	35	165	2	8	0,9524	0,8140	0,8140	0,8494
1FXR	A	10	49	0	3	0,9516	0,7692	0,7692	0,8514

1FXZ	A	119	181	7	23	0,9091	0,8380	0,8380	0,8161
1FYF	A	47	300	3	6	0,9747	0,8868	0,8868	0,8984
1FZG	B	85	161	5	13	0,9318	0,8673	0,8673	0,8533
1G08	A	38	92	1	4	0,9630	0,9048	0,9048	0,9131
1G20	E	27	111	30	51	0,6301	0,3462	0,3462	0,1456
1G2L	B	21	31	1	2	0,9455	0,9130	0,9130	0,8878
1G2O	A	44	183	1	8	0,9619	0,8462	0,8462	0,8870
1G4M	A	17	299	2	3	0,9844	0,8500	0,8500	0,8638
1G57	A	34	135	5	6	0,9389	0,8500	0,8500	0,8217
1G73	A	49	95	4	5	0,9412	0,9074	0,9074	0,8708
1G88	A	46	136	4	7	0,9430	0,8679	0,8679	0,8551
1G8T	A	23	175	1	4	0,9754	0,8519	0,8519	0,8899
1G99	A	70	253	7	6	0,9613	0,9211	0,9211	0,8900
1G9M	C	23	136	1	4	0,9695	0,8519	0,8519	0,8861
1G9M	G	46	222	5	14	0,9338	0,7667	0,7667	0,7921
1GC0	A	84	201	2	9	0,9628	0,9032	0,9032	0,9134
1GCJ	A	0	6	0	0	1,0000	#DIV/0!	#DIV/0!	#DIV/0!
1GD9	A	69	237	4	3	0,9776	0,9583	0,9583	0,9372
1GG1	A	78	189	2	6	0,9709	0,9286	0,9286	0,9311
1GHE	A	11	146	0	1	0,9937	0,9167	0,9167	0,9542
1GLQ	A	29	148	3	4	0,9620	0,8788	0,8788	0,8694
1GOT	B	102	166	5	18	0,9210	0,8500	0,8500	0,8379
1GPU	A	111	425	6	10	0,9710	0,9174	0,9174	0,9145
1GQA	A	19	97	2	3	0,9587	0,8636	0,8636	0,8590
1GSU	A	28	161	0	7	0,9643	0,8000	0,8000	0,8756
1GTZ	A	56	59	3	3	0,9504	0,9492	0,9492	0,9008
1GU7	A	22	291	1	5	0,9812	0,8148	0,8148	0,8732
1GUD	A	20	213	2	5	0,9708	0,8000	0,8000	0,8371
1GUX	B	41	81	1	1	0,9839	0,9762	0,9762	0,9640
1GVE	A	24	247	2	1	0,9891	0,9600	0,9600	0,9354
1GVJ	A	30	97	1	4	0,9621	0,8824	0,8824	0,8997
1GWW	A	19	221	2	6	0,9677	0,7600	0,7600	0,8122
1GXS	A	127	108	1	13	0,9438	0,9071	0,9071	0,8913
1GXS	B	106	41	2	5	0,9545	0,9550	0,9550	0,8904
1GYO	A	15	87	2	1	0,9714	0,9375	0,9375	0,8927
1H1Y	A	24	155	0	6	0,9676	0,8000	0,8000	0,8776
1H2T	C	48	546	3	5	0,9867	0,9057	0,9057	0,9160
1H3I	A	41	215	1	12	0,9517	0,7736	0,7736	0,8426
1H7Z	A	41	113	4	9	0,9222	0,8200	0,8200	0,8112
1H80	A	19	330	5	3	0,9776	0,8636	0,8636	0,8150
1H8T	C	131	87	3	8	0,9520	0,9424	0,9424	0,9012
1H8X	A	44	68	3	4	0,9412	0,9167	0,9167	0,8775
1H90	A	11	78	3	3	0,9368	0,7857	0,7857	0,7487
1HFE	S	61	15	3	3	0,9268	0,9531	0,9531	0,7865
1HK7	A	20	195	2	7	0,9598	0,7407	0,7407	0,7993
1HN4	A	13	95	1	8	0,9231	0,6190	0,6190	0,7196
1HQO	A	47	150	1	2	0,9850	0,9592	0,9592	0,9593
1HR6	B	51	319	6	9	0,9610	0,8500	0,8500	0,8492
1HWM	A	40	176	2	4	0,9730	0,9091	0,9091	0,9139
1HWM	B	35	180	1	13	0,9389	0,7292	0,7292	0,8092
1HXH	A	73	125	1	12	0,9384	0,8588	0,8588	0,8746
1HXS	1	141	121	6	8	0,9493	0,9463	0,9463	0,8981
1HYR	A	43	64	2	2	0,9640	0,9556	0,9556	0,9253
1IOA	A	160	243	5	9	0,9664	0,9467	0,9467	0,9303

1I0R	A	54	79	3	8	0,9236	0,8710	0,8710	0,8448
1I4S	A	28	90	3	4	0,9440	0,8750	0,8750	0,8517
1IAK	B	66	98	5	5	0,9425	0,9296	0,9296	0,8810
1IAR	B	17	144	3	7	0,9415	0,7083	0,7083	0,7435
1IAU	A	9	159	2	4	0,9655	0,6923	0,6923	0,7346
1ICF	A	91	57	1	15	0,9024	0,8585	0,8585	0,8105
1ICF	I	22	42	1	0	0,9846	1,0000	1,0000	0,9666
1IED	A	38	145	4	3	0,9632	0,9268	0,9268	0,8922
1IIN	A	52	182	4	7	0,9551	0,8814	0,8814	0,8755
1IJG	A	129	92	6	12	0,9247	0,9149	0,9149	0,8469
1IJL	A	21	86	2	5	0,9386	0,8077	0,8077	0,8207
1IKN	A	39	212	4	8	0,9544	0,8298	0,8298	0,8404
1INL	A	79	154	3	8	0,9549	0,9080	0,9080	0,9014
1IQB	A	7	73	0	1	0,9877	0,8750	0,8750	0,9291
1IS2	A	133	389	8	10	0,9667	0,9301	0,9301	0,9140
1ISA	A	21	143	1	2	0,9820	0,9130	0,9130	0,9233
1ISI	A	27	186	1	6	0,9682	0,8182	0,8182	0,8709
1IWE	A	68	292	5	10	0,9600	0,8718	0,8718	0,8764
1IWP	B	45	100	3	4	0,9539	0,9184	0,9184	0,8941
1IXM	A	40	111	2	1	0,9805	0,9756	0,9756	0,9507
1J1A	A	11	100	2	3	0,9569	0,7857	0,7857	0,7912
1J3M	A	33	76	2	7	0,9237	0,8250	0,8250	0,8284
1J3Q	A	45	123	2	4	0,9655	0,9184	0,9184	0,9141
1J49	A	63	204	6	7	0,9536	0,9000	0,9000	0,8756
1J4X	A	13	127	5	1	0,9589	0,9286	0,9286	0,7977
1J7D	A	17	108	0	2	0,9843	0,8947	0,8947	0,9373
1J7D	B	17	116	1	6	0,9500	0,7391	0,7391	0,8087
1J7N	A	32	610	7	7	0,9787	0,8205	0,8205	0,8092
1J8H	D	58	111	2	11	0,9286	0,8406	0,8406	0,8493
1J8H	E	60	140	4	11	0,9302	0,8451	0,8451	0,8406
1J95	A	40	46	2	4	0,9348	0,9091	0,9091	0,8699
1JEQ	A	217	252	13	15	0,9437	0,9353	0,9353	0,8868
1J17	A	14	56	1	0	0,9859	1,0000	1,0000	0,9576
1J1F	A	39	70	1	2	0,9732	0,9512	0,9512	0,9422
1JK0	B	30	202	0	5	0,9789	0,8571	0,8571	0,9146
1JKG	A	50	73	3	6	0,9318	0,8929	0,8929	0,8604
1JMT	A	21	67	2	1	0,9670	0,9545	0,9545	0,9118
1JNR	A	156	317	3	18	0,9575	0,8966	0,8966	0,9071
1JR8	A	25	67	1	7	0,9200	0,7813	0,7813	0,8152
1JRH	I	17	71	2	0	0,9778	1,0000	1,0000	0,9329
1JTD	A	28	179	2	1	0,9857	0,9655	0,9655	0,9410
1JTH	A	46	12	2	2	0,9355	0,9583	0,9583	0,8155
1JUO	A	39	115	2	6	0,9506	0,8667	0,8667	0,8752
1JUQ	A	11	90	1	31	0,7594	0,2619	0,2619	0,4071
1JVK	A	49	138	1	10	0,9444	0,8305	0,8305	0,8667
1JWI	B	37	68	2	7	0,9211	0,8409	0,8409	0,8336
1JWY	A	19	642	0	1	0,9985	0,9500	0,9500	0,9739
1JY5	A	9	172	1	4	0,9731	0,6923	0,6923	0,7761
1JYA	A	29	77	1	7	0,9298	0,8056	0,8056	0,8368
1JZT	A	34	160	5	3	0,9604	0,9189	0,9189	0,8708
1K0Z	A	33	96	2	7	0,9348	0,8250	0,8250	0,8390
1K3E	A	29	100	1	6	0,9485	0,8286	0,8286	0,8632
1K3Y	A	34	161	4	3	0,9653	0,9189	0,9189	0,8855
1K55	A	43	162	3	5	0,9624	0,8958	0,8958	0,8911

1K5J	A	39	42	0	6	0,9310	0,8667	0,8667	0,8708
1K8K	A	80	247	6	14	0,9424	0,8511	0,8511	0,8516
1K8K	B	49	108	5	9	0,9181	0,8448	0,8448	0,8154
1K8K	C	48	241	3	7	0,9666	0,8727	0,8727	0,8863
1K8K	D	78	169	3	11	0,9464	0,8764	0,8764	0,8801
1KA9	H	32	123	1	4	0,9688	0,8889	0,8889	0,9091
1KCF	A	20	186	1	4	0,9763	0,8333	0,8333	0,8781
1KDZ	A	11	259	1	5	0,9783	0,6875	0,6875	0,7834
1KEL	H	41	145	5	11	0,9208	0,7885	0,7885	0,7873
1KFU	L	67	571	4	6	0,9846	0,9178	0,9178	0,9220
1KIJ	A	52	274	3	11	0,9588	0,8254	0,8254	0,8595
1KIL	C	53	16	0	2	0,9718	0,9636	0,9636	0,9255
1KJM	A	69	180	6	3	0,9651	0,9583	0,9583	0,9148
1KJN	A	25	105	1	4	0,9630	0,8621	0,8621	0,8880
1KKC	A	46	127	0	3	0,9830	0,9388	0,9388	0,9577
1KNO	A	122	36	4	9	0,9240	0,9313	0,9313	0,7992
1KNX	A	99	163	4	7	0,9597	0,9340	0,9340	0,9150
1KOF	A	26	123	1	6	0,9551	0,8125	0,8125	0,8586
1KOR	A	115	187	18	11	0,9124	0,9127	0,9127	0,8170
1KPE	A	42	47	5	7	0,8812	0,8571	0,8571	0,7625
1KPS	B	28	108	1	4	0,9645	0,8750	0,8750	0,8972
1KSI	A	174	377	8	25	0,9435	0,8744	0,8744	0,8735
1KU5	A	39	22	0	3	0,9531	0,9286	0,9286	0,9040
1KUK	A	16	146	2	2	0,9759	0,8889	0,8889	0,8754
1L1O	B	48	64	0	1	0,9912	0,9796	0,9796	0,9821
1L3A	A	63	82	1	5	0,9603	0,9265	0,9265	0,9206
1L3C	A	32	120	3	6	0,9441	0,8421	0,8421	0,8418
1LBV	A	29	174	2	3	0,9760	0,9063	0,9063	0,9066
1LCK	A	19	130	0	3	0,9803	0,8636	0,8636	0,9188
1LGQ	A	65	40	1	5	0,9459	0,9286	0,9286	0,8888
1LI1	A	121	64	3	12	0,9250	0,9098	0,9098	0,8411
1LNI	A	2	67	0	0	1,0000	1,0000	1,0000	1,0000
1LPF	A	85	300	5	15	0,9506	0,8500	0,8500	0,8646
1LRW	A	111	341	2	21	0,9516	0,8409	0,8409	0,8786
1LUC	A	50	224	2	8	0,9648	0,8621	0,8621	0,8894
1LUC	B	51	217	0	9	0,9675	0,8500	0,8500	0,9034
1LVW	A	13	157	0	71	0,7054	0,1548	0,1548	0,3264
1LXD	A	10	69	0	0	1,0000	1,0000	1,0000	1,0000
1LYA	A	72	17	1	5	0,9368	0,9351	0,9351	0,8171
1M0S	A	27	154	1	5	0,9679	0,8438	0,8438	0,8838
1M26	A	38	77	3	6	0,9274	0,8636	0,8636	0,8402
1M2D	A	14	72	0	3	0,9663	0,8235	0,8235	0,8892
1M5E	A	36	187	0	3	0,9867	0,9231	0,9231	0,9532
1M6J	A	39	175	2	5	0,9683	0,8864	0,8864	0,8989
1M93	B	82	130	7	20	0,8870	0,8039	0,8039	0,7702
1MEE	A	25	175	2	5	0,9662	0,8333	0,8333	0,8592
1MGQ	A	40	19	3	8	0,8429	0,8333	0,8333	0,6647
1MK4	A	30	102	2	4	0,9565	0,8824	0,8824	0,8813
1MO1	A	48	23	2	6	0,8987	0,8889	0,8889	0,7805
1MPX	A	83	438	4	18	0,9595	0,8218	0,8218	0,8621
1MSA	A	46	46	3	5	0,9200	0,9020	0,9020	0,8407
1MTY	B	172	131	10	18	0,9154	0,9053	0,9053	0,8293
1MTY	G	72	72	1	4	0,9664	0,9474	0,9474	0,9337
1MU2	B	67	278	4	7	0,9691	0,9054	0,9054	0,9051



1MUJ	B	70	94	3	6	0,9480	0,9211	0,9211	0,8945
1MWQ	A	30	53	0	9	0,9022	0,7692	0,7692	0,8109
1MY6	A	20	140	1	3	0,9756	0,8696	0,8696	0,8963
1MYP	C	154	256	11	19	0,9318	0,8902	0,8902	0,8566
1MZH	A	36	141	3	2	0,9725	0,9474	0,9474	0,9178
1N1A	A	18	82	3	2	0,9524	0,9000	0,9000	0,8489
1N1E	A	65	217	5	6	0,9625	0,9155	0,9155	0,8973
1N1J	A	50	28	3	4	0,9176	0,9259	0,9259	0,8238
1N1J	B	46	24	3	3	0,9211	0,9388	0,9388	0,8277
1N2A	A	41	127	0	4	0,9767	0,9111	0,9111	0,9398
1N62	C	45	186	4	9	0,9467	0,8333	0,8333	0,8417
1N7S	A	49	6	4	4	0,8730	0,9245	0,9245	0,5245
1N7S	C	53	20	1	5	0,9241	0,9138	0,9138	0,8227
1NC7	A	42	46	1	5	0,9362	0,8936	0,8936	0,8755
1NCQ	C	123	77	3	16	0,9132	0,8849	0,8849	0,8255
1NDO	A	109	263	6	11	0,9563	0,9083	0,9083	0,8968
1NE7	A	44	175	4	13	0,9280	0,7719	0,7719	0,7970
1NEY	A	39	161	2	6	0,9615	0,8667	0,8667	0,8843
1NF3	C	9	78	7	20	0,7632	0,3103	0,3103	0,2859
1NFD	B	78	120	6	17	0,8959	0,8211	0,8211	0,7888
1NFU	B	15	28	1	5	0,8776	0,7500	0,7500	0,7499
1NGK	A	38	76	1	0	0,9913	1,0000	1,0000	0,9807
1NLN	A	28	138	2	0	0,9881	1,0000	1,0000	0,9592
1NR4	A	20	38	3	4	0,8923	0,8333	0,8333	0,7672
1NTV	A	17	104	3	8	0,9167	0,6800	0,6800	0,7125
1NU0	A	11	107	0	0	1,0000	1,0000	1,0000	1,0000
1NWP	A	14	95	1	2	0,9732	0,8750	0,8750	0,8883
1NWW	A	34	91	4	7	0,9191	0,8293	0,8293	0,8050
1NXU	A	72	200	4	17	0,9283	0,8090	0,8090	0,8282
1NYR	A	57	495	15	6	0,9634	0,9048	0,9048	0,8261
1O0V	A	72	208	6	16	0,9272	0,8182	0,8182	0,8203
1O26	A	83	89	3	14	0,9101	0,8557	0,8557	0,8261
1O5E	H	29	158	4	6	0,9492	0,8286	0,8286	0,8228
1O5E	L	25	76	1	2	0,9712	0,9259	0,9259	0,9243
1O5O	A	72	90	4	9	0,9257	0,8889	0,8889	0,8514
1O5X	A	41	158	1	6	0,9660	0,8723	0,8723	0,9021
1O75	A	50	292	4	11	0,9580	0,8197	0,8197	0,8469
1O7D	A	127	127	6	7	0,9513	0,9478	0,9478	0,9026
1O7D	B	71	0	0	5	0,9342	0,9342	0,9342	#DIV/0!
1O7D	E	56	40	4	6	0,9057	0,9032	0,9032	0,8076
1O81	A	18	111	0	5	0,9627	0,7826	0,7826	0,8654
1O8B	A	24	123	1	6	0,9545	0,8000	0,8000	0,8505
1O9P	A	38	277	3	7	0,9692	0,8444	0,8444	0,8673
1O8B	A	16	79	2	5	0,9314	0,7619	0,7619	0,7819
1OFW	A	34	239	2	10	0,9579	0,7727	0,7727	0,8314
1OFZ	A	26	244	2	0	0,9926	1,0000	1,0000	0,9597
1OI2	A	32	225	2	8	0,9625	0,8000	0,8000	0,8470
1OI6	A	38	135	2	9	0,9402	0,8085	0,8085	0,8394
1OJ5	A	11	87	0	5	0,9515	0,6875	0,6875	0,8063
1OJ7	A	9	222	2	86	0,7241	0,0947	0,0947	0,2151
1ON2	A	29	93	0	3	0,9760	0,9063	0,9063	0,9370
1ONK	A	42	170	0	5	0,9770	0,8936	0,8936	0,9317
1OOP	A	132	117	2	11	0,9504	0,9231	0,9231	0,9027
1OOP	B	84	119	6	20	0,8865	0,8077	0,8077	0,7744

1OOP	C	130	81	3	10	0,9420	0,9286	0,9286	0,8801
1OR4	A	38	114	2	4	0,9620	0,9048	0,9048	0,9017
1OSC	A	59	36	3	2	0,9500	0,9672	0,9672	0,8946
1OUL	A	14	94	1	2	0,9730	0,8750	0,8750	0,8882
1OVM	A	115	298	3	22	0,9429	0,8394	0,8394	0,8668
1OXQ	A	38	44	2	6	0,9111	0,8636	0,8636	0,8251
1OZH	A	101	326	6	12	0,9596	0,8938	0,8938	0,8919
1POF	A	44	251	3	4	0,9768	0,9167	0,9167	0,9126
1P22	A	55	286	2	7	0,9743	0,8871	0,8871	0,9101
1P4R	A	148	336	10	11	0,9584	0,9308	0,9308	0,9035
1P5T	A	15	78	0	2	0,9789	0,8824	0,8824	0,9275
1P65	A	35	18	2	2	0,9298	0,9459	0,9459	0,8459
1P7L	A	94	209	4	15	0,9410	0,8624	0,8624	0,8676
1P8C	A	62	34	2	5	0,9320	0,9254	0,9254	0,8551
1PC6	A	64	63	4	5	0,9338	0,9275	0,9275	0,8677
1PCX	A	17	620	5	1	0,9907	0,9444	0,9444	0,8498
1PD2	1	29	148	1	2	0,9833	0,9355	0,9355	0,9410
1PDK	B	38	95	4	6	0,9301	0,8636	0,8636	0,8342
1PGU	A	23	481	2	5	0,9863	0,8214	0,8214	0,8623
1PK6	A	46	68	1	7	0,9344	0,8679	0,8679	0,8693
1PK6	B	47	70	1	5	0,9512	0,9038	0,9038	0,9011
1PKV	A	20	60	1	3	0,9524	0,8696	0,8696	0,8786
1PLG	H	38	152	1	12	0,9360	0,7600	0,7600	0,8240
1POV	0	106	155	6	9	0,9457	0,9217	0,9217	0,8880
1PP9	F	49	37	4	8	0,8776	0,8596	0,8596	0,7544
1PP9	G	51	21	1	2	0,9600	0,9623	0,9623	0,9052
1PP9	H	27	30	1	7	0,8769	0,7941	0,7941	0,7684
1PP9	J	36	23	1	2	0,9516	0,9474	0,9474	0,8993
1PPL	E	10	259	2	5	0,9746	0,6667	0,6667	0,7326
1PPV	A	22	135	0	0	1,0000	1,0000	1,0000	1,0000
1PQ1	A	28	98	1	1	0,9844	0,9655	0,9655	0,9554
1PS1	A	22	245	3	3	0,9780	0,8800	0,8800	0,8679
1PVM	A	8	109	3	37	0,7452	0,1778	0,1778	0,2675
1PWX	A	92	119	5	6	0,9505	0,9388	0,9388	0,8995
1PXR	A	17	181	2	3	0,9754	0,8500	0,8500	0,8585
1Q06	A	42	67	3	6	0,9237	0,8750	0,8750	0,8416
1Q0E	A	18	114	0	3	0,9778	0,8571	0,8571	0,9139
1Q5X	A	25	98	2	6	0,9389	0,8065	0,8065	0,8263
1Q7Z	A	25	410	1	4	0,9886	0,8621	0,8621	0,9046
1Q95	G	42	96	4	2	0,9583	0,9545	0,9545	0,9035
1QD6	C	57	156	0	16	0,9301	0,7808	0,7808	0,8415
1QGE	D	72	122	3	4	0,9652	0,9474	0,9474	0,9258
1QGE	E	67	22	3	4	0,9271	0,9437	0,9437	0,8134
1QGU	A	131	247	4	12	0,9594	0,9161	0,9161	0,9120
1QHF	A	23	183	1	1	0,9904	0,9583	0,9583	0,9529
1QHH	B	187	51	6	17	0,9119	0,9167	0,9167	0,7638
1QHH	C	49	44	2	6	0,9208	0,8909	0,8909	0,8441
1QHH	D	73	16	4	1	0,9468	0,9865	0,9865	0,8363
1QHI	A	43	228	2	6	0,9713	0,8776	0,8776	0,8989
1QIN	A	91	63	5	9	0,9167	0,9100	0,9100	0,8297
1QKS	A	36	400	4	10	0,9689	0,7826	0,7826	0,8226
1QO0	D	64	99	3	5	0,9532	0,9275	0,9275	0,9026
1QOR	A	31	246	0	2	0,9928	0,9394	0,9394	0,9653
1QPW	A	40	83	5	5	0,9248	0,8889	0,8889	0,8321

1QPW	B	37	82	6	5	0,9154	0,8810	0,8810	0,8079
1QSN	A	25	114	1	5	0,9586	0,8333	0,8333	0,8708
1QU9	A	43	65	3	3	0,9474	0,9348	0,9348	0,8907
1QVC	A	45	83	4	8	0,9143	0,8491	0,8491	0,8167
1R28	A	41	63	4	2	0,9455	0,9535	0,9535	0,8870
1R9J	A	13	411	5	100	0,8015	0,1150	0,1150	0,2329
1RIW	B	91	41	0	5	0,9635	0,9479	0,9479	0,9192
1RKT	A	38	143	2	4	0,9679	0,9048	0,9048	0,9067
1RL4	A	14	124	0	4	0,9718	0,7778	0,7778	0,8680
1RLI	A	41	85	2	5	0,9474	0,8913	0,8913	0,8830
1RQ2	A	29	212	1	5	0,9757	0,8529	0,8529	0,8947
1RU8	A	31	175	0	6	0,9717	0,8378	0,8378	0,9000
1RV0	H	85	183	1	13	0,9504	0,8673	0,8673	0,8915
1RV0	I	105	38	9	7	0,8994	0,9375	0,9375	0,7557
1RXZ	A	23	190	1	4	0,9771	0,8519	0,8519	0,8910
1RYI	A	19	289	0	3	0,9904	0,8636	0,8636	0,9245
1RYP	D	80	107	8	7	0,9257	0,9195	0,9195	0,8488
1RYP	H	79	58	3	17	0,8726	0,8229	0,8229	0,7550
1RYP	J	79	50	5	25	0,8113	0,7596	0,7596	0,6372
1RZH	L	114	149	3	6	0,9669	0,9500	0,9500	0,9330
1RZH	M	142	133	7	12	0,9354	0,9221	0,9221	0,8712
1S3R	A	20	411	1	3	0,9908	0,8696	0,8696	0,9053
1S4N	A	36	239	2	10	0,9582	0,7826	0,7826	0,8381
1S5U	A	8	73	1	37	0,6807	0,1778	0,1778	0,3013
1S7M	A	100	42	4	11	0,9045	0,9009	0,9009	0,7834
1S7Q	A	71	178	4	6	0,9614	0,9221	0,9221	0,9071
1SBQ	A	9	141	0	1	0,9934	0,9000	0,9000	0,9453
1SCF	A	23	83	1	4	0,9550	0,8519	0,8519	0,8754
1SDD	A	41	183	2	6	0,9655	0,8723	0,8723	0,8911
1SEI	A	5	109	2	2	0,9661	0,7143	0,7143	0,6963
1SFF	A	128	216	1	15	0,9556	0,8951	0,8951	0,9087
1SGM	A	44	126	1	2	0,9827	0,9565	0,9565	0,9554
1SGP	E	20	107	4	4	0,9407	0,8333	0,8333	0,7973
1SHD	A	16	71	1	1	0,9775	0,9412	0,9412	0,9273
1SHZ	C	41	133	2	8	0,9457	0,8367	0,8367	0,8585
1SJN	A	71	43	5	6	0,9120	0,9221	0,9221	0,8148
1SMP	A	31	357	2	6	0,9798	0,8378	0,8378	0,8764
1SMP	I	18	68	3	2	0,9451	0,9000	0,9000	0,8430
1SQS	A	30	165	5	3	0,9606	0,9091	0,9091	0,8592
1SW0	A	41	166	2	3	0,9764	0,9318	0,9318	0,9278
1SZ2	A	43	224	1	7	0,9709	0,8600	0,8600	0,9001
1SZH	A	12	119	2	2	0,9704	0,8571	0,8571	0,8406
1SZQ	A	58	335	5	6	0,9728	0,9063	0,9063	0,8973
1T15	A	19	159	0	0	1,0000	1,0000	1,0000	1,0000
1T3M	B	78	30	4	10	0,8852	0,8864	0,8864	0,7342
1T3Q	C	44	202	1	4	0,9801	0,9167	0,9167	0,9348
1T6G	A	29	291	1	7	0,9756	0,8056	0,8056	0,8698
1T6G	C	33	116	3	4	0,9551	0,8919	0,8919	0,8750
1T72	A	35	151	0	6	0,9688	0,8537	0,8537	0,9061
1T8Z	A	38	10	0	2	0,9600	0,9500	0,9500	0,8898
1T92	A	30	67	0	5	0,9510	0,8571	0,8571	0,8931
1TCR	A	52	123	3	13	0,9162	0,8000	0,8000	0,8122
1TCV	A	54	170	4	7	0,9532	0,8852	0,8852	0,8768
1TD3	A	34	57	0	3	0,9681	0,9189	0,9189	0,9343

1TDT	A	66	140	3	12	0,9321	0,8462	0,8462	0,8510
1TID	A	3	84	1	38	0,6905	0,0732	0,0732	0,1641
1TID	B	4	78	5	17	0,7885	0,1905	0,1905	0,1859
1TIQ	A	16	130	2	8	0,9359	0,6667	0,6667	0,7358
1TK2	A	25	183	0	4	0,9811	0,8621	0,8621	0,9185
1TK4	A	14	92	0	4	0,9636	0,7778	0,7778	0,8633
1TMX	A	72	183	5	7	0,9551	0,9114	0,9114	0,8915
1TOK	A	79	254	4	9	0,9624	0,8977	0,8977	0,8997
1TR8	A	18	65	1	1	0,9765	0,9474	0,9474	0,9322
1TRE	A	39	180	0	2	0,9910	0,9512	0,9512	0,9699
1TU1	A	6	89	2	40	0,6934	0,1304	0,1304	0,2184
1TVK	A	44	298	4	2	0,9828	0,9565	0,9565	0,9265
1TW3	A	80	216	1	7	0,9737	0,9195	0,9195	0,9354
1TWJ	A	47	25	4	1	0,9351	0,9792	0,9792	0,8619
1TWS	A	9	226	1	4	0,9792	0,6923	0,6923	0,7792
1TYO	A	68	269	7	5	0,9656	0,9315	0,9315	0,8972
1TZY	C	75	8	3	7	0,8925	0,9146	0,9146	0,5636
1TZY	D	69	12	0	1	0,9878	0,9857	0,9857	0,9539
1U0K	A	32	205	4	4	0,9673	0,8889	0,8889	0,8698
1U0R	A	65	170	1	8	0,9631	0,8904	0,8904	0,9118
1U2G	C	27	114	1	7	0,9463	0,7941	0,7941	0,8438
1U3T	A	46	257	3	5	0,9743	0,9020	0,9020	0,9049
1U3U	A	41	250	2	7	0,9700	0,8542	0,8542	0,8853
1U55	A	18	154	0	1	0,9942	0,9474	0,9474	0,9702
1U6L	A	32	81	4	4	0,9339	0,8889	0,8889	0,8418
1U7T	A	82	125	5	11	0,9283	0,8817	0,8817	0,8524
1UD9	B	19	195	0	3	0,9862	0,8636	0,8636	0,9223
1UEH	A	36	150	5	4	0,9538	0,9000	0,9000	0,8599
1UF3	A	61	109	1	10	0,9392	0,8592	0,8592	0,8746
1UFH	A	11	132	0	5	0,9662	0,6875	0,6875	0,8139
1UII	A	35	24	1	1	0,9672	0,9722	0,9722	0,9322
1UIK	A	111	205	5	22	0,9213	0,8346	0,8346	0,8350
1UIX	A	32	28	1	1	0,9677	0,9697	0,9697	0,9352
1UJ0	A	13	40	0	2	0,9636	0,8667	0,8667	0,9085
1UJK	A	24	102	1	3	0,9692	0,8889	0,8889	0,9049
1UJN	A	24	258	4	3	0,9758	0,8889	0,8889	0,8595
1UL1	X	40	225	2	7	0,9672	0,8511	0,8511	0,8813
1UMD	B	106	144	1	13	0,9470	0,8908	0,8908	0,8958
1UMY	A	107	203	2	12	0,9568	0,8992	0,8992	0,9074
1UN8	A	116	334	8	17	0,9474	0,8722	0,8722	0,8677
1UPT	B	32	17	2	3	0,9074	0,9143	0,9143	0,8000
1UQR	A	55	69	2	2	0,9688	0,9649	0,9649	0,9367
1USC	A	72	63	2	8	0,9310	0,9000	0,9000	0,8648
1USP	A	66	55	2	8	0,9237	0,8919	0,8919	0,8502
1UTY	A	49	81	2	10	0,9155	0,8305	0,8305	0,8284
1UW4	A	29	51	2	4	0,9302	0,8788	0,8788	0,8518
1UZH	C	63	45	2	3	0,9558	0,9545	0,9545	0,9094
1V0Z	A	81	235	4	14	0,9461	0,8526	0,8526	0,8658
1V2F	A	75	231	4	8	0,9623	0,9036	0,9036	0,9011
1V54	C	138	89	6	7	0,9458	0,9517	0,9517	0,8870
1V54	J	36	14	5	3	0,8621	0,9231	0,9231	0,6804
1V5V	A	43	271	2	9	0,9662	0,8269	0,8269	0,8700
1V6I	A	45	152	1	8	0,9563	0,8491	0,8491	0,8844
1V75	A	20	107	1	2	0,9769	0,9091	0,9091	0,9168

1V8C	A	36	100	3	10	0,9128	0,7826	0,7826	0,7918
1V96	A	28	97	0	3	0,9766	0,9032	0,9032	0,9360
1V9S	A	72	95	4	11	0,9176	0,8675	0,8675	0,8353
1V9Y	A	24	69	1	4	0,9490	0,8571	0,8571	0,8735
1VA0	A	57	134	4	4	0,9598	0,9344	0,9344	0,9054
1VDH	A	81	134	1	15	0,9307	0,8438	0,8438	0,8614
1VE2	A	55	131	3	6	0,9538	0,9016	0,9016	0,8918
1VEF	A	109	197	5	10	0,9533	0,9160	0,9160	0,8995
1VF2	A	35	160	3	4	0,9653	0,8974	0,8974	0,8878
1VGL	A	24	54	2	9	0,8764	0,7273	0,7273	0,7346
1VGM	A	94	221	1	8	0,9722	0,9216	0,9216	0,9356
1VGT	A	38	132	2	4	0,9659	0,9048	0,9048	0,9051
1VGZ	A	46	128	4	4	0,9560	0,9200	0,9200	0,8897
1VH5	A	29	87	1	6	0,9431	0,8286	0,8286	0,8586
1VHD	A	28	277	0	1	0,9967	0,9655	0,9655	0,9808
1VHQ	A	35	144	5	4	0,9521	0,8974	0,8974	0,8559
1VHZ	A	64	95	1	7	0,9521	0,9014	0,9014	0,9034
1VI2	A	17	236	0	4	0,9844	0,8095	0,8095	0,8922
1VIM	A	14	75	1	77	0,5329	0,1538	0,1538	0,2450
1VIX	A	2	300	8	37	0,8703	0,0513	0,0513	0,0478
1VJU	A	5	257	1	1	0,9924	0,8333	0,8333	0,8295
1VKH	A	4	182	6	17	0,8900	0,1905	0,1905	0,2233
1VL2	A	112	206	6	20	0,9244	0,8485	0,8485	0,8402
1VL4	A	27	326	4	4	0,9778	0,8710	0,8710	0,8588
1VLJ	A	4	275	4	39	0,8665	0,0930	0,0930	0,1720
1VM7	A	28	205	1	5	0,9749	0,8485	0,8485	0,8913
1VME	A	10	269	2	43	0,8611	0,1887	0,1887	0,3551
1VMK	A	39	163	4	9	0,9395	0,8125	0,8125	0,8209
1VPV	A	22	209	1	2	0,9872	0,9167	0,9167	0,9293
1VPZ	A	40	8	1	2	0,9412	0,9524	0,9524	0,8078
1VQV	A	53	194	3	10	0,9500	0,8413	0,8413	0,8610
1VR6	A	15	160	4	93	0,6434	0,1389	0,1389	0,2198
1VS3	A	30	183	2	5	0,9682	0,8571	0,8571	0,8780
1VZ5	A	40	173	4	2	0,9726	0,9524	0,9524	0,9136
1W5A	A	27	230	3	5	0,9698	0,8438	0,8438	0,8545
1W7J	B	41	80	4	3	0,9453	0,9318	0,9318	0,8796
1W7W	A	62	64	2	7	0,9333	0,8986	0,8986	0,8692
1W9E	A	52	81	4	14	0,8808	0,7879	0,7879	0,7607
1WCG	A	26	335	0	2	0,9945	0,9286	0,9286	0,9608
1WEK	A	71	107	1	5	0,9674	0,9342	0,9342	0,9332
1WHT	B	28	45	4	54	0,5573	0,3415	0,3415	0,2926
1WKQ	A	85	48	3	5	0,9433	0,9444	0,9444	0,8786
1WLT	A	40	108	1	4	0,9673	0,9091	0,9091	0,9197
1WOV	A	29	188	2	7	0,9602	0,8056	0,8056	0,8457
1WPN	A	18	119	0	4	0,9716	0,8182	0,8182	0,8897
1WTD	A	58	166	7	9	0,9333	0,8657	0,8657	0,8330
1WU7	A	89	282	3	14	0,9562	0,8641	0,8641	0,8862
1WUF	A	33	265	2	4	0,9803	0,8919	0,8919	0,9059
1WUI	L	84	342	4	13	0,9616	0,8660	0,8660	0,8856
1WV4	A	32	106	3	3	0,9583	0,9143	0,9143	0,8868
1WWP	A	22	85	0	0	1,0000	1,0000	1,0000	1,0000
1WWW	V	55	40	3	8	0,8962	0,8730	0,8730	0,7924
1WX0	A	67	105	3	4	0,9609	0,9437	0,9437	0,9182
1WYE	A	55	174	1	10	0,9542	0,8462	0,8462	0,8830

1WYT	A	201	162	3	15	0,9528	0,9306	0,9306	0,9064
1WYT	B	205	179	5	18	0,9435	0,9193	0,9193	0,8885
1WZ3	A	56	17	3	8	0,8690	0,8750	0,8750	0,6754
1X13	A	45	253	2	7	0,9707	0,8654	0,8654	0,8933
1X1Q	A	39	278	2	9	0,9665	0,8125	0,8125	0,8607
1X1V	A	20	104	0	2	0,9841	0,9091	0,9091	0,9444
1X7D	A	83	216	1	8	0,9708	0,9121	0,9121	0,9297
1X7O	A	34	197	2	2	0,9830	0,9444	0,9444	0,9344
1X9F	B	38	87	3	3	0,9542	0,9268	0,9268	0,8935
1X9F	D	54	66	3	7	0,9231	0,8852	0,8852	0,8466
1XB2	B	43	192	2	4	0,9751	0,9149	0,9149	0,9197
1XDI	A	84	303	8	9	0,9579	0,9032	0,9032	0,8808
1XDS	A	76	218	5	10	0,9515	0,8837	0,8837	0,8777
1XEA	A	60	199	1	12	0,9522	0,8333	0,8333	0,8762
1XEW	X	52	76	2	5	0,9481	0,9123	0,9123	0,8939
1XFS	A	24	117	0	3	0,9792	0,8889	0,8889	0,9309
1XG0	A	67	5	1	2	0,9600	0,9710	0,9710	0,7501
1XI8	A	46	175	3	5	0,9651	0,9020	0,9020	0,8980
1XIM	A	166	161	6	12	0,9478	0,9326	0,9326	0,8962
1XIY	A	22	126	0	2	0,9867	0,9167	0,9167	0,9499
1XJ5	A	63	173	3	13	0,9365	0,8289	0,8289	0,8475
1XKZ	A	9	192	3	3	0,9710	0,7500	0,7500	0,7346
1XMA	A	10	81	2	1	0,9681	0,9091	0,9091	0,8525
1XOU	B	47	32	1	3	0,9518	0,9400	0,9400	0,9015
1XPJ	A	58	51	0	5	0,9561	0,9206	0,9206	0,9157
1XRK	A	38	66	0	2	0,9811	0,9500	0,9500	0,9602
1XRP	A	14	211	13	1	0,9414	0,9333	0,9333	0,6706
1XRU	A	33	198	2	7	0,9625	0,8250	0,8250	0,8606
1XX9	A	39	140	3	4	0,9624	0,9070	0,9070	0,8934
1Y1L	A	34	73	2	0	0,9817	1,0000	1,0000	0,9588
1Y1X	A	36	127	1	4	0,9702	0,9000	0,9000	0,9170
1Y44	A	34	199	1	1	0,9915	0,9714	0,9714	0,9664
1Y4U	A	63	337	4	12	0,9615	0,8400	0,8400	0,8662
1Y56	B	44	266	1	7	0,9748	0,8627	0,8627	0,9044
1Y7B	A	66	359	8	18	0,9424	0,7857	0,7857	0,8030
1Y97	A	4	135	2	37	0,7809	0,0976	0,0976	0,1936
1Y9E	A	57	215	1	11	0,9577	0,8382	0,8382	0,8824
1Y9K	A	1	104	4	27	0,7721	0,0357	0,0357	-0,0028
1YA0	A	38	373	6	4	0,9762	0,9048	0,9048	0,8708
1YBG	A	58	254	1	17	0,9455	0,7733	0,7733	0,8415
1YDW	A	59	228	5	7	0,9599	0,8939	0,8939	0,8822
1YEE	H	41	131	1	13	0,9247	0,7593	0,7593	0,8160
1YGA	A	17	250	2	5	0,9745	0,7727	0,7727	0,8181
1YIF	A	70	358	4	9	0,9705	0,8861	0,8861	0,8979
1YKD	A	78	276	3	5	0,9779	0,9398	0,9398	0,9371
1YKF	A	86	192	3	14	0,9424	0,8600	0,8600	0,8711
1YKU	A	29	91	0	6	0,9524	0,8286	0,8286	0,8817
1YLM	A	28	100	0	2	0,9846	0,9333	0,9333	0,9566
1YMP	A	16	100	2	1	0,9748	0,9412	0,9412	0,9000
1YMT	A	12	178	0	0	1,0000	1,0000	1,0000	1,0000
1YPH	E	63	18	1	6	0,9205	0,9130	0,9130	0,7949
1YQ6	A	45	81	1	5	0,9545	0,9000	0,9000	0,9038
1YRR	A	22	278	2	2	0,9868	0,9167	0,9167	0,9095
1YRT	B	37	26	3	3	0,9130	0,9250	0,9250	0,8216

1YRZ	A	17	420	4	1	0,9887	0,9444	0,9444	0,8687
1YSJ	A	25	276	0	5	0,9837	0,8333	0,8333	0,9047
1YU4	A	101	208	4	16	0,9392	0,8632	0,8632	0,8671
1YUC	A	36	180	3	4	0,9686	0,9000	0,9000	0,8924
1YWH	A	65	157	3	22	0,8988	0,7471	0,7471	0,7789
1YXM	A	96	149	6	9	0,9423	0,9143	0,9143	0,8799
1YY7	A	25	153	1	0	0,9944	1,0000	1,0000	0,9774
1Z0A	A	41	89	2	4	0,9559	0,9111	0,9111	0,8997
1Z0J	A	17	119	3	3	0,9577	0,8500	0,8500	0,8254
1Z0S	A	54	143	5	9	0,9336	0,8571	0,8571	0,8395
1Z3A	A	32	92	3	3	0,9538	0,9143	0,9143	0,8827
1Z4E	A	63	73	2	2	0,9714	0,9692	0,9692	0,9426
1Z5G	A	55	121	1	8	0,9514	0,8730	0,8730	0,8920
1Z68	A	48	578	1	6	0,9889	0,8889	0,8889	0,9273
1Z6X	A	14	135	1	1	0,9868	0,9333	0,9333	0,9260
1Z7G	A	18	149	0	4	0,9766	0,8182	0,8182	0,8926
1Z8G	A	15	308	0	2	0,9938	0,8824	0,8824	0,9363
1Z9M	A	18	74	0	5	0,9485	0,7826	0,7826	0,8562
1ZAN	L	48	131	2	13	0,9227	0,7869	0,7869	0,8193
1ZB9	A	71	50	1	9	0,9237	0,8875	0,8875	0,8506
1ZC6	A	20	223	0	3	0,9878	0,8696	0,8696	0,9263
1ZE3	H	41	67	4	6	0,9153	0,8723	0,8723	0,8224
1ZEM	A	71	140	5	9	0,9378	0,8875	0,8875	0,8633
1ZGC	A	21	406	0	3	0,9930	0,8750	0,8750	0,9320
1ZGY	A	17	226	0	2	0,9918	0,8947	0,8947	0,9417
1ZKR	A	12	109	2	3	0,9603	0,8000	0,8000	0,8058
1ZLH	B	25	39	1	4	0,9275	0,8621	0,8621	0,8526
1ZOY	A	67	425	6	12	0,9647	0,8481	0,8481	0,8618
1ZOY	B	115	97	5	12	0,9258	0,9055	0,9055	0,8524
1ZPE	A	45	193	1	11	0,9520	0,8036	0,8036	0,8591
1ZRS	A	32	212	1	8	0,9644	0,8000	0,8000	0,8616
1ZT4	A	67	176	4	12	0,9382	0,8481	0,8481	0,8524
1ZTC	A	13	115	0	51	0,7151	0,2031	0,2031	0,3751
1ZTD	A	21	86	1	1	0,9817	0,9545	0,9545	0,9431
1ZXJ	A	25	154	3	2	0,9728	0,9259	0,9259	0,8933
1ZXM	A	55	264	1	16	0,9494	0,7746	0,7746	0,8444
1ZXZ	A	23	135	4	6	0,9405	0,7931	0,7931	0,7865
1ZZ1	A	43	237	1	11	0,9589	0,7963	0,7963	0,8596
25C8	H	41	131	1	14	0,9198	0,7455	0,7455	0,8056
2A0U	A	57	253	2	6	0,9748	0,9048	0,9048	0,9196
2A10	A	51	35	3	7	0,8958	0,8793	0,8793	0,7890
2A2Q	T	48	113	3	9	0,9306	0,8421	0,8421	0,8414
2A40	B	33	157	2	8	0,9500	0,8049	0,8049	0,8418
2A4C	A	31	59	1	1	0,9783	0,9688	0,9688	0,9521
2A5D	B	25	131	1	7	0,9512	0,7813	0,7813	0,8394
2A6Q	A	59	23	3	1	0,9535	0,9833	0,9833	0,8887
2A6Q	E	46	22	3	5	0,8947	0,9020	0,9020	0,7676
2A7R	A	70	212	7	7	0,9527	0,9091	0,9091	0,8771
2A87	A	59	211	3	5	0,9712	0,9219	0,9219	0,9181
2A9D	A	36	269	1	15	0,9502	0,7059	0,7059	0,8038
2A9U	A	41	82	4	2	0,9535	0,9535	0,9535	0,8971
2AAO	A	32	102	1	4	0,9640	0,8889	0,8889	0,9052
2ABQ	A	29	229	1	6	0,9736	0,8286	0,8286	0,8807
2ACA	A	24	137	0	2	0,9877	0,9231	0,9231	0,9538

2ADF	A	22	129	3	6	0,9438	0,7857	0,7857	0,7984
2AE7	A	14	205	3	4	0,9690	0,7778	0,7778	0,7836
2AF4	C	39	237	2	2	0,9857	0,9512	0,9512	0,9429
2AG5	A	69	132	4	6	0,9526	0,9200	0,9200	0,8961
2AGJ	H	46	147	2	8	0,9507	0,8519	0,8519	0,8719
2AIO	H	48	137	2	11	0,9343	0,8136	0,8136	0,8413
2AKZ	A	47	273	3	9	0,9639	0,8393	0,8393	0,8673
2AP6	A	55	39	2	2	0,9592	0,9649	0,9649	0,9161
2AP9	A	60	201	6	11	0,9388	0,8451	0,8451	0,8364
2APO	A	31	224	2	3	0,9808	0,9118	0,9118	0,9145
2AQS	A	22	111	0	5	0,9638	0,8148	0,8148	0,8830
2ARR	A	51	245	3	4	0,9769	0,9273	0,9273	0,9218
2ARV	A	36	71	4	5	0,9224	0,8780	0,8780	0,8294
2ARZ	A	6	179	4	33	0,8333	0,1538	0,1538	0,2422
2AUK	A	61	93	8	15	0,8701	0,8026	0,8026	0,7342
2AW6	A	94	153	4	9	0,9500	0,9126	0,9126	0,8953
2AWP	A	5	138	1	19	0,8773	0,2083	0,2083	0,3785
2AZ1	A	63	60	2	8	0,9248	0,8873	0,8873	0,8533
2B2C	A	41	184	1	7	0,9657	0,8542	0,8542	0,8930
2B3D	A	45	121	4	4	0,9540	0,9184	0,9184	0,8864
2B7Y	A	106	55	5	12	0,9045	0,8983	0,8983	0,7952
2B8N	A	27	310	4	1	0,9854	0,9643	0,9643	0,9087
2B9C	A	70	45	4	9	0,8984	0,8861	0,8861	0,7918
2B9H	A	21	266	3	5	0,9729	0,8077	0,8077	0,8260
2BA0	G	76	141	4	10	0,9394	0,8837	0,8837	0,8698
2BB0	A	9	313	2	5	0,9787	0,6429	0,6429	0,7147
2BB3	A	12	142	2	2	0,9747	0,8571	0,8571	0,8433
2BBA	A	24	138	1	6	0,9586	0,8000	0,8000	0,8533
2BBK	L	64	49	2	4	0,9496	0,9412	0,9412	0,8981
2BC0	A	72	309	1	18	0,9525	0,8000	0,8000	0,8614
2BDR	A	43	97	2	6	0,9459	0,8776	0,8776	0,8771
2BHV	A	46	130	3	4	0,9617	0,9200	0,9200	0,9032
2BHW	A	45	167	3	6	0,9593	0,8824	0,8824	0,8836
2BJD	A	13	57	1	1	0,9722	0,9286	0,9286	0,9113
2BKM	A	17	97	3	2	0,9580	0,8947	0,8947	0,8471
2BNU	A	57	124	3	3	0,9679	0,9500	0,9500	0,9264
2BSJ	A	34	74	2	5	0,9391	0,8718	0,8718	0,8631
2BSZ	A	19	205	1	4	0,9782	0,8261	0,8261	0,8743
2BV8	B	60	90	6	6	0,9259	0,9091	0,9091	0,8466
2BVE	A	9	99	1	4	0,9558	0,6923	0,6923	0,7665
2BVJ	A	37	308	1	6	0,9801	0,8605	0,8605	0,9046
2BZ0	A	37	110	2	7	0,9423	0,8409	0,8409	0,8553
2BZ6	H	23	158	2	3	0,9731	0,8846	0,8846	0,8866
2C0P	A	22	401	1	3	0,9906	0,8800	0,8800	0,9126
2C3N	A	35	162	1	1	0,9899	0,9722	0,9722	0,9661
2C41	A	76	56	1	8	0,9362	0,9048	0,9048	0,8745
2C4B	A	20	103	3	6	0,9318	0,7692	0,7692	0,7769
2C4F	L	50	67	5	5	0,9213	0,9091	0,9091	0,8396
2C4U	A	77	151	2	16	0,9268	0,8280	0,8280	0,8463
2C63	A	43	157	3	6	0,9569	0,8776	0,8776	0,8782
2C7M	A	14	43	0	1	0,9828	0,9333	0,9333	0,9551
2C7W	A	31	58	1	6	0,9271	0,8378	0,8378	0,8475
2CCJ	A	24	139	3	2	0,9702	0,9231	0,9231	0,8882
2CCL	A	7	71	2	49	0,6047	0,1250	0,1250	0,1899



2CCL	B	3	34	2	15	0,6852	0,1667	0,1667	0,1807
2CHC	A	12	93	2	49	0,6731	0,1967	0,1967	0,2999
2CJA	A	97	289	5	14	0,9531	0,8739	0,8739	0,8805
2CJW	A	19	127	0	3	0,9799	0,8636	0,8636	0,9185
2CKF	B	75	65	4	15	0,8805	0,8333	0,8333	0,7686
2CKI	A	28	192	0	5	0,9778	0,8485	0,8485	0,9094
2CLY	B	72	39	3	6	0,9250	0,9231	0,9231	0,8391
2CNM	A	40	83	0	13	0,9044	0,7547	0,7547	0,8078
2CQS	A	73	549	14	18	0,9511	0,8022	0,8022	0,7922
2CQZ	A	75	68	3	8	0,9286	0,9036	0,9036	0,8588
2CU3	A	7	45	0	5	0,9123	0,5833	0,5833	0,7246
2CVI	A	22	51	0	3	0,9605	0,8800	0,8800	0,9117
2CX9	A	90	217	6	5	0,9654	0,9474	0,9474	0,9177
2CZV	A	44	126	3	3	0,9659	0,9362	0,9362	0,9129
2D0O	B	31	53	1	6	0,9231	0,8378	0,8378	0,8428
2D1C	A	131	274	4	12	0,9620	0,9161	0,9161	0,9150
2D3A	A	99	193	6	14	0,9359	0,8761	0,8761	0,8605
2D40	A	73	179	5	7	0,9545	0,9125	0,9125	0,8918
2D4R	A	28	106	0	5	0,9640	0,8485	0,8485	0,9001
2D5C	A	22	179	0	5	0,9757	0,8148	0,8148	0,8903
2D5X	A	26	108	0	1	0,9926	0,9630	0,9630	0,9768
2D6C	A	10	127	0	4	0,9716	0,7143	0,7143	0,8322
2DBS	A	27	40	2	4	0,9178	0,8710	0,8710	0,8317
2DC1	A	52	159	2	6	0,9635	0,8966	0,8966	0,9051
2DC4	A	20	126	2	2	0,9733	0,9091	0,9091	0,8935
2DDM	A	40	182	3	4	0,9694	0,9091	0,9091	0,9008
2DDZ	A	58	94	1	13	0,9157	0,8169	0,8169	0,8335
2DFD	A	56	194	10	2	0,9542	0,9655	0,9655	0,8765
2DGM	A	148	215	6	18	0,9380	0,8916	0,8916	0,8740
2DKJ	A	88	241	2	9	0,9676	0,9072	0,9072	0,9201
2DKN	A	45	166	6	4	0,9548	0,9184	0,9184	0,8711
2DLF	H	15	75	0	6	0,9375	0,7143	0,7143	0,8133
2DOQ	D	58	19	0	4	0,9506	0,9355	0,9355	0,8791
2DQ3	A	56	317	2	5	0,9816	0,9180	0,9180	0,9307
2DR1	A	78	215	1	14	0,9513	0,8478	0,8478	0,8837
2DR3	A	53	125	2	13	0,9223	0,8030	0,8030	0,8273
2DSR	G	27	46	2	2	0,9481	0,9310	0,9310	0,8894
2DUC	A	35	214	3	7	0,9614	0,8333	0,8333	0,8537
2DVT	A	89	172	4	9	0,9526	0,9082	0,9082	0,8963
2DZO	A	6	147	0	40	0,7927	0,1304	0,1304	0,3202
2,00E+18	A	62	148	2	9	0,9502	0,8732	0,8732	0,8853
2E2A	A	42	45	2	5	0,9255	0,8936	0,8936	0,8528
2E2N	A	43	194	1	3	0,9834	0,9348	0,9348	0,9457
2,00E+31	A	36	173	2	5	0,9676	0,8780	0,8780	0,8926
2E3D	A	66	163	5	10	0,9385	0,8684	0,8684	0,8551
2E3X	A	1	342	3	22	0,9321	0,0435	0,0435	0,0812
2,00E+55	A	7	85	0	76	0,5476	0,0843	0,0843	0,2110
2E7A	A	2	82	0	52	0,6176	0,0370	0,0370	0,1505
2,00E+85	A	4	114	5	17	0,8429	0,1905	0,1905	0,2161
2E9X	A	87	38	3	8	0,9191	0,9158	0,9158	0,8173
2E9X	B	79	80	1	6	0,9578	0,9294	0,9294	0,9174
2EAV	A	1	121	1	19	0,8592	0,0500	0,0500	0,1234
2ECU	A	4	80	0	57	0,5957	0,0656	0,0656	0,1957
2EEN	A	1	140	0	27	0,8393	0,0357	0,0357	0,1730

2EFY	A	2	201	3	42	0,8185	0,0455	0,0455	0,0836
2EGV	A	1	166	1	35	0,8227	0,0278	0,0278	0,0843
2EGX	A	1	154	1	39	0,7949	0,0250	0,0250	0,0743
2EGZ	A	2	162	2	10	0,9318	0,1667	0,1667	0,2613
2EHB	A	2	123	1	53	0,6983	0,0364	0,0364	0,1017
2EHP	A	0	73	2	32	0,6822	0,0000	0,0000	-0,0901
2EIG	A	5	129	0	55	0,7090	0,0833	0,0833	0,2417
2EJ8	A	1	87	0	27	0,7652	0,0357	0,0357	0,1651
2EK6	A	2	43	1	23	0,6522	0,0800	0,0800	0,1350
2EMQ	A	2	88	0	32	0,7377	0,0588	0,0588	0,2077
2EO0	A	1	103	3	7	0,9123	0,1250	0,1250	0,1342
2EPI	A	4	32	0	56	0,3913	0,0667	0,0667	0,1557
2EQB	B	60	26	5	2	0,9247	0,9677	0,9677	0,8287
2ESR	A	21	117	0	6	0,9583	0,7778	0,7778	0,8601
2EUL	A	34	110	1	3	0,9730	0,9189	0,9189	0,9272
2EW2	A	21	232	1	3	0,9844	0,8750	0,8750	0,9055
2EWS	A	43	176	2	6	0,9648	0,8776	0,8776	0,8940
2F02	A	33	237	2	6	0,9712	0,8462	0,8462	0,8771
2F20	A	19	187	1	2	0,9856	0,9048	0,9048	0,9192
2F22	A	25	94	1	2	0,9754	0,9259	0,9259	0,9280
2F2C	A	50	160	1	4	0,9767	0,9259	0,9259	0,9377
2F3O	A	14	586	0	4	0,9934	0,7778	0,7778	0,8789
2F62	A	43	91	3	5	0,9437	0,8958	0,8958	0,8733
2F8M	A	4	152	7	35	0,7879	0,1026	0,1026	0,1016
2F9I	A	9	158	9	68	0,6844	0,1169	0,1169	0,1120
2F9W	A	44	159	1	11	0,9442	0,8000	0,8000	0,8513
2FBK	A	72	71	6	2	0,9470	0,9730	0,9730	0,8954
2FBM	A	55	158	1	13	0,9383	0,8088	0,8088	0,8528
2FBN	A	28	105	1	1	0,9852	0,9655	0,9655	0,9561
2FFF	B	17	343	1	5	0,9836	0,7727	0,7727	0,8462
2FHQ	A	28	89	2	4	0,9512	0,8750	0,8750	0,8715
2FHX	A	38	141	2	6	0,9572	0,8636	0,8636	0,8789
2FHZ	A	31	54	4	4	0,9140	0,8857	0,8857	0,8167
2FJK	A	57	192	5	9	0,9468	0,8636	0,8636	0,8562
2FK3	A	21	31	0	4	0,9286	0,8400	0,8400	0,8626
2FKL	A	8	49	0	3	0,9500	0,7273	0,7273	0,8278
2FP4	A	60	176	3	5	0,9672	0,9231	0,9231	0,9155
2FPB	A	22	239	2	2	0,9849	0,9167	0,9167	0,9084
2FQ1	A	34	201	0	5	0,9792	0,8718	0,8718	0,9223
2FS2	A	20	88	3	2	0,9558	0,9091	0,9091	0,8616
2FTR	A	26	63	0	3	0,9674	0,8966	0,8966	0,9251
2FTY	A	77	361	2	8	0,9777	0,9059	0,9059	0,9263
2FU3	A	90	241	8	20	0,9220	0,8182	0,8182	0,8134
2FU4	A	13	58	1	0	0,9861	1,0000	1,0000	0,9554
2FVH	A	24	61	2	4	0,9341	0,8571	0,8571	0,8433
2FVU	A	9	133	2	25	0,8402	0,2647	0,2647	0,4061
2FYX	A	26	82	1	3	0,9643	0,8966	0,8966	0,9058
2G2N	A	38	55	2	2	0,9588	0,9500	0,9500	0,9149
2G72	A	14	210	2	3	0,9782	0,8235	0,8235	0,8372
2G8L	A	22	205	0	2	0,9913	0,9167	0,9167	0,9528
2G9H	B	72	97	6	5	0,9389	0,9351	0,9351	0,8754
2GAG	D	12	30	2	43	0,4828	0,2182	0,2182	0,2043
2GAH	B	16	232	1	103	0,7045	0,1345	0,1345	0,2872
2GAH	C	10	111	1	53	0,6914	0,1587	0,1587	0,2963

2GF2	A	78	159	3	7	0,9595	0,9176	0,9176	0,9099
2GF6	A	32	84	5	4	0,9280	0,8889	0,8889	0,8260
2GHJ	A	43	48	1	2	0,9681	0,9556	0,9556	0,9362
2GHV	C	22	139	1	8	0,9471	0,7333	0,7333	0,8094
2GK4	A	6	146	3	32	0,8128	0,1579	0,1579	0,2590
2GL9	B	79	40	5	4	0,9297	0,9518	0,9518	0,8451
2GM5	A	23	83	2	5	0,9381	0,8214	0,8214	0,8299
2GPY	A	27	123	2	6	0,9494	0,8182	0,8182	0,8423
2GQ2	A	79	131	6	9	0,9333	0,8977	0,8977	0,8595
2GRU	A	29	266	0	8	0,9736	0,7838	0,7838	0,8723
2GSC	A	51	41	2	5	0,9293	0,9107	0,9107	0,8589
2GSK	A	38	519	4	9	0,9772	0,8085	0,8085	0,8432
2GTA	A	56	29	5	2	0,9239	0,9655	0,9655	0,8358
2GUD	A	49	63	2	2	0,9655	0,9608	0,9608	0,9300
2GV8	A	41	346	2	5	0,9822	0,8913	0,8913	0,9120
2GZ1	A	63	212	2	11	0,9549	0,8514	0,8514	0,8801
2GZ6	A	16	283	1	9	0,9676	0,6400	0,6400	0,7612
2H0D	A	53	38	0	4	0,9579	0,9298	0,9298	0,9172
2H21	A	70	301	7	7	0,9636	0,9091	0,9091	0,8864
2H3N	A	43	46	2	4	0,9368	0,9149	0,9149	0,8744
2H3R	A	45	34	3	3	0,9294	0,9375	0,9375	0,8564
2H4U	A	44	56	3	2	0,9524	0,9565	0,9565	0,9037
2H88	B	114	87	9	13	0,9013	0,8976	0,8976	0,8004
2H8F	A	41	79	6	6	0,9091	0,8723	0,8723	0,8018
2H8N	A	48	17	1	2	0,9559	0,9600	0,9600	0,8893
2HAL	A	16	159	1	4	0,9722	0,8000	0,8000	0,8530
2HCY	A	37	242	1	6	0,9755	0,8605	0,8605	0,9017
2HIM	A	10	196	3	75	0,7254	0,1176	0,1176	0,2248
2HKU	A	37	131	0	4	0,9767	0,9024	0,9024	0,9358
2HNT	E	55	5	2	1	0,9524	0,9821	0,9821	0,7456
2HPG	A	56	210	4	8	0,9568	0,8750	0,8750	0,8762
2HQU	A	76	40	4	11	0,8855	0,8736	0,8736	0,7581
2HT9	A	12	77	1	4	0,9468	0,7500	0,7500	0,8025
2HTD	A	4	80	1	26	0,7568	0,1333	0,1333	0,2591
2HUF	A	97	225	5	11	0,9527	0,8981	0,8981	0,8903
2HWG	A	49	452	0	7	0,9862	0,8750	0,8750	0,9283
2HX1	A	46	177	2	16	0,9253	0,7419	0,7419	0,7998
2HYD	A	165	327	17	17	0,9354	0,9066	0,9066	0,8572
2HZL	A	78	195	5	8	0,9545	0,9070	0,9070	0,8911
2I04	A	31	44	3	4	0,9146	0,8857	0,8857	0,8252
2I14	A	128	187	4	20	0,9292	0,8649	0,8649	0,8584
2I39	A	9	58	3	37	0,6262	0,1957	0,1957	0,2298
2I6T	A	35	179	3	2	0,9772	0,9459	0,9459	0,9197
2I8D	A	77	31	3	5	0,9310	0,9390	0,9390	0,8371
2I9U	A	50	283	3	6	0,9737	0,8929	0,8929	0,9023
2IA0	A	59	83	2	8	0,9342	0,8806	0,8806	0,8681
2IA1	A	22	109	1	3	0,9704	0,8800	0,8800	0,8998
2IAM	D	66	139	3	7	0,9535	0,9041	0,9041	0,8957
2IBP	A	116	212	8	6	0,9591	0,9508	0,9508	0,9112
2IEQ	A	51	31	3	2	0,9425	0,9623	0,9623	0,8789
2IFR	A	21	149	0	3	0,9827	0,8750	0,8750	0,9261
2IG7	A	28	264	1	2	0,9898	0,9333	0,9333	0,9437
2IGI	A	4	122	0	32	0,7975	0,1111	0,1111	0,2967
2IHC	A	8	59	2	41	0,6091	0,1633	0,1633	0,2256

2IK8	B	24	85	1	1	0,9820	0,9600	0,9600	0,9484
2IMZ	A	21	101	0	4	0,9683	0,8400	0,8400	0,8989
2INC	C	26	43	2	5	0,9079	0,8387	0,8387	0,8092
2IOO	B	25	52	1	3	0,9506	0,8929	0,8929	0,8903
2IO5	A	33	106	1	7	0,9456	0,8250	0,8250	0,8609
2IPR	A	8	97	1	2	0,9722	0,8000	0,8000	0,8283
2IQQ	A	14	108	1	5	0,9531	0,7368	0,7368	0,8043
2ISW	A	46	193	5	7	0,9522	0,8679	0,8679	0,8547
2ISY	A	21	111	0	1	0,9925	0,9545	0,9545	0,9726
2ITM	A	27	354	1	4	0,9870	0,8710	0,8710	0,9096
2IUT	A	34	304	1	4	0,9854	0,8947	0,8947	0,9243
2IUY	A	36	255	1	7	0,9732	0,8372	0,8372	0,8880
2IVS	A	31	212	2	3	0,9798	0,9118	0,9118	0,9139
2IX2	C	27	185	0	2	0,9907	0,9310	0,9310	0,9597
2IX5	A	132	198	5	11	0,9538	0,9231	0,9231	0,9046
2IXE	A	28	188	1	4	0,9774	0,8750	0,8750	0,9064
2IXK	A	40	117	2	7	0,9458	0,8511	0,8511	0,8645
2IZ5	A	40	82	3	10	0,9037	0,8000	0,8000	0,7926
2IZW	A	7	112	0	43	0,7346	0,1400	0,1400	0,3181
2J1M	A	13	368	2	4	0,9845	0,7647	0,7647	0,8062
2J3H	A	30	240	0	4	0,9854	0,8824	0,8824	0,9316
2J3T	D	52	144	1	3	0,9800	0,9455	0,9455	0,9496
2J41	A	30	111	1	7	0,9463	0,8108	0,8108	0,8535
2J5C	A	16	402	2	3	0,9882	0,8421	0,8421	0,8590
2J5U	A	59	128	1	9	0,9492	0,8676	0,8676	0,8883
2J62	A	30	439	3	8	0,9771	0,7895	0,7895	0,8352
2J6H	A	81	398	4	11	0,9696	0,8804	0,8804	0,8978
2J6P	A	5	75	1	44	0,6400	0,1020	0,1020	0,2030
2J6R	A	47	166	1	9	0,9552	0,8393	0,8393	0,8793
2JBR	A	98	232	6	9	0,9565	0,9159	0,9159	0,8978
2JBY	A	26	82	3	3	0,9474	0,8966	0,8966	0,8613
2JCA	A	36	73	0	4	0,9646	0,9000	0,9000	0,9237
2JDI	A	115	264	4	17	0,9475	0,8712	0,8712	0,8807
2JFK	A	18	316	2	4	0,9824	0,8182	0,8182	0,8489
2JHF	A	42	251	1	5	0,9799	0,8936	0,8936	0,9228
2JIF	A	83	208	6	5	0,9636	0,9432	0,9432	0,9121
2JIW	A	0	476	3	11	0,9714	0,0000	0,0000	-0,0119
2JJY	A	7	113	0	93	0,5634	0,0700	0,0700	0,1960
2JK2	A	2	169	0	42	0,8028	0,0455	0,0455	0,1908
2JLA	A	9	340	1	124	0,7363	0,0677	0,0677	0,2024
2JLM	A	7	111	1	47	0,7108	0,1296	0,1296	0,2640
2MEV	2	63	141	3	11	0,9358	0,8514	0,8514	0,8560
2MIN	A	125	243	7	18	0,9364	0,8741	0,8741	0,8619
2NN3	C	2	153	6	45	0,7524	0,0426	0,0426	0,0105
2NPS	A	48	8	4	3	0,8889	0,9412	0,9412	0,6287
2NPS	C	53	20	4	2	0,9241	0,9636	0,9636	0,8177
2NR4	A	38	119	3	8	0,9345	0,8261	0,8261	0,8321
2NRU	A	28	224	3	7	0,9618	0,8000	0,8000	0,8287
2NU6	B	76	231	7	16	0,9303	0,8261	0,8261	0,8233
2NV4	A	55	61	4	4	0,9355	0,9322	0,9322	0,8707
2NWA	A	23	49	1	3	0,9474	0,8846	0,8846	0,8824
2NWU	A	13	99	0	5	0,9573	0,7222	0,7222	0,8292
2NXE	A	16	192	1	4	0,9765	0,8000	0,8000	0,8555
2NZ7	A	5	54	1	26	0,6860	0,1613	0,1613	0,2697

201Z	A	34	221	0	3	0,9884	0,9189	0,9189	0,9522
2027	A	19	99	2	1	0,9752	0,9500	0,9500	0,9123
2028	A	65	93	3	7	0,9405	0,9028	0,9028	0,8787
202Y	A	92	149	1	11	0,9526	0,8932	0,8932	0,9034
2038	A	32	39	1	5	0,9221	0,8649	0,8649	0,8479
205G	A	1	98	0	47	0,6781	0,0208	0,0208	0,1187
206F	A	40	91	1	5	0,9562	0,8889	0,8889	0,9005
20A9	A	3	185	4	23	0,8744	0,1154	0,1154	0,1731
20B3	A	34	230	2	1	0,9888	0,9714	0,9714	0,9514
20BP	A	12	62	1	2	0,9610	0,8571	0,8571	0,8662
20EZ	A	31	146	3	12	0,9219	0,7209	0,7209	0,7653
20IZ	A	71	218	4	12	0,9475	0,8554	0,8554	0,8655
20IZ	D	51	33	6	6	0,8750	0,8947	0,8947	0,7409
20K2	A	19	335	1	3	0,9888	0,8636	0,8636	0,9000
20KA	A	31	40	0	4	0,9467	0,8857	0,8857	0,8973
20KF	A	29	85	0	2	0,9828	0,9355	0,9355	0,9560
20KU	A	27	83	3	2	0,9565	0,9310	0,9310	0,8863
20MW	B	2	49	1	33	0,6000	0,0571	0,0571	0,0991
20MX	B	3	52	1	35	0,6044	0,0789	0,0789	0,1445
20O0	A	16	266	4	50	0,8393	0,2424	0,2424	0,3822
20OI	A	27	116	0	5	0,9662	0,8438	0,8438	0,8994
20OK	A	30	76	2	4	0,9464	0,8824	0,8824	0,8720
20OQ	A	10	223	4	2	0,9749	0,8333	0,8333	0,7586
20PT	A	4	142	2	46	0,7526	0,0800	0,0800	0,1670
20Q1	A	30	189	3	6	0,9605	0,8333	0,8333	0,8475
20QG	A	36	37	3	10	0,8488	0,7826	0,7826	0,7090
20QY	A	86	206	0	9	0,9701	0,9053	0,9053	0,9313
20U5	A	31	117	3	9	0,9250	0,7750	0,7750	0,7939
20UU	A	23	257	0	4	0,9859	0,8519	0,8519	0,9159
20VR	B	56	308	2	10	0,9681	0,8485	0,8485	0,8869
20WQ	A	3	151	9	21	0,8370	0,1250	0,1250	0,0938
20WY	A	3	242	2	33	0,8750	0,0833	0,0833	0,1899
20X4	A	107	179	2	17	0,9377	0,8629	0,8629	0,8731
20Y9	A	16	58	1	1	0,9737	0,9412	0,9412	0,9242
20YA	A	29	54	1	7	0,9121	0,8056	0,8056	0,8190
20ZN	A	1	96	0	23	0,8083	0,0417	0,0417	0,1833
2P0M	A	2	530	3	33	0,9366	0,0571	0,0571	0,1326
2P0R	A	41	231	1	5	0,9784	0,8913	0,8913	0,9204
2P0V	A	35	322	5	8	0,9649	0,8140	0,8140	0,8243
2P10	A	83	134	7	16	0,9042	0,8384	0,8384	0,8020
2P1J	A	59	93	1	4	0,9682	0,9365	0,9365	0,9339
2P1N	A	32	80	2	7	0,9256	0,8205	0,8205	0,8278
2P35	A	32	184	1	4	0,9774	0,8889	0,8889	0,9154
2P58	A	1	27	1	21	0,5600	0,0455	0,0455	0,0247
2P5S	A	13	119	1	2	0,9778	0,8667	0,8667	0,8848
2P6P	A	2	288	0	41	0,8761	0,0465	0,0465	0,2018
2P7O	A	45	67	4	4	0,9333	0,9184	0,9184	0,8620
2P8J	A	29	154	1	1	0,9892	0,9667	0,9667	0,9602
2P9X	A	1	61	0	27	0,6966	0,0357	0,0357	0,1573
2PA7	A	38	82	3	3	0,9524	0,9268	0,9268	0,8915
2PAQ	A	46	110	0	6	0,9630	0,8846	0,8846	0,9159
2PBF	A	27	152	0	4	0,9781	0,8710	0,8710	0,9212
2PBR	A	1	162	0	9	0,9477	0,1000	0,1000	0,3078
2PDO	A	56	55	0	2	0,9823	0,9655	0,9655	0,9652

2PEH	A	23	65	2	5	0,9263	0,8214	0,8214	0,8196
2PEQ	A	55	44	2	2	0,9612	0,9649	0,9649	0,9214
2PEX	A	5	68	0	58	0,5573	0,0794	0,0794	0,2070
2PFR	A	16	233	0	5	0,9803	0,7619	0,7619	0,8637
2PGW	A	106	203	4	9	0,9596	0,9217	0,9217	0,9118
2PH4	A	1	97	0	13	0,8829	0,0714	0,0714	0,2510
2PHD	A	11	164	2	134	0,5627	0,0759	0,0759	0,1591
2PHL	A	128	186	2	20	0,9345	0,8649	0,8649	0,8707
2PID	A	1	246	0	36	0,8728	0,0270	0,0270	0,1535
2PIH	A	44	71	3	2	0,9583	0,9565	0,9565	0,9124
2PK7	A	18	42	0	0	1,0000	1,0000	1,0000	1,0000
2PKA	A	66	9	2	1	0,9615	0,9851	0,9851	0,8362
2PM7	A	113	196	6	9	0,9537	0,9262	0,9262	0,9011
2POK	A	69	315	2	7	0,9771	0,9079	0,9079	0,9255
2PQ5	A	7	94	4	34	0,7266	0,1707	0,1707	0,2195
2PQN	A	2	88	1	23	0,7895	0,0800	0,0800	0,1777
2PQV	A	30	101	1	3	0,9704	0,9091	0,9091	0,9189
2PS1	A	2	152	0	42	0,7857	0,0455	0,0455	0,1887
2PT7	A	7	188	0	99	0,6633	0,0660	0,0660	0,2080
2PVA	A	103	165	5	14	0,9338	0,8803	0,8803	0,8631
2PVP	A	5	239	3	42	0,8443	0,1064	0,1064	0,2114
2PX7	A	48	117	3	5	0,9538	0,9057	0,9057	0,8904
2PYW	A	2	302	0	50	0,8588	0,0385	0,0385	0,1817
2PZ0	A	1	163	0	26	0,8632	0,0370	0,0370	0,1787
2PZM	A	4	244	5	24	0,8953	0,1429	0,1429	0,2087
2Q0H	A	3	186	1	19	0,9043	0,1364	0,1364	0,2935
2Q0L	A	55	200	5	5	0,9623	0,9167	0,9167	0,8923
2Q0Q	A	3	152	0	37	0,8073	0,0750	0,0750	0,2456
2Q2I	A	6	60	0	41	0,6168	0,1277	0,1277	0,2754
2Q2J	A	4	91	1	17	0,8407	0,1905	0,1905	0,3397
2Q2L	A	1	117	1	12	0,9008	0,0769	0,0769	0,1669
2Q2V	A	2	187	1	31	0,8552	0,0606	0,0606	0,1703
2Q5B	A	3	68	0	19	0,7889	0,1364	0,1364	0,3265
2Q78	A	8	59	2	43	0,5982	0,1569	0,1569	0,2167
2Q86	B	5	149	2	53	0,7368	0,0862	0,0862	0,1816
2Q8U	A	5	227	7	11	0,9280	0,3125	0,3125	0,3235
2QAI	A	24	62	1	2	0,9663	0,9231	0,9231	0,9179
2QC1	A	1	52	1	18	0,7361	0,0526	0,0526	0,0906
2QCQ	A	3	62	2	36	0,6311	0,0769	0,0769	0,1031
2QCZ	A	6	228	0	66	0,7800	0,0833	0,0833	0,2542
2QFA	A	2	85	1	41	0,6744	0,0465	0,0465	0,1091
2QFR	A	2	301	2	44	0,8682	0,0435	0,0435	0,1172
2QHx	A	6	138	0	86	0,6261	0,0652	0,0652	0,2004
2QIB	A	5	153	2	50	0,7524	0,0909	0,0909	0,1911
2QJ2	A	2	125	0	29	0,8141	0,0645	0,0645	0,2288
2QJ4	A	2	123	1	36	0,7716	0,0526	0,0526	0,1401
2QJ8	A	30	240	0	7	0,9747	0,8108	0,8108	0,8876
2QJT	A	3	272	3	26	0,9046	0,1034	0,1034	0,1954
2QKL	A	2	88	0	20	0,8182	0,0909	0,0909	0,2722
2QM8	A	4	240	2	52	0,8188	0,0714	0,0714	0,1757
2QM9	A	3	108	0	16	0,8740	0,1579	0,1579	0,3708
2QN6	A	7	295	1	46	0,8653	0,1321	0,1321	0,3086
2QQD	B	7	28	0	68	0,3398	0,0933	0,0933	0,1650
2QQD	D	2	0	0	48	0,0400	0,0400	0,0400	#DIV/0!

2QQN	A	1	116	2	12	0,8931	0,0769	0,0769	0,1199
2QR4	A	2	420	4	50	0,8866	0,0385	0,0385	0,0812
2QRD	A	8	32	1	71	0,3571	0,1013	0,1013	0,1190
2QSI	A	4	53	0	50	0,5327	0,0741	0,0741	0,1952
2QSX	A	1	117	4	31	0,7712	0,0313	0,0313	-0,0041
2QTW	A	4	53	1	25	0,6867	0,1379	0,1379	0,2393
2QUF	A	7	105	2	94	0,5385	0,0693	0,0693	0,1243
2QWO	B	1	66	1	14	0,8171	0,0667	0,0667	0,1297
2QYO	A	10	206	0	104	0,6750	0,0877	0,0877	0,2414
2QZX	A	3	266	1	29	0,8997	0,0938	0,0938	0,2422
2ROL	A	2	155	2	40	0,7889	0,0476	0,0476	0,1014
2R2C	A	1	85	1	22	0,7890	0,0435	0,0435	0,0968
2R2D	A	1	219	1	13	0,9402	0,0714	0,0714	0,1723
2R3U	A	3	120	2	46	0,7193	0,0612	0,0612	0,1203
2R50	A	4	107	1	20	0,8409	0,1667	0,1667	0,3180
2R6I	A	4	206	4	13	0,9251	0,2353	0,2353	0,3087
2R6Z	A	2	165	0	30	0,8477	0,0625	0,0625	0,2300
2R9A	A	7	125	2	63	0,6701	0,1000	0,1000	0,1931
2R9S	A	1	274	3	14	0,9418	0,0667	0,0667	0,1060
2RAO	A	3	87	1	46	0,6569	0,0612	0,0612	0,1420
2RBB	A	5	58	0	53	0,5431	0,0862	0,0862	0,2122
2RCD	A	3	80	1	30	0,7281	0,0909	0,0909	0,1936
2RD7	A	7	260	3	23	0,9113	0,2333	0,2333	0,3706
2RET	A	3	41	1	31	0,5789	0,0882	0,0882	0,1435
2RFQ	A	10	272	1	57	0,8294	0,1493	0,1493	0,3273
2RIE	A	3	90	1	35	0,7209	0,0789	0,0789	0,1787
2RJI	A	3	57	2	12	0,8108	0,2000	0,2000	0,2660
2TEC	E	30	176	3	1	0,9810	0,9677	0,9677	0,9269
2TMG	A	71	246	3	11	0,9577	0,8659	0,8659	0,8847
2TNF	A	46	77	4	6	0,9248	0,8846	0,8846	0,8414
2UUZ	A	39	47	1	2	0,9663	0,9512	0,9512	0,9323
2UVK	A	3	274	0	27	0,9112	0,1000	0,1000	0,3017
2UWG	A	5	88	2	64	0,5849	0,0725	0,0725	0,1214
2UWI	A	16	107	0	4	0,9685	0,8000	0,8000	0,8782
2UXR	A	11	237	1	95	0,7209	0,1038	0,1038	0,2506
2UXU	A	51	134	2	4	0,9686	0,9273	0,9273	0,9229
2UXX	A	7	529	1	55	0,9054	0,1129	0,1129	0,2944
2UY6	B	53	85	2	7	0,9388	0,8833	0,8833	0,8738
2UYK	A	40	62	4	8	0,8947	0,8333	0,8333	0,7837
2UZ1	A	117	313	7	20	0,9409	0,8540	0,8540	0,8574
2UZI	R	22	117	0	0	1,0000	1,0000	1,0000	1,0000
2V4J	A	26	148	0	217	0,4450	0,1070	0,1070	0,2083
2V54	A	3	152	0	20	0,8857	0,1304	0,1304	0,3395
2V66	B	16	25	2	68	0,3694	0,1905	0,1905	0,1355
2V8D	A	9	292	2	56	0,8384	0,1385	0,1385	0,2942
2V9D	A	7	195	0	64	0,7594	0,0986	0,0986	0,2725
2V9U	A	4	70	3	43	0,6167	0,0851	0,0851	0,0917
2V9Y	A	0	245	6	11	0,9351	0,0000	0,0000	-0,0320
2VFX	A	5	147	5	23	0,8444	0,1786	0,1786	0,2305
2VGL	S	20	25	2	79	0,3571	0,2020	0,2020	0,1383
2VH0	B	4	27	2	16	0,6327	0,2000	0,2000	0,1965
2VH3	A	5	74	0	22	0,7822	0,1852	0,1852	0,3778
2VHF	A	3	275	7	19	0,9145	0,1364	0,1364	0,1620
2VKW	A	3	146	2	34	0,8054	0,0811	0,0811	0,1667

2VLD	A	10	131	5	67	0,6620	0,1299	0,1299	0,1748
2VPQ	A	4	318	10	39	0,8679	0,0930	0,0930	0,1051
2VQQ	A	5	252	5	26	0,8924	0,1613	0,1613	0,2401
2VRO	A	9	342	4	76	0,8144	0,1059	0,1059	0,2194
2VSM	A	4	292	3	39	0,8757	0,0930	0,0930	0,1939
2VT8	A	3	106	5	8	0,8934	0,2727	0,2727	0,2634
2VU1	A	14	226	2	72	0,7643	0,1628	0,1628	0,3123
2VV6	A	4	71	4	21	0,7500	0,1600	0,1600	0,1703
2VVH	A	9	127	1	57	0,7010	0,1364	0,1364	0,2755
2VVW	A	4	106	3	15	0,8594	0,2105	0,2105	0,2862
2VXG	A	2	86	4	23	0,7652	0,0800	0,0800	0,0659
2VXO	A	12	430	6	96	0,8125	0,1111	0,1111	0,2171
2VYN	A	21	185	3	68	0,7437	0,2360	0,2360	0,3652
2WOP	A	0	61	2	24	0,7011	0,0000	0,0000	-0,0947
2W2G	A	5	189	3	29	0,8584	0,1471	0,1471	0,2543
2W2Q	P	4	52	1	27	0,6667	0,1290	0,1290	0,2247
2W41	A	1	381	2	29	0,9249	0,0333	0,0333	0,0859
2W43	A	7	141	1	21	0,8706	0,2500	0,2500	0,4256
2W4E	A	5	83	4	28	0,7333	0,1515	0,1515	0,1789
2W4L	A	17	50	4	74	0,4621	0,1868	0,1868	0,1549
2W4S	A	4	49	4	23	0,6625	0,1481	0,1481	0,1146
2W57	A	5	86	2	30	0,7398	0,1429	0,1429	0,2340
2W72	A	8	79	2	38	0,6850	0,1739	0,1739	0,2663
2W7R	A	4	58	3	21	0,7209	0,1600	0,1600	0,1840
2W7W	A	3	140	1	24	0,8512	0,1111	0,1111	0,2506
2WAD	A	2	491	9	19	0,9463	0,0952	0,0952	0,1057
2WBA	A	13	340	3	76	0,8171	0,1461	0,1461	0,2941
2WBJ	B	13	92	1	63	0,6213	0,1711	0,1711	0,2893
2WBT	A	17	55	5	41	0,6102	0,2931	0,2931	0,2693
2WCB	A	3	50	1	31	0,6235	0,0882	0,0882	0,1588
2WCU	A	4	95	6	23	0,7734	0,1481	0,1481	0,1349
2WER	A	3	161	2	14	0,9111	0,1765	0,1765	0,2922
2WFH	A	7	117	3	31	0,7848	0,1842	0,1842	0,2795
2WHD	A	8	241	4	28	0,8861	0,2222	0,2222	0,3403
2WHN	A	5	80	2	13	0,8500	0,2778	0,2778	0,3815
2WHX	A	4	465	9	34	0,9160	0,1053	0,1053	0,1438
2WIU	A	6	333	7	36	0,8874	0,1429	0,1429	0,2110
2WKJ	A	10	175	1	52	0,7773	0,1613	0,1613	0,3253
2WN3	A	14	122	3	84	0,6099	0,1429	0,1429	0,2223
2WNR	A	9	142	7	64	0,6802	0,1233	0,1233	0,1386
2WNR	B	13	114	3	58	0,6755	0,1831	0,1831	0,2736
2WNT	A	3	225	6	16	0,9120	0,1579	0,1579	0,1877
2WO3	B	4	96	7	23	0,7692	0,1481	0,1481	0,1169
2WOX	A	16	286	1	101	0,7475	0,1368	0,1368	0,3011
2WPN	B	12	300	3	76	0,7980	0,1364	0,1364	0,2750
2WPQ	A	29	32	2	36	0,6162	0,4462	0,4462	0,3966
2WR7	A	17	305	5	106	0,7436	0,1382	0,1382	0,2507
2WR9	A	6	69	0	33	0,6944	0,1538	0,1538	0,3226
2WS2	A	6	153	2	25	0,8548	0,1935	0,1935	0,3318
2WU9	A	10	233	4	56	0,8020	0,1515	0,1515	0,2647
2WUA	A	12	226	1	60	0,7960	0,1667	0,1667	0,3402
2WUJ	A	14	16	2	18	0,6000	0,4375	0,4375	0,3359
2WUL	A	4	77	2	16	0,8182	0,2000	0,2000	0,2939
2WUQ	A	5	184	3	33	0,8400	0,1316	0,1316	0,2338



2WV1	A	12	114	3	32	0,7826	0,2727	0,2727	0,3788
2WVG	A	21	296	1	139	0,6937	0,1313	0,1313	0,2850
2WVU	A	2	346	5	11	0,9560	0,1538	0,1538	0,1886
2WYT	A	3	106	2	18	0,8450	0,1429	0,1429	0,2378
2WZT	A	4	178	3	36	0,8235	0,1000	0,1000	0,1834
2X0G	A	11	232	5	36	0,8556	0,2340	0,2340	0,3432
2X3H	A	53	208	2	178	0,5918	0,2294	0,2294	0,3324
2X3L	A	9	279	1	74	0,7934	0,1084	0,1084	0,2691
2X3O	A	15	71	1	41	0,6719	0,2679	0,2679	0,3810
2X4H	A	8	83	5	19	0,7913	0,2963	0,2963	0,3206
2X7J	A	13	310	2	134	0,7037	0,0884	0,0884	0,2153
2XBL	A	15	77	2	76	0,5412	0,1648	0,1648	0,2320
2XGF	A	53	24	0	137	0,3598	0,2789	0,2789	0,2039
2XHN	A	2	413	4	27	0,9305	0,0690	0,0690	0,1271
2XL9	A	2	186	1	24	0,8826	0,0769	0,0769	0,1988
2YQY	A	2	105	5	7	0,8992	0,2222	0,2222	0,1986
2YR5	A	17	455	10	93	0,8209	0,1545	0,1545	0,2474
2YWVW	A	4	103	2	17	0,8492	0,1905	0,1905	0,3000
2YXH	A	21	43	0	45	0,5872	0,3182	0,3182	0,3943
2YY7	A	6	225	4	26	0,8851	0,1875	0,1875	0,2905
2YY9	A	6	63	4	20	0,7419	0,2308	0,2308	0,2478
2YYV	A	4	146	1	33	0,8152	0,1081	0,1081	0,2497
2YZI	A	7	89	3	26	0,7680	0,2121	0,2121	0,2917
2YZJ	A	10	64	0	70	0,5139	0,1250	0,1250	0,2443
2Z0F	A	2	376	5	21	0,9356	0,0870	0,0870	0,1311
2Z1M	A	9	202	2	62	0,7673	0,1268	0,1268	0,2612
2Z34	A	8	92	5	39	0,6944	0,1702	0,1702	0,1942
2Z36	A	2	337	5	12	0,9522	0,1429	0,1429	0,1795
2Z5B	A	4	76	4	26	0,7273	0,1333	0,1333	0,1429
2Z6D	A	6	83	3	23	0,7739	0,2069	0,2069	0,2781
2Z76	A	4	90	3	27	0,7581	0,1290	0,1290	0,1816
2Z9D	A	5	141	4	28	0,8202	0,1515	0,1515	0,2198
2ZD1	A	9	413	14	71	0,8323	0,1125	0,1125	0,1396
2ZD1	B	9	293	8	63	0,8097	0,1250	0,1250	0,1862
2ZDH	A	8	229	2	34	0,8681	0,1905	0,1905	0,3492
2ZDS	A	16	174	0	72	0,7252	0,1818	0,1818	0,3586
2ZFZ	A	6	32	0	29	0,5672	0,1714	0,1714	0,2999
2ZGI	A	14	145	6	53	0,7294	0,2090	0,2090	0,2705
2ZHY	A	11	71	1	42	0,6560	0,2075	0,2075	0,3249
2ZJT	A	3	135	3	34	0,7886	0,0811	0,0811	0,1332
2ZJU	A	9	108	1	73	0,6126	0,1098	0,1098	0,2235
2ZJX	A	2	39	6	8	0,7455	0,2000	0,2000	0,0729
2ZMF	A	8	124	5	29	0,7952	0,2162	0,2162	0,2749
2ZON	A	22	154	1	112	0,6090	0,1642	0,1642	0,2906
2ZRU	A	11	213	3	66	0,7645	0,1429	0,1429	0,2661
2ZSH	A	13	252	11	22	0,8893	0,3714	0,3714	0,3900
2ZSH	B	9	26	1	23	0,5932	0,2813	0,2813	0,3243
2ZU9	A	3	294	2	32	0,8973	0,0857	0,0857	0,1991
2ZUB	A	2	208	10	30	0,8400	0,0625	0,0625	0,0260
2ZW2	A	8	37	2	33	0,5625	0,1951	0,1951	0,2174
2ZXX	A	13	18	6	32	0,4493	0,2889	0,2889	0,0415
2ZXX	C	8	128	5	34	0,7771	0,1905	0,1905	0,2490
2ZY9	A	29	252	19	98	0,7060	0,2283	0,2283	0,2265
2ZYZ	A	22	37	1	30	0,6556	0,4231	0,4231	0,4493

2ZYZ	B	10	91	3	65	0,5976	0,1333	0,1333	0,1891
3A1Y	A	6	24	0	26	0,5357	0,1875	0,1875	0,3000
3A2V	A	18	76	2	109	0,4585	0,1417	0,1417	0,1900
3A3D	A	8	315	6	54	0,8433	0,1290	0,1290	0,2166
3A4U	B	8	38	5	12	0,7302	0,4000	0,4000	0,3264
3A4V	A	3	222	2	31	0,8721	0,0882	0,0882	0,1946
3A5R	A	3	233	5	63	0,7763	0,0455	0,0455	0,0630
3A5Z	A	17	176	1	95	0,6678	0,1518	0,1518	0,2946
3A6S	A	2	87	0	24	0,7876	0,0769	0,0769	0,2455
3A7H	A	1	230	7	19	0,8988	0,0500	0,0500	0,0316
3A9F	A	1	59	6	11	0,7792	0,0833	0,0833	-0,0113
3AB4	B	9	73	1	58	0,5816	0,1343	0,1343	0,2350
3ABI	A	4	270	3	15	0,9384	0,2105	0,2105	0,3217
3AEH	A	12	212	18	26	0,8358	0,3158	0,3158	0,2628
3AEK	A	11	256	1	80	0,7672	0,1209	0,1209	0,2818
3B48	A	8	64	1	40	0,6372	0,1667	0,1667	0,2762
3B4N	A	6	237	3	25	0,8967	0,1935	0,1935	0,3216
3B4Y	A	5	243	0	47	0,8407	0,0962	0,0962	0,2838
3B6V	A	2	240	7	24	0,8864	0,0769	0,0769	0,0799
3B7S	A	0	482	18	28	0,9129	0,0000	0,0000	-0,0445
3BAL	A	10	74	2	59	0,5793	0,1449	0,1449	0,2150
3BBJ	A	8	175	2	54	0,7657	0,1290	0,1290	0,2577
3BDW	A	2	85	1	23	0,7838	0,0800	0,0800	0,1761
3BEO	A	7	276	1	30	0,9013	0,1892	0,1892	0,3797
3BF0	A	17	318	4	78	0,8034	0,1789	0,1789	0,3194
3BF5	A	2	201	4	28	0,8638	0,0667	0,0667	0,0998
3BH1	A	9	326	5	62	0,8333	0,1268	0,1268	0,2322
3BH7	A	7	104	5	26	0,7817	0,2121	0,2121	0,2524
3BHP	A	6	20	0	23	0,5306	0,2069	0,2069	0,3102
3BIL	A	3	176	0	38	0,8249	0,0732	0,0732	0,2453
3BJ6	A	16	70	0	44	0,6615	0,2667	0,2667	0,4047
3BJE	A	12	193	2	60	0,7678	0,1667	0,1667	0,3114
3BJQ	A	19	156	4	85	0,6629	0,1827	0,1827	0,2732
3BJZ	A	10	92	3	55	0,6375	0,1538	0,1538	0,2198
3BK3	C	3	39	4	18	0,6563	0,1429	0,1429	0,0750
3BL9	A	20	141	5	93	0,6216	0,1770	0,1770	0,2397
3BLJ	A	6	142	1	22	0,8655	0,2143	0,2143	0,3871
3BM1	A	16	85	1	54	0,6474	0,2286	0,2286	0,3463
3BMA	A	9	275	3	62	0,8138	0,1268	0,1268	0,2562
3BNI	A	7	135	4	20	0,8554	0,2593	0,2593	0,3420
3BNW	A	2	112	6	23	0,7972	0,0800	0,0800	0,0482
3BOO	A	2	321	10	23	0,9073	0,0800	0,0800	0,0705
3BOS	A	11	151	1	40	0,7980	0,2157	0,2157	0,3846
3BOW	A	27	420	11	139	0,7487	0,1627	0,1627	0,2517
3BPD	A	7	29	0	48	0,4286	0,1273	0,1273	0,2189
3BQA	A	3	86	1	32	0,7295	0,0857	0,0857	0,1885
3BRJ	A	6	109	8	33	0,7372	0,1538	0,1538	0,1295
3BRV	B	18	15	3	25	0,5410	0,4186	0,4186	0,2418
3BSF	A	15	161	4	40	0,8000	0,2727	0,2727	0,3830
3BSY	A	5	115	4	42	0,7229	0,1064	0,1064	0,1448
3BU2	A	16	105	0	60	0,6685	0,2105	0,2105	0,3660
3BUM	B	5	254	6	13	0,9317	0,2778	0,2778	0,3215
3BWU	D	7	87	4	20	0,7966	0,2593	0,2593	0,3111
3BWU	F	6	80	0	39	0,6880	0,1333	0,1333	0,2994

3BXU	A	1	51	6	13	0,7324	0,0714	0,0714	-0,0452
3BXZ	A	5	340	9	22	0,9176	0,1852	0,1852	0,2173
3C17	A	4	186	1	42	0,8155	0,0870	0,0870	0,2242
3C31	A	7	185	4	25	0,8688	0,2188	0,2188	0,3197
3C3Y	A	5	145	5	42	0,7614	0,1064	0,1064	0,1418
3C5O	A	11	79	0	44	0,6716	0,2000	0,2000	0,3584
3C66	A	5	410	7	38	0,9022	0,1163	0,1163	0,1817
3C9D	A	11	154	4	22	0,8639	0,3333	0,3333	0,4329
3CBU	A	4	141	2	36	0,7923	0,1000	0,1000	0,1996
3CBY	A	9	63	2	24	0,7347	0,2727	0,2727	0,3622
3CD3	A	4	290	8	14	0,9304	0,2222	0,2222	0,2369
3CDK	A	15	106	1	77	0,6080	0,1630	0,1630	0,2818
3CDW	A	1	373	9	18	0,9327	0,0526	0,0526	0,0396
3CE6	A	23	278	1	104	0,7414	0,1811	0,1811	0,3490
3CHH	A	5	240	5	24	0,8942	0,1724	0,1724	0,2494
3CI0	I	6	41	4	25	0,6184	0,1935	0,1935	0,1522
3CI6	A	4	91	1	30	0,7540	0,1176	0,1176	0,2428
3CIJ	A	2	191	2	24	0,8813	0,0769	0,0769	0,1608
3CIK	A	5	510	10	23	0,9398	0,1786	0,1786	0,2150
3CKA	A	6	258	1	40	0,8656	0,1304	0,1304	0,3025
3CLH	A	7	230	7	17	0,9080	0,2917	0,2917	0,3362
3CLK	A	3	169	7	31	0,8190	0,0882	0,0882	0,0838
3CNK	A	2	52	3	27	0,6429	0,0690	0,0690	0,0290
3CNV	A	7	98	1	39	0,7241	0,1522	0,1522	0,2896
3CO8	A	11	241	5	53	0,8129	0,1719	0,1719	0,2773
3COG	A	12	190	5	91	0,6779	0,1165	0,1165	0,1863
3COL	A	5	121	3	33	0,7778	0,1316	0,1316	0,2100
3COO	A	1	103	4	22	0,8000	0,0435	0,0435	0,0121
3CQG	A	6	173	8	21	0,8606	0,2222	0,2222	0,2388
3CR3	A	4	150	3	23	0,8556	0,1481	0,1481	0,2374
3CR7	A	7	122	1	26	0,8269	0,2121	0,2121	0,3777
3CRK	C	2	52	3	23	0,6750	0,0800	0,0800	0,0487
3CSU	A	42	187	1	5	0,9745	0,8936	0,8936	0,9190
3CTW	B	5	84	4	21	0,7807	0,1923	0,1923	0,2285
3CTY	A	4	216	6	42	0,8209	0,0870	0,0870	0,1192
3CUQ	A	24	118	6	55	0,6995	0,3038	0,3038	0,3509
3CVF	A	15	23	2	29	0,5507	0,3409	0,3409	0,2910
3CYY	A	12	18	0	49	0,3797	0,1967	0,1967	0,2299
3CZ1	A	10	71	2	31	0,7105	0,2439	0,2439	0,3386
3CZU	A	6	112	5	32	0,7613	0,1579	0,1579	0,1929
3CZU	B	5	92	4	19	0,8083	0,2083	0,2083	0,2531
3D03	A	20	126	1	77	0,6518	0,2062	0,2062	0,3371
3D0F	A	14	43	0	43	0,5700	0,2456	0,2456	0,3504
3D0Y	A	3	46	3	31	0,5904	0,0882	0,0882	0,0513
3D1A	B	3	91	1	41	0,6912	0,0682	0,0682	0,1587
3D1L	A	19	145	2	67	0,7039	0,2209	0,2209	0,3494
3D25	A	9	182	2	63	0,7461	0,1250	0,1250	0,2530
3D31	A	3	252	2	63	0,7969	0,0455	0,0455	0,1226
3D3L	A	0	373	9	21	0,9256	0,0000	0,0000	-0,0354
3D45	A	0	1	0	0	1,0000	#DIV/0!	#DIV/0!	#DIV/0!
3D4J	A	6	268	2	30	0,8954	0,1667	0,1667	0,3216
3D4O	A	6	200	4	32	0,8512	0,1579	0,1579	0,2528
3D55	A	19	13	0	36	0,4706	0,3455	0,3455	0,3027
3D57	A	0	194	8	19	0,8778	0,0000	0,0000	-0,0594

3D5O	F	8	107	3	44	0,7099	0,1538	0,1538	0,2349
3D63	A	3	131	10	14	0,8481	0,1765	0,1765	0,1190
3D72	A	5	116	6	13	0,8643	0,2778	0,2778	0,2844
3D7J	A	11	57	0	45	0,6018	0,1964	0,1964	0,3313
3D7T	A	2	203	4	18	0,9031	0,1000	0,1000	0,1426
3D7T	B	3	213	5	18	0,9038	0,1429	0,1429	0,1887
3D85	C	7	67	4	44	0,6066	0,1373	0,1373	0,1393
3D9X	A	32	6	1	75	0,3333	0,2991	0,2991	0,0827
3DA0	A	16	52	0	54	0,5574	0,2286	0,2286	0,3349
3DA3	A	3	208	4	30	0,8612	0,0909	0,0909	0,1476
3DAL	A	1	113	6	7	0,8976	0,1250	0,1250	0,0794
3DAV	A	3	87	5	13	0,8333	0,1875	0,1875	0,1806
3DDV	A	2	73	2	41	0,6356	0,0465	0,0465	0,0528
3DE8	A	3	72	0	24	0,7576	0,1111	0,1111	0,2887
3DGP	A	9	29	1	19	0,6552	0,3214	0,3214	0,3811
3DH9	A	7	334	5	84	0,7930	0,0769	0,0769	0,1542
3DHF	A	17	267	4	92	0,7474	0,1560	0,1560	0,2795
3DHZ	A	13	168	3	59	0,7449	0,1806	0,1806	0,3001
3DI4	A	24	137	3	70	0,6880	0,2553	0,2553	0,3589
3DLJ	A	8	314	2	83	0,7912	0,0879	0,0879	0,2196
3DMY	A	10	320	11	51	0,8418	0,1639	0,1639	0,2104
3DNH	A	9	153	8	51	0,7330	0,1500	0,1500	0,1674
3DNM	A	2	206	5	25	0,8739	0,0741	0,0741	0,0946
3DO6	A	13	345	7	56	0,8504	0,1884	0,1884	0,2933
3DOE	A	11	135	4	18	0,8690	0,3793	0,3793	0,4646
3DOE	B	6	74	4	27	0,7207	0,1818	0,1818	0,2084
3DOR	A	11	398	6	51	0,8777	0,1774	0,1774	0,2945
3DR9	A	1	116	0	5	0,9590	0,1667	0,1667	0,3997
3DTN	A	4	161	5	27	0,8376	0,1290	0,1290	0,1725
3DTZ	A	9	109	4	68	0,6211	0,1169	0,1169	0,1585
3DXB	A	7	114	2	71	0,6237	0,0897	0,0897	0,1690
3DXV	A	9	252	2	98	0,7230	0,0841	0,0841	0,2026
3DZC	A	7	266	2	26	0,9070	0,2121	0,2121	0,3754
3E10	A	11	76	1	58	0,5959	0,1594	0,1594	0,2662
3E1E9	A	7	73	2	39	0,6612	0,1522	0,1522	0,2322
3E39	A	12	87	2	65	0,5964	0,1558	0,1558	0,2393
3E4P	A	8	174	10	27	0,8311	0,2286	0,2286	0,2324
3E5R	A	17	204	3	71	0,7492	0,1932	0,1932	0,3252
3E5W	A	10	133	2	63	0,6875	0,1370	0,1370	0,2501
3E60	A	14	239	4	66	0,7833	0,1750	0,1750	0,2983
3E6G	A	16	212	1	75	0,7500	0,1758	0,1758	0,3411
3E7M	A	1	340	14	8	0,9394	0,1111	0,1111	0,0559
3E9C	A	2	135	4	22	0,8405	0,0833	0,0833	0,1027
3ECH	A	20	55	5	41	0,6198	0,3279	0,3279	0,3020
3ECY	A	7	67	3	35	0,6607	0,1667	0,1667	0,2102
3EDF	A	6	473	2	41	0,9176	0,1277	0,1277	0,2876
3EDQ	A	13	60	2	48	0,5935	0,2131	0,2131	0,2763
3EDV	A	1	258	10	21	0,8931	0,0455	0,0455	0,0113
3EEY	A	9	104	2	51	0,6807	0,1500	0,1500	0,2533
3EGG	C	15	92	2	42	0,7086	0,2632	0,2632	0,3710
3EGJ	A	4	294	6	9	0,9521	0,3077	0,3077	0,3264
3EGV	B	2	28	2	39	0,4225	0,0488	0,0488	-0,0383
3EI3	B	16	243	8	39	0,8464	0,2909	0,2909	0,3700
3EIP	A	10	60	3	6	0,8861	0,6250	0,6250	0,6258

3EIT	A	8	166	2	44	0,7909	0,1538	0,1538	0,2895
3EKO	A	0	173	7	9	0,9153	0,0000	0,0000	-0,0439
3EMH	A	3	220	0	21	0,9139	0,1250	0,1250	0,3378
3ENM	A	13	192	2	52	0,7915	0,2000	0,2000	0,3521
3ENP	A	6	136	4	12	0,8987	0,3333	0,3333	0,3977
3EOQ	A	5	316	4	29	0,9068	0,1471	0,1471	0,2519
3EP8	A	15	135	6	66	0,6757	0,1852	0,1852	0,2346
3EQU	A	1	356	5	8	0,9649	0,1111	0,1111	0,1186
3ERR	A	0	182	4	0	0,9785	#DIV/0!	#DIV/0!	#DIV/0!
3EUC	A	19	245	4	67	0,7881	0,2209	0,2209	0,3539
3EUH	C	11	95	5	79	0,5579	0,1222	0,1222	0,1299
3EWW	A	6	150	3	49	0,7500	0,1091	0,1091	0,1940
3EXR	A	5	136	2	38	0,7790	0,1163	0,1163	0,2247
3EYB	A	7	122	5	47	0,7127	0,1296	0,1296	0,1660
3EZ2	A	14	246	7	64	0,7855	0,1795	0,1795	0,2643
3EZS	A	8	252	5	51	0,8228	0,1356	0,1356	0,2279
3F0N	A	4	267	1	29	0,9003	0,1212	0,1212	0,2872
3F1C	A	13	133	3	38	0,7807	0,2549	0,2549	0,3707
3F1P	B	7	56	6	19	0,7159	0,2692	0,2692	0,2217
3F3H	A	13	65	4	21	0,7573	0,3824	0,3824	0,4109
3F4B	A	18	155	2	74	0,6948	0,1957	0,1957	0,3248
3F4F	A	12	39	0	76	0,4016	0,1364	0,1364	0,2150
3F69	A	4	205	3	25	0,8819	0,1379	0,1379	0,2391
3F6A	A	7	87	5	39	0,6812	0,1522	0,1522	0,1637
3F6D	A	7	139	6	34	0,7849	0,1707	0,1707	0,2103
3F6Z	B	11	27	1	49	0,4318	0,1833	0,1833	0,2004
3F75	P	8	42	3	22	0,6667	0,2667	0,2667	0,2770
3F8H	A	5	78	4	32	0,6975	0,1351	0,1351	0,1512
3F8M	A	6	163	3	42	0,7897	0,1250	0,1250	0,2222
3FBZ	A	23	52	1	36	0,6696	0,3898	0,3898	0,4514
3FEA	A	6	55	3	15	0,7722	0,2857	0,2857	0,3254
3FFH	A	6	240	2	50	0,8255	0,1071	0,1071	0,2390
3FFU	A	1	95	2	24	0,7869	0,0400	0,0400	0,0505
3FG1	A	13	522	5	49	0,9083	0,2097	0,2097	0,3569
3FGV	A	3	58	6	18	0,7176	0,1429	0,1429	0,0688
3FHW	A	4	51	8	30	0,5914	0,1176	0,1176	-0,0258
3FIJ	A	7	147	0	43	0,7817	0,1400	0,1400	0,3291
3FJU	A	3	218	4	23	0,8911	0,1154	0,1154	0,1801
3FK4	A	12	238	5	68	0,7740	0,1500	0,1500	0,2502
3FKF	A	5	91	2	31	0,7442	0,1389	0,1389	0,2324
3FLH	A	6	68	2	32	0,6852	0,1579	0,1579	0,2358
3FLP	A	11	116	0	54	0,7017	0,1692	0,1692	0,3398
3FN9	A	9	477	9	78	0,8482	0,1034	0,1034	0,1747
3FPA	A	5	138	1	45	0,7566	0,1000	0,1000	0,2335
3FPC	A	20	181	1	80	0,7128	0,2000	0,2000	0,3544
3FPF	A	6	213	2	19	0,9125	0,2400	0,2400	0,3926
3FPK	A	4	168	6	28	0,8350	0,1250	0,1250	0,1526
3FPQ	A	3	223	5	19	0,9040	0,1364	0,1364	0,1842
3FPV	A	10	39	0	71	0,4083	0,1235	0,1235	0,2092
3FS4	A	5	92	2	35	0,7239	0,1250	0,1250	0,2133
3FTP	A	11	142	2	67	0,6892	0,1410	0,1410	0,2585
3FUB	A	7	135	6	32	0,7889	0,1795	0,1795	0,2179
3FUB	B	5	55	12	19	0,6593	0,2083	0,2083	0,0330
3FV9	A	14	186	0	93	0,6826	0,1308	0,1308	0,2953

3FWB	B	24	7	1	22	0,5741	0,5217	0,5217	0,2827
3FXG	A	23	216	0	111	0,6829	0,1716	0,1716	0,3367
3FY7	A	4	130	2	26	0,8272	0,1333	0,1333	0,2431
3FZV	A	11	167	3	77	0,6899	0,1250	0,1250	0,2247
3FZZ	A	1	167	1	20	0,8889	0,0476	0,0476	0,1280
3G12	A	4	61	2	32	0,6566	0,1111	0,1111	0,1600
3G13	A	5	95	11	26	0,7299	0,1613	0,1613	0,0749
3G2P	A	3	174	5	41	0,7937	0,0682	0,0682	0,0861
3G3D	A	8	140	1	46	0,7590	0,1481	0,1481	0,3008
3G3Z	A	22	72	2	36	0,7121	0,3793	0,3793	0,4533
3G5O	A	21	26	3	42	0,5109	0,3333	0,3333	0,2432
3G5W	A	19	163	0	81	0,6920	0,1900	0,1900	0,3563
3G67	A	28	91	4	89	0,5613	0,2393	0,2393	0,2740
3G72	A	3	259	4	23	0,9066	0,1154	0,1154	0,1864
3G7R	A	4	121	0	29	0,8117	0,1212	0,1212	0,3127
3G80	A	11	34	5	23	0,6164	0,3235	0,3235	0,2355
3GA1	A	10	53	2	39	0,6058	0,2041	0,2041	0,2620
3GB9	A	7	186	2	52	0,7814	0,1186	0,1186	0,2458
3GBR	A	3	245	4	22	0,9051	0,1200	0,1200	0,1897
3GC9	A	1	288	4	11	0,9507	0,0833	0,0833	0,1066
3GDG	A	14	128	3	93	0,5966	0,1308	0,1308	0,2085
3GE6	A	17	112	1	55	0,6973	0,2361	0,2361	0,3739
3GE8	C	9	42	2	24	0,6623	0,2727	0,2727	0,3214
3GET	A	14	227	2	60	0,7954	0,1892	0,1892	0,3467
3GFA	A	10	91	3	78	0,5549	0,1136	0,1136	0,1586
3GFD	A	21	77	5	95	0,4949	0,1810	0,1810	0,1751
3GFF	A	5	253	2	23	0,9117	0,1786	0,1786	0,3282
3GIO	A	7	88	3	34	0,7197	0,1707	0,1707	0,2409
3GJ0	A	2	167	4	10	0,9235	0,1667	0,1667	0,1992
3GJZ	A	3	250	4	41	0,8490	0,0682	0,0682	0,1228
3GK3	A	9	123	5	78	0,6140	0,1034	0,1034	0,1281
3GKB	A	10	175	0	62	0,7490	0,1389	0,1389	0,3202
3GL1	A	4	292	1	41	0,8757	0,0889	0,0889	0,2405
3GMG	A	8	86	1	31	0,7460	0,2051	0,2051	0,3476
3GMY	A	6	96	7	30	0,7338	0,1667	0,1667	0,1485
3GN5	A	0	100	4	21	0,8000	0,0000	0,0000	-0,0817
3GNI	B	2	182	5	29	0,8440	0,0645	0,0645	0,0748
3GOV	B	2	189	2	27	0,8682	0,0690	0,0690	0,1481
3GQP	A	7	85	2	41	0,6815	0,1458	0,1458	0,2357
3GR1	A	3	150	12	19	0,8315	0,1364	0,1364	0,0739
3GRI	A	5	338	2	23	0,9321	0,1786	0,1786	0,3352
3GRN	A	3	97	2	20	0,8197	0,1304	0,1304	0,2175
3GSN	A	10	109	2	61	0,6538	0,1408	0,1408	0,2414
3GTU	B	35	151	6	2	0,9588	0,9459	0,9459	0,8735
3GUZ	A	6	110	5	28	0,7785	0,1765	0,1765	0,2134
3GVH	A	16	178	1	75	0,7185	0,1758	0,1758	0,3313
3GVP	A	15	237	3	87	0,7368	0,1471	0,1471	0,2757
3GWJ	A	16	459	3	116	0,7997	0,1212	0,1212	0,2710
3GWO	A	10	23	9	11	0,6226	0,4762	0,4762	0,1988
3GZC	A	7	277	7	55	0,8208	0,1129	0,1129	0,1718
3H1Z	A	2	174	7	24	0,8502	0,0769	0,0769	0,0622
3H2D	A	1	93	2	29	0,7520	0,0333	0,0333	0,0343
3H43	A	11	21	1	36	0,4638	0,2340	0,2340	0,2319
3H4S	E	11	35	2	46	0,4894	0,1930	0,1930	0,1966

3H5D	A	1	233	2	24	0,9000	0,0400	0,0400	0,0869
3H6P	C	12	25	2	17	0,6607	0,4138	0,4138	0,3920
3H7D	A	0	162	2	12	0,9205	0,0000	0,0000	-0,0290
3H8K	A	3	102	7	33	0,7241	0,0833	0,0833	0,0326
3H91	A	6	16	1	24	0,4681	0,2000	0,2000	0,1905
3H9A	A	2	191	4	10	0,9324	0,1667	0,1667	0,2036
3HAH	A	27	148	5	81	0,6705	0,2500	0,2500	0,3263
3HB3	A	21	304	39	79	0,7336	0,2100	0,2100	0,1176
3HCG	A	3	97	3	22	0,8000	0,1200	0,1200	0,1684
3HCN	A	3	235	4	46	0,8264	0,0612	0,0612	0,1086
3HCY	A	3	103	2	23	0,8092	0,1154	0,1154	0,2005
3HDQ	A	10	246	1	77	0,7665	0,1149	0,1149	0,2727
3HF4	A	5	86	2	43	0,6691	0,1042	0,1042	0,1761
3HH1	A	3	61	3	37	0,6154	0,0750	0,0750	0,0587
3HHE	A	2	152	5	35	0,7938	0,0541	0,0541	0,0468
3HHJ	A	3	91	3	20	0,8034	0,1304	0,1304	0,1775
3HHR	B	44	120	7	9	0,9111	0,8302	0,8302	0,7840
3HHS	A	10	532	4	53	0,9048	0,1587	0,1587	0,3072
3HIO	A	5	363	3	69	0,8364	0,0676	0,0676	0,1662
3HI7	A	34	409	6	188	0,6954	0,1532	0,1532	0,2724
3HIM	A	7	127	3	35	0,7791	0,1667	0,1667	0,2636
3HJA	A	17	193	1	76	0,7317	0,1828	0,1828	0,3429
3HJV	A	5	201	0	34	0,8583	0,1282	0,1282	0,3311
3HKV	A	6	136	5	22	0,8402	0,2143	0,2143	0,2695
3HLO	A	3	251	5	34	0,8669	0,0811	0,0811	0,1255
3HLK	A	0	347	8	9	0,9533	0,0000	0,0000	-0,0239
3HLU	A	6	40	3	18	0,6866	0,2500	0,2500	0,2534
3HMK	A	4	227	12	19	0,8817	0,1739	0,1739	0,1462
3HMU	A	13	239	1	122	0,6720	0,0963	0,0963	0,2333
3HNO	A	0	298	13	15	0,9141	0,0000	0,0000	-0,0448
3HPM	A	3	67	2	27	0,7071	0,1000	0,1000	0,1490
3HPX	A	26	184	5	133	0,6034	0,1635	0,1635	0,2397
3HQ9	A	9	246	2	38	0,8644	0,1915	0,1915	0,3543
3HQI	A	19	173	6	62	0,7385	0,2346	0,2346	0,3158
3HR7	A	9	92	5	28	0,7537	0,2432	0,2432	0,2802
3HRD	B	29	95	4	167	0,4203	0,1480	0,1480	0,1611
3HS3	A	4	184	4	32	0,8393	0,1111	0,1111	0,1778
3HY2	X	6	50	0	40	0,5833	0,1304	0,1304	0,2692
3HYH	A	4	174	8	22	0,8558	0,1538	0,1538	0,1559
3HYM	B	14	151	4	95	0,6250	0,1284	0,1284	0,2005
3HYU	A	3	107	3	22	0,8148	0,1200	0,1200	0,1748
3HYW	A	10	308	1	38	0,8908	0,2083	0,2083	0,4049
3HZ4	A	5	65	1	36	0,6542	0,1220	0,1220	0,2257
3HZE	A	3	67	3	28	0,6931	0,0968	0,0968	0,1052
3HZH	A	5	88	7	15	0,8087	0,2500	0,2500	0,2186
3I05	A	6	250	5	30	0,8797	0,1667	0,1667	0,2539
3I1I	A	1	305	10	17	0,9189	0,0556	0,0556	0,0301
3I2W	A	26	154	4	76	0,6923	0,2549	0,2549	0,3509
3I3W	A	3	326	1	28	0,9190	0,0968	0,0968	0,2507
3I5C	A	9	126	8	26	0,7988	0,2571	0,2571	0,2660
3I5G	C	16	91	1	42	0,7133	0,2759	0,2759	0,4071
3I5T	A	17	241	3	112	0,6917	0,1318	0,1318	0,2523
3I6L	D	8	173	3	70	0,7126	0,1026	0,1026	0,1938
3I7F	A	16	304	9	110	0,7289	0,1270	0,1270	0,1917

3IA7	A	5	287	9	42	0,8513	0,1064	0,1064	0,1320
3IAC	A	15	312	3	84	0,7899	0,1515	0,1515	0,2970
3IAL	A	5	348	2	61	0,8486	0,0758	0,0758	0,1990
3IB6	A	7	104	1	34	0,7603	0,1707	0,1707	0,3183
3IBM	A	3	91	1	43	0,6812	0,0652	0,0652	0,1527
3ICF	A	3	234	4	18	0,9151	0,1429	0,1429	0,2122
3ICS	A	11	387	3	81	0,8257	0,1196	0,1196	0,2618
3IE4	A	0	76	3	11	0,8444	0,0000	0,0000	-0,0693
3IEL	A	4	290	4	57	0,8282	0,0656	0,0656	0,1321
3IIW	A	2	289	6	9	0,9510	0,1818	0,1818	0,1884
3ILW	A	7	351	7	42	0,8796	0,1429	0,1429	0,2202
3ILX	A	9	87	2	32	0,7385	0,2195	0,2195	0,3290
3IMF	A	9	119	3	74	0,6244	0,1084	0,1084	0,1753
3INQ	A	13	76	4	37	0,6846	0,2600	0,2600	0,3030
3IOY	A	14	132	3	88	0,6160	0,1373	0,1373	0,2207
3IP4	A	16	267	6	82	0,7628	0,1633	0,1633	0,2637
3IPF	A	3	43	1	17	0,7188	0,1500	0,1500	0,2437
3IR2	A	1	153	7	7	0,9167	0,1250	0,1250	0,0813
3ISO	A	9	151	3	30	0,8290	0,2308	0,2308	0,3513
3IT3	A	7	242	2	38	0,8616	0,1556	0,1556	0,3076
3IT4	B	18	82	1	84	0,5405	0,1765	0,1765	0,2694
3IVR	A	7	293	1	39	0,8824	0,1522	0,1522	0,3357
3IWJ	A	16	311	3	84	0,7899	0,1600	0,1600	0,3077
3JR8	A	2	81	3	15	0,8218	0,1176	0,1176	0,1413
3JTF	A	7	85	1	23	0,7931	0,2333	0,2333	0,3831
3JWR	A	5	235	4	40	0,8451	0,1111	0,1111	0,1967
3JZ4	A	9	300	3	90	0,7687	0,0909	0,0909	0,2051
3JZA	B	7	116	5	42	0,7235	0,1429	0,1429	0,1796
3JZQ	A	4	42	4	28	0,5897	0,1250	0,1250	0,0617
3K0Z	A	7	107	1	23	0,8261	0,2333	0,2333	0,3955
3K12	A	12	36	0	62	0,4364	0,1622	0,1622	0,2441
3K2I	A	5	309	3	42	0,8747	0,1064	0,1064	0,2211
3K4I	A	14	119	2	42	0,7514	0,2500	0,2500	0,3787
3K8A	A	10	46	4	37	0,5773	0,2128	0,2128	0,1888
3K8P	C	8	219	11	21	0,8764	0,2759	0,2759	0,2758
3K8P	D	4	497	14	24	0,9295	0,1429	0,1429	0,1426
3K8R	A	2	59	0	17	0,7821	0,1053	0,1053	0,2859
3K8V	A	3	249	10	24	0,8811	0,1111	0,1111	0,1018
3K8X	A	5	532	33	35	0,8876	0,1250	0,1250	0,0682
3K90	A	5	125	3	29	0,8025	0,1471	0,1471	0,2323
3K9U	A	1	129	6	8	0,9028	0,1111	0,1111	0,0750
3K9V	A	1	401	10	11	0,9504	0,0833	0,0833	0,0616
3KB1	A	7	190	5	11	0,9249	0,3889	0,3889	0,4382
3KCP	A	4	247	5	16	0,9228	0,2000	0,2000	0,2629
3KCP	B	4	115	0	25	0,8264	0,1379	0,1379	0,3366
3KDG	A	4	133	8	28	0,7919	0,1250	0,1250	0,1043
3KDN	A	12	231	3	124	0,6568	0,0882	0,0882	0,1844
3KDV	A	7	67	2	39	0,6435	0,1522	0,1522	0,2247
3KF8	A	7	149	11	24	0,8168	0,2258	0,2258	0,1982
3KG8	A	6	230	4	15	0,9255	0,2857	0,2857	0,3804
3KGW	A	15	205	4	93	0,6940	0,1389	0,1389	0,2391
3KHN	A	10	111	1	25	0,8231	0,2857	0,2857	0,4480
3KHP	A	8	197	8	43	0,8008	0,1569	0,1569	0,1944
3KHU	A	18	280	7	86	0,7621	0,1731	0,1731	0,2685



3KIO	C	24	31	1	56	0,4911	0,3000	0,3000	0,2916
3KIP	A	11	56	1	55	0,5447	0,1667	0,1667	0,2506
3KJI	A	3	184	12	14	0,8779	0,1765	0,1765	0,1221
3KKB	A	11	64	2	46	0,6098	0,1930	0,1930	0,2639
3KOM	A	16	470	3	79	0,8556	0,1684	0,1684	0,3364
3KOP	A	8	81	1	47	0,6496	0,1455	0,1455	0,2637
3KP7	A	24	63	3	44	0,6493	0,3529	0,3529	0,3832
3KQG	A	10	104	3	33	0,7600	0,2326	0,2326	0,3287
3KR3	D	4	26	4	20	0,5556	0,1667	0,1667	0,0466
3KRM	A	3	79	5	50	0,5985	0,0566	0,0566	-0,0061
3KRY	A	6	113	2	24	0,8207	0,2000	0,2000	0,3240
3KSU	A	6	168	1	27	0,8614	0,1818	0,1818	0,3556
3KV4	A	8	336	10	38	0,8776	0,1739	0,1739	0,2230
3KW0	A	6	110	0	42	0,7342	0,1250	0,1250	0,3008
3KW2	A	6	173	4	29	0,8443	0,1714	0,1714	0,2606
3KYD	B	12	344	7	71	0,8203	0,1446	0,1446	0,2396
3KYI	B	4	72	4	11	0,8352	0,2667	0,2667	0,2804
3L07	A	5	200	2	38	0,8367	0,1163	0,1163	0,2429
3L0L	A	6	170	7	43	0,7788	0,1224	0,1224	0,1467
3L28	A	6	72	2	36	0,6724	0,1429	0,1429	0,2197
3L4Q	C	7	114	13	26	0,7563	0,2121	0,2121	0,1343
3L5O	A	5	172	3	19	0,8894	0,2083	0,2083	0,3170
3L6D	A	22	161	7	76	0,6880	0,2245	0,2245	0,2830
3L6U	A	5	196	4	31	0,8517	0,1389	0,1389	0,2232
3L6X	A	8	298	6	31	0,8921	0,2051	0,2051	0,2974
3L70	A	46	206	35	96	0,6580	0,3239	0,3239	0,2114
3L70	C	68	152	60	87	0,5995	0,4387	0,4387	0,1614
3L70	G	35	10	12	23	0,5625	0,6034	0,6034	0,0526
3L70	J	20	14	10	17	0,5574	0,5405	0,5405	0,1210
3L9Y	A	2	101	3	22	0,8047	0,0833	0,0833	0,1098
3LBX	A	9	103	9	16	0,8175	0,3600	0,3600	0,3197
3LBX	B	11	120	14	28	0,7572	0,2821	0,2821	0,2110
3LFV	A	11	274	4	63	0,8097	0,1486	0,1486	0,2708
3LJS	A	6	245	9	21	0,8932	0,2222	0,2222	0,2449
3LKK	A	7	144	5	39	0,7744	0,1522	0,1522	0,2096
3LXX	A	2	28	2	32	0,4688	0,0588	0,0588	-0,0162
3LL3	A	4	370	5	31	0,9122	0,1143	0,1143	0,1925
3LM4	A	16	210	3	72	0,7508	0,1818	0,1818	0,3137
3LM6	A	6	208	3	56	0,7839	0,0968	0,0968	0,1937
3LN4	A	9	174	0	62	0,7469	0,1268	0,1268	0,3057
3LPM	A	5	172	6	21	0,8676	0,1923	0,1923	0,2342
3LR5	A	6	81	6	22	0,7565	0,2143	0,2143	0,2040
3LRU	A	15	72	8	60	0,5613	0,2000	0,2000	0,1406
3LVM	A	8	262	5	65	0,7941	0,1096	0,1096	0,1946
3LYP	A	7	144	3	14	0,8988	0,3333	0,3333	0,4374
3LYX	A	8	84	2	22	0,7931	0,2667	0,2667	0,3797
3LZ7	A	4	73	2	46	0,6160	0,0800	0,0800	0,1222
3MOA	D	3	44	5	17	0,6812	0,1500	0,1500	0,0680
3MOD	C	21	13	3	26	0,5397	0,4468	0,4468	0,2324
3M1C	A	20	545	16	74	0,8626	0,2128	0,2128	0,2834
3M1C	B	20	44	3	71	0,4638	0,2198	0,2198	0,1983
3M1D	A	2	49	8	10	0,7391	0,1667	0,1667	0,0283
3M1I	B	13	59	1	51	0,5806	0,2031	0,2031	0,2944
3M1R	A	20	155	0	94	0,6506	0,1754	0,1754	0,3305

3M21	A	10	14	0	41	0,3692	0,1961	0,1961	0,2234
3M52	A	6	67	3	32	0,6759	0,1579	0,1579	0,1988
3M5V	A	8	187	2	58	0,7647	0,1212	0,1212	0,2496
3M84	A	9	199	5	66	0,7455	0,1200	0,1200	0,1939
3M8E	A	6	40	3	25	0,6216	0,1935	0,1935	0,1868
3M8J	A	6	59	5	15	0,7647	0,2857	0,2857	0,2667
3M92	A	11	32	2	34	0,5443	0,2444	0,2444	0,2479
3MBQ	A	10	51	2	66	0,4729	0,1316	0,1316	0,1590
3MCE	A	9	19	0	29	0,4912	0,2368	0,2368	0,3062
3MDO	A	9	266	3	48	0,8436	0,1579	0,1579	0,2960
3MEQ	A	12	205	4	67	0,7535	0,1519	0,1519	0,2586
3MGD	A	6	118	7	11	0,8732	0,3529	0,3529	0,3343
3MGJ	A	4	67	3	14	0,8068	0,2222	0,2222	0,2674
3MI9	A	4	240	12	31	0,8502	0,1143	0,1143	0,0951
3MIZ	A	5	206	4	32	0,8543	0,1351	0,1351	0,2211
3MKR	A	16	175	8	62	0,7318	0,2051	0,2051	0,2557
3MM5	B	43	126	1	149	0,5298	0,2240	0,2240	0,3067
3MMH	A	3	104	4	29	0,7643	0,0938	0,0938	0,1093
3MNP	A	2	200	10	15	0,8899	0,1176	0,1176	0,0824
3MOL	A	4	137	0	16	0,8981	0,2000	0,2000	0,4232
3MP7	A	14	249	63	31	0,7367	0,3111	0,3111	0,0881
3MRU	A	3	377	5	38	0,8983	0,0732	0,0732	0,1305
3MT1	A	14	232	1	65	0,7885	0,1772	0,1772	0,3515
3MTI	A	4	106	3	21	0,8209	0,1600	0,1600	0,2319
3MVP	A	8	145	5	31	0,8095	0,2051	0,2051	0,2747
3MWH	A	1	66	5	21	0,7204	0,0455	0,0455	-0,0432
3MXN	B	8	81	2	28	0,7479	0,2222	0,2222	0,3280
3MYB	A	17	128	3	84	0,6250	0,1683	0,1683	0,2569
3N07	A	2	126	6	18	0,8421	0,1000	0,1000	0,0826
3N2Y	A	4	232	11	25	0,8676	0,1379	0,1379	0,1253
3N6J	A	11	323	2	55	0,8542	0,1667	0,1667	0,3354
3N74	A	15	117	1	62	0,6769	0,1948	0,1948	0,3319
3N7Z	A	6	264	1	75	0,7803	0,0741	0,0741	0,2114
3NAQ	A	13	227	3	59	0,7947	0,1806	0,1806	0,3187
3NB0	A	11	445	11	71	0,8476	0,1341	0,1341	0,1997
3NEK	A	4	164	4	21	0,8705	0,1600	0,1600	0,2294
3NFE	B	6	100	3	33	0,7465	0,1538	0,1538	0,2285
3NFU	A	3	313	4	34	0,8927	0,0811	0,0811	0,1504
3NHE	B	2	21	0	44	0,3433	0,0435	0,0435	0,1185
3NI7	A	4	136	3	21	0,8537	0,1600	0,1600	0,2461
3NJA	A	13	33	1	57	0,4423	0,1857	0,1857	0,2148
3NJC	A	8	82	2	34	0,7143	0,1905	0,1905	0,2907
3NK6	A	4	188	7	32	0,8312	0,1111	0,1111	0,1281
3NKD	A	4	176	5	46	0,7792	0,0800	0,0800	0,1115
3NKZ	A	16	30	1	39	0,5349	0,2909	0,2909	0,3118
3NMT	B	3	206	8	19	0,8856	0,1364	0,1364	0,1365
3NMV	B	3	189	6	20	0,8807	0,1304	0,1304	0,1539
3NNM	A	6	245	2	20	0,9194	0,2308	0,2308	0,3876
3NO9	A	22	246	4	119	0,6854	0,1560	0,1560	0,2699
3NOQ	A	9	104	1	54	0,6726	0,1429	0,1429	0,2728
3NPK	A	11	178	5	47	0,7842	0,1897	0,1897	0,2787
3NQO	A	23	93	2	56	0,6667	0,2911	0,2911	0,3834
3NQW	A	7	116	3	28	0,7987	0,2000	0,2000	0,2973
3NRB	A	15	167	2	69	0,7194	0,1786	0,1786	0,3136

3NRX	A	3	82	8	28	0,7025	0,0968	0,0968	0,0120
3NUW	A	4	175	4	67	0,7160	0,0563	0,0563	0,0871
3NWO	B	8	141	15	18	0,8187	0,3077	0,3077	0,2228
3NWZ	A	12	76	3	53	0,6111	0,1846	0,1846	0,2389
3NYM	A	11	60	3	26	0,7100	0,2973	0,2973	0,3474
3NZE	A	2	205	4	14	0,9200	0,1250	0,1250	0,1689
3OOH	A	19	306	5	67	0,8186	0,2209	0,2209	0,3541
3OOT	A	2	122	5	29	0,7848	0,0645	0,0645	0,0485
3O4W	A	10	349	0	42	0,8953	0,1923	0,1923	0,4143
3O53	A	0	242	8	14	0,9167	0,0000	0,0000	-0,0418
3O78	A	15	259	3	92	0,7425	0,1402	0,1402	0,2712
3O7M	A	10	88	3	37	0,7101	0,2128	0,2128	0,2917
3OA2	A	11	209	2	45	0,8240	0,1964	0,1964	0,3536
3OBK	A	26	146	0	137	0,5566	0,1595	0,1595	0,2869
3OD1	A	18	229	6	91	0,7180	0,1651	0,1651	0,2550
3ODU	A	8	269	28	22	0,8471	0,2667	0,2667	0,1590
3OHE	A	5	69	1	20	0,7789	0,2000	0,2000	0,3362
3OHM	A	3	228	5	42	0,8309	0,0667	0,0667	0,0996
3OJA	A	41	306	14	63	0,8184	0,3942	0,3942	0,4488
3OLZ	A	11	262	8	36	0,8612	0,2340	0,2340	0,3060
3ONQ	A	6	172	5	46	0,7773	0,1154	0,1154	0,1707
3OO2	A	9	234	1	73	0,7666	0,1098	0,1098	0,2643
3OOW	A	15	46	2	72	0,4519	0,1724	0,1724	0,1886
3OQB	A	25	240	1	74	0,7794	0,2525	0,2525	0,4246
3OQL	A	12	171	1	39	0,8206	0,2353	0,2353	0,4114
3OT1	A	2	135	1	31	0,8107	0,0606	0,0606	0,1599
3OZ9	H	8	136	5	36	0,7784	0,1818	0,1818	0,2438
3OZ9	L	10	131	5	46	0,7344	0,1786	0,1786	0,2402
3POW	A	2	295	4	43	0,8634	0,0444	0,0444	0,0800
3P2O	A	6	188	7	37	0,8151	0,1395	0,1395	0,1755
3P6K	A	8	246	0	64	0,7987	0,1111	0,1111	0,2969
3P7H	A	12	73	2	30	0,7265	0,2857	0,2857	0,3829
3P7M	A	11	187	4	76	0,7122	0,1264	0,1264	0,2165
3P8L	A	6	178	3	43	0,8000	0,1224	0,1224	0,2236
3PBL	A	11	251	17	33	0,8397	0,2500	0,2500	0,2272
4FBP	A	87	151	15	9	0,9084	0,9063	0,9063	0,8062
6FAB	L	53	120	5	13	0,9058	0,8030	0,8030	0,7891
7FAB	L	38	137	4	8	0,9358	0,8261	0,8261	0,8232