UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL


# Previsão da Localização Subcelular de Proteínas Humanas com base em Aprendizagem Automática

**Pedro de Almeida Martins**

MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIZAÇÃO EM BIOINFORMÁTICA

Dissertação orientada por:
Prof. Doutor Francisco José Moreira Couto
Prof.ª Doutora Margarida Sofia Pereira Duarte Amaral

2017

# Agradecimentos

A realização desta dissertação não teria sido possível sem o apoio de algumas pessoas que, de forma directa ou indirecta, contribuíram para a sua execução. Como tal, é com grande agrado que expresso o meu profundo e sincero agradecimento:

Aos meus orientadores, Professor Doutor Francisco M Couto e Professora Doutora Margarida D Amaral, por terem aceite supervisionar o meu trabalho e por todo o apoio dado no desenvolvimento do mesmo;

Ao LaSIGE (Large-Scale Informatics Systems Laboratory), pela oportunidade de fazer parte do grupo de investigação durante a realização deste projecto;

Ao BioISI (Instituto de Biossistemas e Ciências Integrativas), em nome da Professora Doutora Margarida D Amaral, Doutor Luka A Clarke e Doutor Hugo Botelho, pela disponibilização da lista de genes essenciais para a execução deste projecto;

Aos meus pais, Fátima e Fernando, aos meus irmãos, João e Mariana, e a todos os meus amigos que, de uma forma ou de outra, contribuíram para aquilo que eu sou hoje; pela disponibilidade, amizade, paciência e apoio.

# Resumo

Conhecer a localização subcelular de um dado produto génico (i.e., onde a proteína codificada pelo gene está localizada) é particularmente importante para a anotação funcional das proteínas. Para lidar com o aumento exponencial do número de proteínas descobertas recentemente, foram desenvolvidos métodos computacionais capazes de prever a localização subcelular de proteínas. Uma vez que as proteínas localizadas em determinados compartimentos intracelulares possuem características em comum, os algoritmos de aprendizagem automática podem ser úteis para essa previsão.

O objectivo principal deste estudo foi prever a localização subcelular de proteínas codificadas por 800 genes humanos envolvidos no tráfego da CFTR (regulador de condutância transmembranar de fibrose quística), uma proteína que, quando mutada, causa a doença genética Fibrose Quística.

Neste projecto foram analisados os resultados de diferentes algoritmos de classificação disponíveis no MEKA, assim como diferentes métodos de construção de vectores representativos de proteínas. Por um lado, estes vectores foram construídos seguindo duas abordagens baseadas em *Gene Ontology* (GO): (1) valor 1-0 (presença ou ausência do termo GO) e (2) frequência dos termos GO. Por outro lado, foram consideradas três dimensões distintas dos vectores - 10165-D (todos os termos GO distintos para as proteínas em estudo), 429-D (termos GO essenciais obtidos pelo classificador mEN) e 87-D (termos GO essenciais obtidos pelo classificador mLASSO). Após a extracção dos termos GO e construção dos vectores representativos das proteínas, a localização subcelular das proteínas foi prevista através de três métodos de transformação do problema - *Binary Relevance* (BR), *Classifier Chain* (CC) e *Label Cardinality* (LC) - juntamente com três classificadores single-label - SMO, PART e J48. Estes classificadores foram avaliados através dos métodos *10-fold cross-validation* e *Leave-one-out cross-validation*. Os sete melhores modelos de previsão criados pelo MEKA atingiram uma

taxa global de sucesso entre 69,2 e 72,3% (*overall actual accuracy*) e 76,1 e 80,3% (*overall locative accuracy*).

**Palavras Chave:** Aprendizagem Automática, Localização Subcelular de Proteínas, Gene Ontology (GO), MEKA, Métodos de Transformação do Problema

# Abstract

To know the subcellular localization of a given gene product (i.e., where the protein codified by the gene is located) is particularly helpful to the functional annotation of proteins. In order to better deal with the exponential increase of newly discovered proteins, several computational methods, capable of predicting proteins' subcellular localization, were developed. Since proteins located in particular intracellular compartments share certain common features, Machine Learning (ML) algorithms are useful to predict it.

The goal of this study was to predict the subcellular localization of proteins encoded by 800 human genes involved in CFTR (cystic fibrosis transmembrane conductance regulator) traffic, a protein that, when mutated, causes Cystic Fibrosis, a genetic disease.

On this project we analyzed different classification algorithms available in MEKA, as well as different methods of construction of vectors representative of proteins. On one hand, the vectors were built following two approaches based on Gene Ontology (GO): (1) 1-0 Value (presence or absence of GO terms) and (2) term-frequency (number of occurences of individual go terms). On the other hand, three different dimensions of the vectors were considered: 10165-D (all distinct GO terms), 429-D (essencial GO terms selected by mEN classifier) and 87-D (essencial GO terms selected by mLASSO classifier). After extracting the GO terms and building the vectors, the subcellular localization of proteins was predicted using three methods of problem transformation - Binary Relevance (BR), Classifier Chain (CC) and Label Cardinality (LC) – along with three single-label classifiers - SMO, PART and J48. These classifiers were evaluated by the methods of the 10-fold cross-validation and Leave-one-out cross-validation. The seven best predictive models created by MEKA achieved an overall success rate between 69.2 and 72.3% (overall actual accuracy) and between 76.1 and 80.3% (overall locative accuracy).

**Keywords:** Machine Learning, Protein Subcellular Localization, Gene Ontology (GO), MEKA, Problem Transformation Methods

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas

**AL-KNN** Accumulation-label K-nearest neighbor.

**BLAST** Basic Local Alignment Search Tool.

**BR** Binary Relevance.

**CC** Classifier Chain.

**CFTR** Cystic Fibrosis Transmembrane Conductance Regulator.

**DNA** Deoxyribonucleic acid.

**EN** Elastic net.

**GO** Gene Ontology.

**GOA** Gene Ontology Annotation.

**ISF** Inverse sequence frequency.

**LASSO** Least absolute shrinkage and selection operator.

**LC** Label Cardinality ou Label Powerset (LP).

**LOOCV** Leave-one-out cross-validation.

**OAA** Overall Actual Accuracy.

**OET-KNN** Optimized evidence-theoretic K-nearest neighbor.

**OLA** Overall Locative Accuracy.

**PSI-BLAST** Position-Specific Iterated BLAST.

**RPS-BLAST** Reverse PSI-BLAST.

**SMO** Sequential minimal optimization.

**SVM** Support Vector Machines.

**TF** Term Frequency.

**TF-ISF** Term frequency-inverse sequence frequency.

**UniProtKB** UniProt Knowledgebase.

# Capítulo 1

# Introdução

A célula é a unidade básica e estrutural de todos os organismos vivos e é capaz de crescer e de reproduzir-se de forma independente. Um ser humano adulto é composto por aproximadamente $10^{14}$ células [4] e cada célula contém cerca de $10^9$ moléculas proteicas localizadas em diferentes compartimentos ou organelos celulares [5].

As proteínas, que são macromoléculas biológicas essenciais para todos os organismos, possuem uma grande variedade de funções biológicas e participam em praticamente todos os processos celulares: servem como componentes estruturais de células e tecidos; participam no transporte e armazenamento de pequenas moléculas (e.g. transporte de oxigénio pela hemoglobina); transmitem informações entre as células (e.g. hormonas) e contribuem na protecção contra infecções (e.g. anticorpos) [6]. No entanto, a propriedade fundamental das proteínas é a sua habilidade de actuar como enzimas, catalisando praticamente todas as reacções dos sistemas biológicos [6, 7].

A maioria das actividades biológicas realizadas pelas proteínas ocorre nos organelos. Estes são componentes celulares ou localizações subcelulares dentro da célula que possuem funções específicas [7]. Por exemplo, o núcleo celular, que contém a maior parte do material genético (DNA), é responsável pelo controlo das actividades da célula através da regulação da expressão génica; a membrana celular ou plasmática, que separa o ambiente intracelular do espaço extracelular, tem a função de revestir e proteger a célula, possuindo uma permeabilidade selectiva; o citoplasma, que ocupa a maior parte do volume celular, é o local onde a maioria das actividades celulares, como a divisão celular e as vias metabólicas, ocorrem; a mitocôndria é responsável pela produção e fornecimento da maior parte da energia utilizada nas actividades celulares e o complexo

de Golgi é um organelo particularmente importante na secreção celular [5, 7]. Praticamente todas estas funções, que são críticas para a sobrevivência da célula, são realizadas pelas proteínas [5].

## 1.1 Motivação

Um dos objectivos fundamentais da biologia celular e da proteómica é a identificação da localização subcelular das proteínas, pois o seu papel na célula está intimamente correlacionado com o compartimento ou organelo em que esta reside [8, 9]. A informação sobre a localização subcelular é, assim, importante para a anotação das proteínas, identificação dos alvos dos fármacos e para o desenvolvimento dos mesmos [7]. A localização das proteínas em contextos fisiológicos apropriados dentro da célula é essencial para que estas possam exercer as suas funções biológicas [3, 7, 10–12]. Por outro lado, de acordo com Wan et al. [11], uma localização diferente da desejada está intimamente relacionada com um grande leque de doenças, como Alzheimer [13], pedra nos rins [14], tumores hepáticos primários [15], cancro da mama [16], pré-eclampsia [17], síndrome de Bartter [18] e fibrose quística [19].

Com o crescimento exponencial do número de novas sequências proteicas descobertas na era pós-genómica, surgiu a necessidade de desenvolver métodos computacionais capazes de prever a localização subcelular das proteínas, tornando este processo mais rápido e eficiente. Uma vez que as proteínas localizadas em determinados compartimentos intracelulares possuem características comuns, os algoritmos de aprendizagem automática podem ser úteis para essa previsão.

## 1.2 Objectivos

**Objectivo Principal:** Prever a localização subcelular de proteínas codificadas por 800 genes humanos envolvidos no tráfego da CFTR (Cystic Fibrosis Transmembrane Conductance Regulator), uma proteína que, quando mutada, causa a doença genética Fibrose Quística.

Complementarmente, são, também, objectivos deste trabalho a identificação e comparação dos diferentes serviços web capazes de prever a localização subcelular de proteínas

humanas, dos algoritmos de classificação e dos métodos de construção de vectores representativos das proteínas.

## 1.3 Metodologia

Este estudo consistiu na previsão da localização subcelular de proteínas humanas através de algoritmos disponíveis no software MEKA [20, 21], uma extensão do WEKA [20] para previsão *multi-label*. Para isso, foi utilizado um conjunto de dados de treino constituído por 3106 proteínas humanas, distribuídas por 14 localizações subcelulares, e um conjunto de dados para previsão constituído por 799 proteínas humanas. Os vectores representativos das proteínas foram construídos seguindo duas abordagens baseadas nas anotações do Gene Ontology (GO): (1) valor 1-0 (presença ou ausência do termo GO) e (2) frequência dos termos GO. Por outro lado, foram consideradas três dimensões distintas dos vectores: 10165-D (todos os termos GO distintos para as proteínas em estudo), 429-D (termos GO essenciais obtidos pelo classificador mEN [10]) e 87-D (termos GO essenciais obtidos pelo classificador mLASSO [10]). No processo de aprendizagem foram utilizados três métodos de transformação do problema - Binary Relevance (BR), Classifier Chain (CC) e Label Cardinality/Label Powerset (LC) - juntamente com três classificadores *single-label* - SMO (Sequential minimal optimization), PART e J48. Estes classificadores foram avaliados através do método *10-fold cross-validation* e, os que apresentaram melhor desempenho, foram re-avaliados pelo método *Leave-one-out cross-validation*.

## 1.4 Contribuições

Este projecto teve as seguintes contribuições:

**Modelos de previsão:** foram gerados 7 modelos de previsão capazes de prever a localização subcelular de proteínas humanas, com uma taxa de sucesso global entre 69,2 e 72,3% (*overall actual accuracy*) e entre 76,1 e 80,3% (*overall locative accuracy*), de acordo com os 14 compartimentos subcelulares em estudo. Estes modelos foram desenvolvidos através de métodos de transformação do problema e têm como base a informação *Gene Ontology*. Link GitHub: `https://github.com/p-a-martins/7pred`.

**Previsão da localização de 799 proteínas:** foram previstas as localizações subcelulares de 799 proteínas codificadas por genes humanos envolvidos no tráfego da CFTR.

## 1.5 Estrutura

Este trabalho está dividido em vários capítulos e secções, nomeadamente:

**O Capítulo 2** introduz os diferentes métodos computacionais existentes para previsão da localização subcelular de proteínas: métodos baseados na sequência (Secção 2.1) e métodos baseados no conhecimento (Secção 2.1.1); distingue os conceitos de previsão *single-label* e *multi-label* (Secção 2.2); enumera os métodos de classificação (Secção 2.3) e de avaliação estatística (Secção 2.4), assim como as métricas de desempenho (Secção 2.5); descreve os principais serviços web capazes de prever a localização proteínas (Secção 2.6);

**O Capítulo 3** descreve os conjuntos de dados utilizados e as metodologias adoptadas para extracção dos termos GO e construção dos vectores representativos das proteínas; analisa os resultados da previsão da localização subcelular das proteínas;

**O Capítulo 4** apresenta os principais resultados e conclusões do estudo.

# Capítulo 2

# Enquadramento Teórico

A localização subcelular das proteínas pode ser determinada tanto por **métodos laboratoriais convencionais** como por **métodos computacionais**. Os primeiros são essenciais para a construção de bases de dados de localização de elevada qualidade, como a Human Protein Atlas[1] [3, 7, 10]. Assim, recorrendo a técnicas de engenharia genética é possível determinar a localização subcelular de proteínas através de diferentes técnicas [7], como a (1) **microscopia de fluorescência** (criação de uma proteína de fusão, constituída pela proteína de interesse ligada a um gene "repórter") [7, 22–24]; (2) a **microscopia imunoelectrónica** (utilização de anticorpos conjugados com ouro coloidal) [7, 25] ou a (3) **marcação fluorescente com biomarcadores** (utilização de marcadores compartimentais conhecidos para diferentes regiões celulares, marcados com fluorescência) [7, 26]. No entanto, estes são procedimentos demorados, laboriosos e com elevado custo [3, 7, 10, 27, 28].

## 2.1 Métodos Computacionais

Com a avalanche de proteínas geradas na era pós-genómica, surgiu a necessidade de criar métodos computacionais capazes de identificar rápida e eficazmente várias características biológicas das novas proteínas descobertas, nomeadamente a localização subcelular das mesmas. O rápido progresso da aprendizagem automática, juntamente com aumento do número de proteínas com localização determinada experimentalmente, tornou possível e promissora a previsão da localização subcelular de proteínas através

[1] `http://www.proteinatlas.org/`

de métodos computacionais [7]. De acordo com Chou et al. [28] e Xiao et al. [29], esses métodos foram desenvolvidos seguindo, principalmente, três direcções:

1. Aumentar o leque de localizações abrangidas de 2 [30], para 5 [31], para 12 [32, 33] e, finalmente, para 22 localizações [34];

2. Desenvolver métodos capazes de prever a localização com ênfase em diferentes organismos: humanos [1–3, 9, 10, 35–39], eucariotas [5, 8, 11, 12, 28, 34–36, 38, 39], plantas [11, 12, 27, 35, 36, 40–43], vírus [29, 35, 36, 42–45] e/ou bactérias gram-positivas [35, 36, 46–48] e gram-negativas [35, 36, 49–51];

3. Extrair informações úteis das proteínas através de diferentes métodos baseados na sequência (secção 2.1.1) ou no conhecimento/anotação (secção 2.1.2).

### 2.1.1 Métodos Baseados na Sequência

Os métodos baseados na sequência, que utilizam apenas a sequência aminoacídica da proteína como *input* [7], podem ser baseados na **composição** (secção 2.1.1.1) [30, 32, 52–57], na **homologia** (secção 2.1.1.2) [58–62], em **péptidos sinal** (secção 2.1.1.3) [63–65], no **domínio funcional** (secção 2.1.1.4) ou na **evolução sequencial** (secção 2.1.1.5).

#### 2.1.1.1 Métodos Baseados na Composição

Os métodos baseados na composição tomam partido da relação entre a localização subcelular das proteínas e a informação da composição incorporada na sequência dos aminoácidos [7]. Por sua vez, estes podem ser baseados na **composição aminoacídica (AA)** (secção 2.1.1.1.1) [30, 66], na **composição aminoacídica pareada (PairAA)** (secção 2.1.1.1.2) [30], na **composição aminoacídica pareada com *gaps* (GapAA)** (secção 2.1.1.1.3) [33, 67] ou na **composição pseudo-aminoacídica (PseAA)** (secção 2.1.1.1.4) [49, 54, 66, 68, 69].

Por outro lado, desde que o conceito de PseAA foi introduzido por Chou [54], este tem sido utilizado em vários estudos de proteómica. Assim, existem vários modelos discretos derivados do PseAA, como os métodos baseados no **domínio funcional** (secção 2.1.1.4) e os métodos baseados na **evolução sequencial** (secção 2.1.1.5). Actualmente, existe um serviço web, PseAA[1] [70], capaz de gerar 63 tipos diferentes de composição

[1] http://chou.med.harvard.edu/bioinf/PseAAC/

PseAA [71].

#### 2.1.1.1.1 Composição Aminoacídica (AA)

No método baseado na composição dos aminoácidos, a sequência proteíca é representada por um vector com 20 elementos, onde cada elemento corresponde à frequência da ocorrência de cada um dos 20 aminoácidos na sequência [7]. Sendo os resíduos de aminoácidos da proteína $P_i$ representados por [7]:

$$A_1 A_2 A_3 A_4 A_5 A_6 A_7 ... A_{l_i} ... A_{L_i} \tag{2.1}$$

onde $A_{l_i}$ representa o aminoácido na posição $l_i$ da proteína $i$, então, o vector $q_i^{AA}$ pode ser representado por

$$q_i^{AA} = [f_{i,1}, f_{i,2}, ..., f_{i,u}, ..., f_{i,20}]^T \tag{2.2}$$

onde $f_{i,u}$ representa a frequência de ocorrência do aminoácido $u$ ($u \in 1, 2, ..., 20$) na proteína $P_i$ e

$$\sum_{u=1}^{20} f_{i,u} = L_i \tag{2.3}$$

Existem vários estudos que aplicam este método, como por exemplo:

- Cedano et al. [31] aplicaram este método para prever a localização subcelular de 5 classes de proteínas;
- Reinhardt and Hubbard [72] usaram um algoritmo baseado na composição aminoacídica para prever a localização de proteínas procarióticas e eucarióticas;
- Chou and Elrod [32] recorreram a um algoritmo baseado na composição aminoacídica para prever a localização de proteínas em 12 organelos;

#### 2.1.1.1.2 Composição Aminoacídica Pareada (PairAA)

O método baseado na composição aminoacídica pareada (PairAA) utiliza a informação relativa à ordem dos aminoácidos na sequência, fazendo, para isso, a contagem do número de ocorrências dos pares de aminoácidos na proteína. Este método incorpora,

assim, a informação das frequências das co-ocorrências dos dipéptidos na sequência proteica [7]. Especificamente, para a proteína $i$, o vector $q_i^{PairAA}$ é definido por [7]:

$$q_i^{PairAA} = [f_{i,21}, f_{i,22}, ..., f_{i,(20+u\times v)}, ..., f_{i,420}]^T \tag{2.4}$$

onde $f_{i,(20+u\times v)}$ é o número de co-ocorrências dos aminoácidos $u$ e $v$ $(u, v \in 1, 2, ..., 20)$ em dipéptidos na proteína $i$, e

$$\sum_{u=1}^{20}\sum_{v=1}^{20} f_{i,(20+u\times v)} = L_i - 1 \tag{2.5}$$

Existem vários estudos que aplicam este método, como por exemplo:

- Nakashima and Nishikawa [30] usaram a razão de probabilidades para descriminar entre proteínas intra- e extracelulars solúveis, através de métodos baseados na composição AA e PairAA;
- Garg et al. [73] criaram o HSCLPred que utiliza métodos baseados na composição AA e PairAA para prever a localização subcelular de proteínas humanas.

#### 2.1.1.1.3 Composição Aminoacídica Pareada com Gaps (GapAA)

Com base no método anterior, surgiu uma nova abordagem que conta a frequência de pares de aminoácidos cujos os resíduos estão separados por uma ou mais posições de resíduos (*gaps*), conhecido como GapAA [7]. Nomeadamente, para a proteína $i$, o vector $q_i^{GapAA(k)}$ com $k$ *gaps* é definido por [7]:

$$q_i^{GapAA(k)} = [f_{i,(400\times k+21)}, f_{i,(400\times k+22)}, ..., f_{i,(400\times k+20+u\times v)}, ..., f_{i,(400\times k+420)}]^T \tag{2.6}$$

onde $f_{i,(400\times k+20+u\times v)}$ é o número de ocorrências dos aminoácidos $u$ e $v$ $(u, v \in 1, 2, ..., 20)$ que estão separados por $k$ resíduos de aminoácidos na proteína $i$, e

$$\sum_{u=1}^{20}\sum_{v=1}^{20} f_{i,(400\times k+20+u\times v)} = L_i - k - 1 \tag{2.7}$$

onde $k \in \{1, 2, ..., (L_i - 2)\}$.

Existem vários estudos que aplicam este método, como por exemplo:

- Park and Kanehisa [33] foram os primeiros a usar o GapAA para contar as frequências de aminoácidos cujos os resíduos estão separados por um ou mais posições (*gaps*);
- Chou and Cai [66] combinaram o AA, PairAA e GapAA para prever a localização proteica em leveduras;
- Park and Kanehisa [33], Chou and Shen [74], Wan et al. [75] demonstraram que a combinação dos métodos AA, PairAA e GapAA apresenta uma performance superior à combinação dos métodos AA e PairAA, que, por sua vez, apresenta melhores resultados do que a utilização do método AA.

#### 2.1.1.1.4 Composição Pseudo-Aminoacídica (PseAA)

O método de composição pseudo-aminoacídica (PseAA) proposto por Chou [54] usa o factor de correlação baseado na ordem da sequência para descobrir mais propriedades bioquímicas (e.g. hidrofobia, hidrofilia e massa da cadeida lateral dos aminoácidos) das sequências proteicas [7]. Concretamente, para a proteína $i$, o vector $q_i^{PseAA(\Omega)}$ é definido por [7]:

$$q_i^{PseAA(\Omega)} = [\widehat{f}_{i,1}, \widehat{f}_{i,2}, ..., \widehat{f}_{i,u}, ..., \widehat{f}_{i,20}, ..., p_{i,1}, p_{i,2}, ..., p_{i,m}, ..., p_{i,\Omega}]^T \tag{2.8}$$

onde $\widehat{f}_{i,u} = \frac{f_{i,u}}{L_i}$ é a frequência de ocorrência normalizada do aminoácido $u$ na proteína $i$, $\{p_{i,m}\}_{m=1}^{\Omega}$ é o factor de correlação da *m-tier* e $\Omega \in 1, 2, ..., (L_i - 1)$ é o número de factores de correlação da sequência. Os factores de correlação da *m-tier* incorporam as propriedades bioquímicas de todos os possíveis dipéptidos com $(m-1)$ *gaps* da sequência [7].

Existem vários estudos que aplicam este método, como por exemplo:

- Chou [54] recorreu ao método PseAA para prever a localização subcelular de proteínas, utilizando três propriedades bioquímicas: hidrofobia, hidrofilia e massa da cadeia lateral dos resíduos de aminoácidos.

#### 2.1.1.2 Métodos Baseados na Homologia

Os métodos baseados na homologia, como o Proteome Analyst [61], o PairProSVM [58], entre outros [59, 76, 77], partem do princípio que proteínas homólogas têm propriedades semelhantes e, que assim sendo, a homologia entre sequências proteicas pode ser usada para transferir anotações. Neste caso, tentam inferir a localização de proteínas através de anotações de proteínas similares [78], considerando que proteínas homólogas têm maior probabilidade de residir na mesma localização subcelular. Neste grupo de métodos, a sequência proteica de interesse é usada para identificar os seus homólogos numa base de dados proteica, através do BLAST (Basic Local Alignment Search Tool), e, então, a localização subcelular dessa proteína é determinada tendo em conta as localizações às quais pertencem as suas homólogas [7, 59, 77, 79, 80]. Este tipo de método pode atingir uma *accuracy* elevada se os homólogos da sequência de interesse forem encontrados nas bases de dados proteicas [7, 76]. No entanto, o mesmo não acontece se existir um grande número de proteínas sem homologia com proteínas com localização determinada experimentalmente.

#### 2.1.1.3 Métodos Baseados em Péptidos Sinal

Os métodos baseados em péptidos sinal, tais como PSORT [63], WoLF PSORT [81] and TargetP [64], prevêem a localização das proteínas através do reconhecimento dos péptidos sinal localizados na região N-terminal da sequência proteica [7, 82]. Após a síntese proteica, a maioria das proteínas são transportadas para localizações apropriadas da célula ou secretadas para o espaço extracelular [83]. A informação sobre para onde a proteína será transportada é, geralmente, encontrada ao nível da sequência primária, na forma de pequenos segmentos de sequência (entre 3 a 70 aminoácidos), denominados de péptidos sinal. Por outras palavras, no processo de triagem das proteínas, os péptidos sinal desempenham um papel essencial, agindo como "código postal" para as proteínas [84]. Depois de entrarem no compartimento celular apropriado, os péptidos sinal são clivados por peptidades de sinal [7, 85, 86]. Um dos passos mais importantes nos métodos baseados em péptidos sinal é prever o sítio de clivagem [7].

As sequências de péptidos sinal têm geralmente uma estrutura tripartida: região flanqueadora N-terminal (região N), região central hidrofóbica (região H) e uma região flanqueadora C-terminal (região C) [87–90]. Os aminoácidos com propriedades similares

podem ser categorizados de acordo com a sua hidrofobia e carga/polaridade, sendo que estas propriedades podem ser usadas para prever o local de clivagem. O grau de hidrofobia também difere dependendo da posição, fazendo com que esta seja uma característica útil para a previsão. Em suma, estas propriedades permitem que os sítios de clivagem sejam previstos computacionalmente. Os métodos para prever os sítios de clivagem podem ser classificados em 3 categorias: matrizes de pesos (PrediSi [91]), redes neuronais (SignalP 1.1 [65]) e modelos escondidos de Markov (SignalP 2.0 e SingalP 3.0 [92, 93]) [7].

Apesar dos métodos baseados nos péptidos sinal serem robustos e plausíveis biologicamente, estes apenas conseguem lidar com proteínas que possuam essa sequência sinal [7].

#### 2.1.1.4 Métodos Baseados no Domínio Funcional (FunD)

De acordo com Chou [71], o conceito de PseAA foi expandido para incorporar a informação do domínio funcional com o objectivo de prever a localização subcelular de proteínas [69, 94], tipos de proteínas membranares [95, 96], classes funcionais de enzimas [97], classes estruturais de proteínas [98] e tipos de proteases [99, 100].

Com base no facto de que as proteínas contêm, frequentemente, vários módulos ou domínios, cada um com origens evolucionárias e origens diferentes, foram desenvolvidas uma série de bases de dados FunD: COG [101], KOG [101], Pfam [102], SMART [103] e CDD [104]. Destas bases de dados, a CDD contém domínios importados da COG, Pfam e SMART e, portanto, é relativamente mais completa [104]. Como a versão 2.11 da CDD contém 17402 domínios característicos, então, usando cada um desses domínios como base do vector, uma dada proteína pode ser definida por um vector com 17402 elementos (17402-D) [1, 45, 47, 50]. Assim, após a utilização do RPS-BLAST (Reverse PSI-BLAST) [105] para fazer um alinhamento da sequência proteica em estudo com cada uma 17402 sequências domínio na base de dados CDD, a proteína $P$ no espaço $FunD$ pode ser definida como [1]

$$P_{FunD} = [\delta_1^D \ \delta_2^D \ ... \ \delta_i^D \ ... \ \delta_{17402}^D]^T \tag{2.9}$$

onde $T$ é o operador de transposição e

$$\delta_i^D = \begin{cases} 1, \ se\ houver\ correspondência \\ 0, \ caso\ contrário \end{cases} \tag{2.10}$$

Desta forma, para além da informação da ordem da sequência, também a informação funcional é incluída. Uma vez que a função da proteína está relacionada com a sua localização subcelular, a formulação $FunD$, que incorpora esses factores, está directamente correlacionada com a localização subcelular das proteínas [34].

#### 2.1.1.5 Métodos Baseados na Evolução Sequencial

De acordo com Chou [71], o conceito de PseAA também foi expandido para incorporar a informação sobre a evolução sequencial de forma a prever tipos de proteínas membranares [106], classes funcionais de enzimas [97] e tipos de proteases [99, 100].

A evolução nas sequências proteicas envolve mudanças em simples resíduos, inserções e delecções de vários resíduos, duplicação ou fusão de genes. No decurso do tempo, estas alterações acumulam-se, fazendo com que muitas similaridades entre as sequências aminoacídicas iniciais e as sequências aminoacídicas resultantes sejam eliminadas. Contudo, as proteínas ainda podem partilhar características em comum, como a localização subcelular [2, 34, 40, 45, 47, 50]. De forma a incorporar este tipo de informações, a evolução sequencial da proteína $P_i$ com $L$ aminoácidos pode ser representada por uma matriz $20 \times L$, onde um dado elemento da matriz, $E_{i\to j}$, é definido por [45]:

$$E_{i\to j} = \frac{E^0_{i\to j} - \overline{E}^0_j}{SD(\overline{E}^0_i)} \ (i = 1, 2, ..., L_i;\ j = 1, 2, ..., 20) \tag{2.11}$$

onde $E^0_{i\to j}$ representa a pontuação da substituição do resíduo aminoacídico $i$ pelo aminoácido $j$, numa determinada posição da sequência proteica durante o processo de evolução, calculada pelo PSI-BLAST (Position-Specific Iterated BLAST) [105] (geralmente inteiros positivos, se a mutação correspondente ocorre com mais frequência do que o esperado pelo acaso, ou negativos, caso contrário); $\overline{E}^0_j$ é a média para $E^0_{i\to j}(i = 1, 2, ..., L_i)$, enquanto que $SD$ corresponde ao desvio padrão [28].

A multiplicação desta matriz pela sua matriz de transposição gera uma matriz de $20 \times 20$ elementos. Por ser uma matriz simétrica, os elementos do triângulo superior são idênticos aos do triângulo inferior e, como tal, só será necessário utilizar os 20 elementos

da diagonal e os 190 elementos do triângulo superior para representar a proteína $P_i$ [45]:

$$P_{Evo} = [\psi_1^E \ \psi_2^E \ ... \ \psi_u^E \ ... \psi_{210}^E]^T \tag{2.12}$$

### 2.1.2 Métodos Baseados no Conhecimento

Os métodos baseados no conhecimento utilizam informações retiradas das bases de conhecimento, como **termos de GO**[1] [2, 7, 39, 69, 74, 75, 107–115], **Swiss-Prot *keywords*** [116, 117] e **PubMed *abstracts*** [61, 76]. Este tipo de métodos faz uso da correlação entre o conhecimento e anotação da proteína e a sua localização subcelular [7]. Diferentes estudos demonstraram que os métodos baseados na informação GO são superiores aos métodos baseados na sequência [38, 74, 118, 119].

GO é um vocabulário controlado usado para anotação de genes e produtos génicos e está dividido em três categorias: componente celular, processo biológico e função molecular [39, 42]. Componente celular refere-se às substâncias que constituem as células e organismos vivos (e.g. proteínas, ácidos nucleicos, membranas e organelos), sendo que a maioria estão localizadas dentro das células. Por sua vez, processo biológico é uma sequência de eventos realizados por conjuntos ordenados de funções moleculares. Por último, função molecular é um conjunto de actividades que podem ser realizadas por produtos génicos ao nível molecular [39, 42]. É importante salientar que a componente celular não é a única informação essencial para uma boa previsão da localização subcelular de proteínas. De acordo com estudos realizados por Lu and Hunter [120], a função molecular e o processo biológico também são importantes, especialmente para a previsão de proteínas localizadas no núcleo, espaço extracelular, membrana, mitocôndria, retículo endoplasmático e complexo de Golgi [10]. Isto é compreensível, pois, apesar dos termos GO pertencentes à categoria de função molecular ou processo biológico não terem implicações directas com a localização subcelular das proteínas, as proteínas apenas podem desempenhar as suas funções em certos contextos fisiológicos e participar em certos processos biológicos dentro dos compartimentos celular adequados. Assim, é lógico e aceitável que todas as categorias dos termos GO sejam consideradas para a previsão da localização de proteínas [7].

[1] http://www.geneontology.org/

Como resultado do GO Consortium, a base de dados Gene Ontology Annotation (GOA)[1] tornou-se um recurso útil para a investigação proteómica [121]. A base de dados fornece anotações estruturais para proteínas não redundantes de muitas espécies que estão presentes na UniProt Knowledgebase (UniProtKB) [122], usando termos GO padrão através de uma combinação de técnicas computacionais e manuais. A atribuição em larga escala de termos GO às entradas (números de acesso) da UniProtKB foi feita através da conversão de parte dos conhecimentos existentes na base de dados UniProtKB em termos GO [121], sendo que a base de dados GOA também inclui uma série de *cross-references* para outras bases de dados. Assim, a integração sistemática das bases de dados GOA e UniProtKB pode ser explorada para a localização subcelular de proteínas. Especificamente, dado um número de acesso de uma proteína, podem ser extraídos os termos GO da base de dados GOA [39, 42]. Na base de dados UniProtKB cada proteína tem um número de acesso e, na base de dados GOA, cada número de acesso pode estar associado a zero, um ou mais termos GO distintos. Reciprocamente, um termo GO pode estar associado a zero, um ou mais números de acesso de proteínas diferentes. Isto significa que o mapeamento entre os números de acesso e os termos GO é de muitos-para-muitos [42].

A previsão baseada em GO pode ser dividida em duas etapas: **extracção dos termos GO** (secção 2.1.2.1) e **construção dos vectores GO** (secção 2.1.2.2).

#### 2.1.2.1 Extracção dos termos GO

De acordo com Wan et al. [42], da perspectiva da extracção dos termos GO, os métodos baseados no conhecimento podem extrair os termos GO de três formas distintas:

- Usando a ferramenta InterProScan [123] para pesquisar contra um conjunto de dados de assinaturas proteicas [69, 75, 108, 124, 125]. Este tipo de método pode ser aplicado a todas as sequências proteicas, no entanto, geralmente, só é possível extrair um pequeno número de termos GO, o que poderá não ser suficiente para boa previsão da localização subcelular [7, 38];
- Usando os números de acesso das proteínas para pesquisar na base de dados GOA [8, 37, 49, 74, 126]. Este método funciona melhor que o anterior, contudo, não pode ser aplicado a proteínas que não estejam anotadas funcionalmente [7, 38];

[1] http://www.ebi.ac.uk/GOA/

- Usando os números de acesso de proteínas homólogas extraídas através do BLAST [127] para pesquisar na base de dados GOA [29, 36, 51, 118]. Este método permite a extensão dos métodos baseados nos termos GO a proteínas descobertas recentemente. Assim, é aplicável a todas as sequências proteicas e capaz de extrair mais termos GO, o que é essencial para uma boa performance [7, 38].

#### 2.1.2.2 Construção dos vectores de termos GO

Após a extracção dos termos GO, a forma como os vectores são construídos é de elevada importância. De acordo com Chou [128], as características de cada proteína podem ser representadas pelo modelo geral da composição pseudo-aminoacídica de Chou [54, 129]:

$$q_i = [\varphi_{i,1}, \varphi_{i,2}, ..., \varphi_{i,u}, ..., \varphi_{i,W}]^T \tag{2.13}$$

onde $T$ é o operador de transposição, $W$ é a dimensão do vector $q_i$ e as definições dos componentes de $W$, $\varphi_{i,u}(u = 1, ..., W)$, dependem das abordagens de extracção utilizadas [7]. Da perspectiva da construção dos vectores de termos GO, os métodos baseados no conhecimento são classificados em duas categorias. A primeira considera cada termo GO como uma base canónica de um espaço Euclidiano. Sendo $\mathbb{W}$ o conjunto de termos GO distintos extraídos (de acordo com os procedimentos descritos na secção 2.1.2.1), então, o vector GO será um espaço Euclidiano de dimensão $\mathbb{W}$. Para cada sequência proteica, o vector GO será construído através da correspondência entre o vector $\mathbb{W}$ e os termos GO associados a essa proteína [7]. Existem diversas abordagens para determinar os elementos do vector GO:

1. **Valor 1-0**: cada termo GO em $\mathbb{W}$ representa uma base canónica do espaço Euclidiano e a proteína é representada por um ponto nesse espaço, onde as coordenadas correspondem a 0 ou 1 [37, 46, 49, 74]. Assim, o vector GO para a proteína $i$ é definido por:

   $$q_i = [a_{i,1} \ ... \ a_{i,u} \ ... \ a_{i,w}] \tag{2.14}$$

   onde $a_{i,u}$ é igual a 1, se o termo GO $u$ estiver associado à proteína $i$, ou 0, caso contrário [7];

2. **Frequência do Termo (TF)**: cada termo GO em $\mathbb{W}$ representa uma base canónica do espaço Euclidiano e a proteína é representada por um ponto nesse espaço,

onde as coordenadas correspondem à frequência do termo GO [38, 39, 42], isto é, ao número de vezes que um termo GO está anotado a uma proteína através de diferentes evidências ou fontes de informação. Assim, o vector GO para a proteína $i$ é definido por:

$$q_i = [b_{i,1} \ \ldots \ b_{i,u} \ \ldots \ b_{i,w}] \tag{2.15}$$

onde $b_{i,u}$ é igual a $f_{i,u}$ (frequência do termo GO $u$ à proteína $i$), se o termo GO $u$ estiver associado à proteína $i$, ou 0, caso contrário [7].

Os métodos ISF (Inverse sequence frequency) e TF-ISF (Term frequency-inverse sequence frequency) são duas alternativas para construção de vectores de termos GO. No entanto, estes dois métodos demonstraram ser inferiores às abordagens Valor 1-0 e TF [39]. Outra possibilidade é atribuir a cada elemento do vector a percentagem de proteínas homólogas à proteína de interesse que contêm determinado termo GO. Esta categoria de métodos fornece uma grande cobertura de termos GO, mas a maioria deles podem ser irrelevantes para a tarefa de classificação. Para além disso, ignoram o facto de que um termo GO pode ser usado para anotar a mesma proteína múltiplas vezes em diferentes entradas na base de dados GOA [7].

A segunda categoria usa algoritmos genéticos para seleccionar os termos GO mais informativos. Por exemplo, ao explorar a semântica do GO é possível verificar a importância de cada termo [130]. Assim, esta categoria selecciona alguns termos GO informativos ou essenciais que estão directamente anotados aos compartimentos subcelulares de interesse. O problema deste tipo de métodos é que pode seleccionar apenas um pequeno número de termos GO, aumentando a probabilidade do vector ser nulo [7].

## 2.2 Previsão *Single-label versus* Previsão *Multi-label*

Muitos dos métodos existentes capazes de prever a localização subcelular de proteínas assumem que as mesmas residem em apenas uma localização subcelular [5, 27, 31, 33, 37, 44, 46, 49, 64, 131–134]. No entanto, as proteinas podem existir simultaneamente em, ou mover-se entre, duas ou mais localizações subcelulares diferentes [7, 135–138]. Estas proteínas com múltiplas localizações são particularmente interessantes, pois podem ter funções biológicas únicas dignas de especial atenção, sendo essenciais para a investigação básica e farmacológica [36, 66, 67, 139–141]. Na verdade, as proteínas com multíplas

localizações desempenham um papel importante em alguns processos metabólicos que acontecem em várias localizações subcelulares [7]. De acordo com Millar et al. [135], um número elevado de proteínas possuem múltiplas localizações subcelulares. Com base nas análises estatísticas da base de dados Swiss-Prot (versão 55.3), este tipo de proteínas podem atingir 20% de todas as proteínas humanas [1].

## 2.3 Métodos de Classificação

Os métodos existentes para classificação *multi-label* podem ser agrupados em duas categorias: **adaptação do algoritmo** ou **transformação do problema**. O primeiro adapta algoritmos específicos de previsão *single-label* de forma a estes poderem ser aplicados directamente na previsão *multi-label* [7]. Por outro lado, no segundo método, o problema de classificação *multi-label* é transformado em vários problemas de classificação *single-label* [142], não sendo, por isso, necessário modificar os classificadores tradicionais [7].

LC, BR e CC são exemplos de métodos de transformação do problema. O método LC considera cada combinação distinta de *labels* como uma única, permitindo que a mesma seja prevista por métodos *single-label*. Porém, o número de combinações aumenta exponencialmente com o número de *labels*. Por sua vez, BR transforma o problema de classificação *multi-label* em várias tarefas de classificação binárias, uma para cada *label*). Contudo, apesar de eficiente, não considera a correlação entre *labels*. Por último, o método CC é semelhante ao BR, no entanto, considera a correlação entre *labels*. Para isso, os classificadores binários são ligados numa cadeia ($c_1 \rightarrow c_2 \rightarrow ... \rightarrow c_i \rightarrow ... \rightarrow c_n$), onde o vector representativo da instância usado pelo classificador $c_i$ incorpora as *labels* previstas pelos classificadores precedentes.

Relativamente aos classificadores *single-label*, SMO é um algoritmo eficiente para a implementação da técnica SVM (Support Vector Machines). Já o J48 é uma implementação em java do algoritmo C4.5 que faz uso de uma estratégia *greedy* para induzir árvores de decisão para posterior classificação. Por sua vez, o PART usa a estratégia de divisão e conquista, construindo de forma parcial uma árvore de decisão C4.5 em cada iteração, transformando a "melhor folha" numa regra.

## 2.4 Métodos de Avaliação Estatística

Em Aprendizagem Automática é essencial especificar o algoritmo de classificação e o conjunto de dados de referência utilizados para testar o desempenho de um predictor [2]. Relativamente aos métodos de avaliação estatística, existem sobretudo três que são utilizados:

- **Método *Holdout***: O conjunto de dados é dividido em dois subconjuntos mutuamente exclusivos (treino e teste), sendo que um é utilizado para estimar os parâmetros de classificação e o outro para avaliar o desempenho do classificador;
- **Método *k-fold cross-validation***: O conjunto de dados é dividido em k subconjuntos mutuamente exclusivos do mesmo tamanho, sendo que k–1 subconjuntos são utilizados para estimar os parâmetros de classificação e o subconjunto restante para testar o classificador; este processo repete-se k vezes, sempre com um conjunto de treino diferente. Em particular, o método *10-fold cross-validation* divide o conjunto de dados em 10 subconjuntos do mesmo tamanho, utilizando 9 subconjuntos para treinar e 1 para testar o classificador; o processo repete-se 10 vezes;
- **Método Leave-one-out cross-validation (LOOCV)**: O conjunto de dados é dividido em N (número total de instâncias) subconjuntos mutuamente exclusivos; sendo que N–1 subconjuntos são utilizados para estimar os parâmetros de classificação e o subconjunto restante para testar o classificador; este processo repete-se N vezes, sempre com um conjunto de treino diferente.

## 2.5 Métricas de Desempenho

Comparativamente com a classificação tradicional *single-label*, a classificação *multi-label* requer métricas de desempenho mais avançadas. Considerando $L(\mathbb{Q}_i)$ e $M(\mathbb{Q}_i)$ como o *label set* real e o *label set* previsto para a proteína $i$, $\mathbb{Q}_i (i = 1, ..., N)$, respectivamente, então a *accuracy*, *precision*, *recall*, *F1-score (F1)* e *Hamming loss* são definidos por [11, 12]:

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{|M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)|}{|M(\mathbb{Q}_i) \cup L(\mathbb{Q}_i)|} \right) \tag{2.16}$$

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)|}{|M(\mathbb{Q}_i)|}\right) \tag{2.17}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)|}{|L(\mathbb{Q}_i)|}\right) \tag{2.18}$$

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2\,|M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)|}{|M(\mathbb{Q}_i)| + |L(\mathbb{Q}_i)|}\right) \tag{2.19}$$

$$HL = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|M(\mathbb{Q}_i) \cup L(\mathbb{Q}_i)| - |M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)|}{M}\right) \tag{2.20}$$

A *accuracy*, *precision*, *recall*, *F1-score* indicam o desempenho da classificação, que será tanto maior quanto maior forem os seus resultados. Por outro lado, quanto menor for o valor de *Hamming loss*, melhor será o desempenho do predictor [11, 12].

Duas métricas adicionais [29, 42], Overall Locative Accuracy (OLA) e Overall Actual Accuracy (OAA) (ou combinação exacta), são frequentemente utilizadas na previsão *multi-label* da localização subcelular de proteínas e são representadas por [11, 12]:

$$OLA = \frac{1}{\sum_{i=1}^{N}|L(\mathbb{Q}_i)|}\sum_{i=1}^{N}|M(\mathbb{Q}_i) \cap L(\mathbb{Q}_i)| \tag{2.21}$$

$$OAA = \frac{1}{N}\sum_{i=1}^{N}\Delta\left[M(\mathbb{Q}_i), L(\mathbb{Q}_i)\right] \tag{2.22}$$

onde

$$\Delta\left[M(\mathbb{Q}_i), L(\mathbb{Q}_i)\right] = \begin{cases} 1\ se\ M(\mathbb{Q}_i) = L(\mathbb{Q}_i) \\ 0\ caso\ contrário \end{cases} \tag{2.23}$$

Para estudar um sistema de proteínas onde estas podem ocorrer simultaneamente em duas ou mais localizações torna-se necessário definir o conceito de **proteína locativa**. Se uma proteína coexiste em duas localizações subcelulares, esta será contabilizada como duas proteínas locativas; se coexiste em três localizações subcelulares será contabilizada como três proteínas locativas; e assim sucessivamente [2]. O número total de proteínas locativas pode ser expressa por:

$$N(loc) = N(seq) + \sum_{m=1}^{M}(m-1)N(m) = \sum_{m=1}^{M} m \times N(m) \tag{2.24}$$

onde $N(loc)$ é o número total de proteínas locativas, $N(seq)$ é o número total de proteínas diferentes, $N(m)$ é o número de proteínas que existem em $m$ localizações e $M$ é o número total de localizações estudadas [29].

Assim, de acordo com a equação 2.21, considera-se que uma proteína locativa foi prevista correctamente se qualquer uma das *labels* previstas corresponder com qualquer uma das *labels* do *label set* real. Por outro lado, a equação 2.22 considera que uma proteína só é considerada correctamente prevista apenas se todas as *labels* previstas corresponderem com exatidão ao *label set* real (apenas quando todas as localizações subcelulares de uma dada proteína são previstas sem qualquer *over-prediction* ou *under-prediction*). Assim, OAA é mais rigorosa e objectiva do que OLA. Tal acontece porque OLA é mais susceptível a fornecer medidas de desempenho mais tendenciosas quando o predictor tende a *over-predict*, isto é, dando um grande $M(\mathbb{Q}_i)$ para muitas $\mathbb{Q}_i$. No caso extremo, se uma proteína for prevista como estando localizada em todas as $M$ localizações subcelulares, então, de acordo com a equação 2.21, OLA é igual a 100%. No entanto, as previsões são incorrectas e sem significado. Pelo contrário, OAA é igual a 0%, o que definitivamente reflete o desempenho real [11, 12].

## 2.6 Exemplos de Predictores

### 2.6.1 Cell-PLoc e Cell-PLoc 2.0

Cell-PLoc[1] [35] é um conjunto de seis predictores web: **Hum-mPLoc**[2] [9] (versão actualizada do Hum-PLoc [37]), Euk-mPLoc[3] [8] (versão actualizada do Euk-PLoc [5]), Plant-PLoc[4] [27], Virus-PLoc[5] [44], Gpos-PLoc[6] [46] e Gneg-PLoc[7] [49]. Estes são capazes de prever a localização subcelular de proteínas em humanos, eucariotas, plantas,

[1] http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc/
[2] http://www.csbio.sjtu.edu.cn/bioinf/hum-multi/
[3] http://www.csbio.sjtu.edu.cn/bioinf/euk-multi/
[4] http://www.csbio.sjtu.edu.cn/bioinf/plant/
[5] http://www.csbio.sjtu.edu.cn/bioinf/virus/
[6] http://www.csbio.sjtu.edu.cn/bioinf/Gpos/
[7] http://www.csbio.sjtu.edu.cn/bioinf/Gneg/

vírus, bactérias gram-positivas e bactérias gram-negativas, respectivamente. Em particular, Hum-mPLoc [9] e Euk-mPLoc [8] são capazes de lidar com proteínas que estão localizadas em, ou que se movem entre, duas ou mais regiões subcelulares (Tabela 2.1).

Tabela 2.1: Cell-PLoc e Cell-PLoc 2.0: predictores para previsão da localização subcelular de proteínas em seis organismos/espécies.

| | Predictor | Organismo/Espécie | Localização Múltipla | Número de Localizações | Referências |
|---|---|---|---|---|---|
| | **Hum-PLoc** | **Humanos** | **Não** | **12** | [37] |
| | Euk-PLoc | Eucariotas | Não | 18 | [5] |
| Cell-PLoc | **Hum-mPloc** | **Humanos** | **Sim** | **14** | [9, 35] |
| Cell-PLoc | Euk-mPLoc | Eucariotas | Sim | 22 | [8, 35] |
| Cell-PLoc | Plant-PLoc | Plantas | Não | 11 | [27, 35] |
| Cell-PLoc | Virus-PLoc | Vírus | Não | 7 | [35, 44] |
| Cell-PLoc | Gpos-PLoc | Bact. Gram+[a] | Não | 5 | [35, 46] |
| Cell-PLoc | Gneg-PLoc | Bact. Gram−[b] | Não | 8 | [35, 49] |
| Cell-PLoc 2.0 | **Hum-mPloc 2.0** | **Humanos** | **Sim** | **14** | [1, 36] |
| Cell-PLoc 2.0 | Euk-mPLoc 2.0 | Eucariotas | Sim | 22 | [34, 36] |
| Cell-PLoc 2.0 | Plant-mPLoc | Plantas | Sim | 12 | [36, 40] |
| Cell-PLoc 2.0 | Virus-mPLoc | Vírus | Sim | 6 | [36, 45] |
| Cell-PLoc 2.0 | Gpos-mPLoc | Bact. Gram+[a] | Sim | 4 | [36, 47] |
| Cell-PLoc 2.0 | Gneg-mPLoc | Bact. Gram−[b] | Sim | 8 | [36, 50] |

[a] bactéria gram-positiva;
[b] bactéria gram-negativa.

Os predictores Cell-PLoc [35], ao contrário do PSORT [143], TargetP [64] e PSORT-B [131], que cobrem 5 ou menos localizações subcelulares, conseguem prever até 22 localizações subcelulares (Tabela 2.1). Para além disso, os conjuntos de dados de referência possuem baixa identidade de sequência. Enquanto que os conjuntos de dados construídos por outros preditores permitem a inclusão de proteínas com 80% [33], 90% [72, 73] ou mais identidade de sequência, os predictores Cell-PLoc [35] não permitem que uma proteína tenha mais de 25% de identidade de sequência com outra que pertença à mesma localização subcelular [35].

Contudo, estes predictores têm a desvantagem de exigir o número de acesso das proteínas como *input*. Uma vez que muitas proteínas (e.g. proteínas sintéticas ou hipotéticas e proteínas descobertas recentemente) não têm número de acesso atribuído, a sua localização não pode ser prevista através da abordagem GO (secção 2.1.2). Nestes casos, os predictores utilizam a abordagem PseAA (secção 2.1.1.1.4). No entanto, apesar desta abordagem ter em conta alguns efeitos parciais da ordem da sequência, não tem

em conta as informações do domínio funcional e da evolução sequencial [35, 36].

Para contornar estes problemas, foi criado o Cell-PLoc 2.0[1] [36] que é constituído, igualmente, por seis predictores: **Hum-mPLoc 2.0**[2] [1], Euk-mPLoc 2.0[3] [34], Plant-mPloc[4] [40], Virus-mPLoc[5] [45], Gpos-mPloc[6] [47] e Gneg-mPLoc[7] [50]. Estes, por sua vez, apenas requerem a sequência proteica como *input*, utilizando os números de acesso das proteínas homólogas à sequência proteica. Para além disso, recorrem a uma abordagem PseAA mais avançada, como complemento da GO, que inclui as informações do domínio funcional (secção 2.1.1.4) e da evolução sequencial (secção 2.1.1.5) (Tabela 2.2). Complementarmente, todos os servidores web conseguem lidar com com proteínas que estão localizadas em, ou que se movem entre, duas ou mais regiões subcelulares. Por fim, usam um conjunto de dados de referência mais recente (versão 55.3 em vez da 50.7) [1, 34, 36, 40, 45, 47, 50].

Da perspectiva de classificação, os predictores CellPLoc 2.0 [36] utilizam um conjunto de classificadores formados pela fusão de muitos classificadores individuais básicos, operados pelo algoritmo OET-KNN (Optimized evidence-theoretic K-nearest neighbor)[8] (Tabela 2.2).

Posto isto, o processo de previsão através dos predictores CellPLoc 2.0 [36] pode ser dividido em duas situações (Figura 2.1):

1. Se a proteína em estudo puder ser representada através da abordagem GO (secção 2.1.2), então $P_{GO}$ será utilizado para determinar as localizações subcelulares das proteínas. O *output* será determinado pela fusão de predictores OET-KNN com diferentes parâmetros $K$;

2. Se a proteína em estudo não tiver homologias significativas com nenhuma proteína na base de dados Swiss-Prot ou se não tiver nenhum termo GO a ela associada, então $P_{FunD}$ (secção 2.1.1.4) e $P_{SeqEvo}$ (secção 2.1.1.5) serão utilizados para a previsão. O *output* será determinado pela fusão de predictores OET-KNN com diferentes parâmetros $K$ e $\lambda$.

[1] http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/
[2] http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/
[3] http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/
[4] http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/
[5] http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/
[6] http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/
[7] http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/
[8] formulação matemática detalhada em [144]

Tabela 2.2: Comparação entre os predictores Hum-mPLoc 2.0 (Cell-PLoc 2.0), iLoc-Hum (iLoc-Cell), mLASSO e mEN (PolyU-Loc) quanto à forma de construção dos vectores e de classificação.

| Predictor | Construção do vector | Classificação |
|---|---|---|
| Hum-mPLoc 2.0 [1] | GO Hom (1-0)[a] + FunD[b] + SeqEvo[c] | OET-KNN[d] |
| iLoc-Hum [2] | GO (%Hom)[e] + SeqEvo[c] | AL-KNN[f] |
| mLASSO [10] | GO (TF)[g] | LASSO[h] |
| mEN [10] | | EN[l] |

[a] Gene Ontology (valor 1-0, presença ou ausência do termo GO nas proteínas homólogas); a dimensão do vector depende do número total de termos GO presentes na versão da base de dados GOA utilizada (60020-D, para a versão 10-Mar-2008, no caso do Hum-mPLoc 2.0);
[b] Domínio Funcional;
[c] Evolução Sequencial;
[d] Optimized Evidence-Theoretic K-Nearest Neighbor;
[e] Gene Ontology (% de proteínas homologas com determinado termo GO); a dimensão do vector depende do número total de termos GO presentes na versão da base de dados GOA utilizada (11118-D, para a versão 30-Jul-2009, no caso do iLoc-Hum);
[f] Accumulation-Label K-Nearest Neighbor.
[g] Gene Ontology (Frequência do Termo);
[h] Least Absolute Shrinkage and Selection Operator;
[i] Elastic Net.

Como referido anteriormente, o Hum-PLoc [37], Hum-mPLoc [9] e Hum-mPLoc 2.0 [1] são capazes de prever a localização subcelular de proteínas humanas. Para além disso, os dois últimos são capazes de lidar com proteínas que existem em, ou que se movem entre, dois ou mais compartimentos subcelulares. Comparando a taxa de sucesso global destes (Tabela 2.3), Shen and Chou [1] determinaram que:

- A inclusão de proteínas com múltiplas localizações dificulta a tarefa de previsão;
- Apesar do Hum-mPLoc [9] ter alcançado uma taxa de sucesso global de cerca de 70%, este utiliza os números de acesso das proteínas para extracção dos termos GO. Nas mesmas condições que o Hum-mPLoc 2.0 [1] (mesmo conjunto de dados e apenas as sequências proteicas como *input*), a taxa de sucesso global desceria para cerca de 30%;
- A utilização das informações de domínio funcional (secção 2.1.1.4) e evolução sequencial (secção 2.1.1.5) demonstraram ser mais eficazes como complemento à abordagem GO (secção 2.1.2) do que a composição pseudo-aminoacídica (secção 2.1.1.1.4).

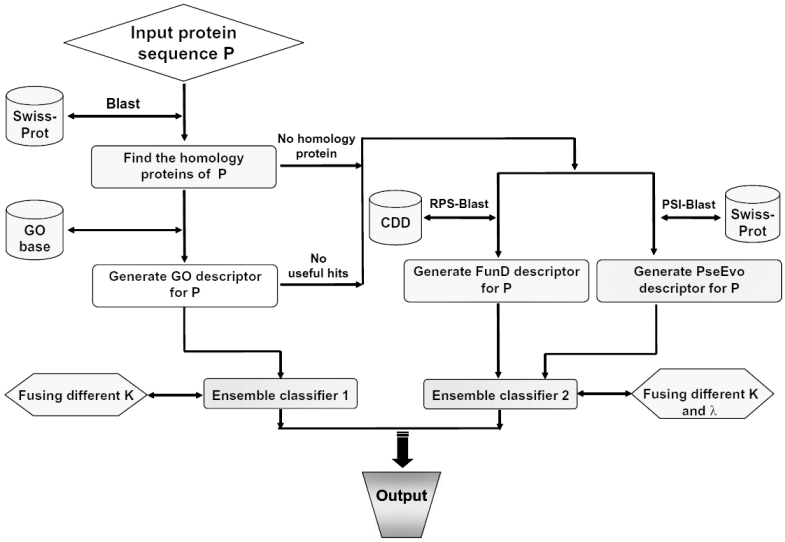


Figura 2.1: Representação do processo de previsão dos predictores Cell-PLoc 2.0 (Hum-mPLoc 2.0, Euk-mPLoc 2.0, Plant-mPLoc, Virus-mPLoc, Gpos-mPLoc e Gneg-mPLoc) (Figura extraída de Shen and Chou [1]).

De salientar também que os conjuntos de dados de referência do Hum-mPLoc [9] e do Hum-mPloc 2.0 [1] são bastante rigorosos, cobrindo 14 localizações subcelulares, onde nenhuma proteína tem mais de 25% de identidade de sequência com outra que esteja no mesmo subconjunto (localização subcelular). Quanto mais rigoroso for o conjunto de dados mais difícil será para melhorar a taxa de sucesso global [1].

### 2.6.2 iLoc-Cell

iLoc-Cell[1] é um conjunto de seis predictores: **iLoc-Hum**[2] [2], iLoc-Euk[3] [28], iLoc-Plant[4] [41], iLoc-Virus[5] [29], iLoc-Gpos[6] [48] e iLoc-Gneg[7] [51]. Estes, à semelhança do Cell-PLoc [35] e Cell-PLoc 2.0 [36] são capazes de prever a localização subcelular de proteínas em humanos, eucariotas, plantas, vírus, bactérias gram-positivas e bactérias gram-negativas, respectivamente. Para além disso, todos eles são capazes de lidar

[1] http://www.jci-bioinfo.cn/iLoc-Cell
[2] http://www.jci-bioinfo.cn/iLoc-Hum
[3] http://www.jci-bioinfo.cn/iLoc-Euk
[4] http://www.jci-bioinfo.cn/iLoc-Plant
[5] http://www.jci-bioinfo.cn/iLoc-Virus
[6] http://www.jci-bioinfo.cn/iLoc-Gpos
[7] http://www.jci-bioinfo.cn/iLoc-Gneg

Tabela 2.3: Comparação entre os diferentes predictores para previsão da localização subcelular de proteínas humanas, através do LOOCV[a].

| Versão | Predictor | Input | OAA[b] | OLA[c] | ACC[d] | HL[e] | Referências |
|---|---|---|---|---|---|---|---|
| 49.3[f] | Hum-PLoc | AC[g] + Seq[h] | - | 0.811 | - | - | [37] |
| 50.7[f] | Hum-mPLoc | | - | 0.708 | - | - | [9] |
| 55.3[f] | Hum-mPLoc | Seq[h] | - | 0.381 | - | - | [9] |
| | Hum-mPLoc 2.0 | | - | 0.627 | - | - | [1] |
| | iLoc-Hum | | 0.682 | 0.763 | - | - | [2] |
| | mLASSO | | 0.729 | 0.820 | 0.814 | 0.029 | [10] |
| | mEN | | 0.743 | 0.836 | 0.827 | 0.028 | [10] |

[a] Leave-one-out cross-validation;
[b] Overall Actual Accuracy;
[c] Overall Locative Accuracy;
[d] Accuracy
[e] Hamming Loss
[f] 49.3 (21-Mar-2006); 50.7 (19-Sept-2006) e 55.3 (29-Apr-2008)
[g] Número de acesso da proteína;
[h] Sequência proteica.

com proteínas que estão localizadas em, ou que se movem entre, duas ou mais regiões subcelulares (Tabela 2.4).

Tabela 2.4: iLoc: predictores para previsão da localização subcelular de proteínas em seis organismos/espécies.

| Predictor | Organismo/Espécie | Localização Múltipla | Número de Localizações | Referências |
|---|---|---|---|---|
| **iLoc-Hum** | **Humanos** | **Sim** | **14** | [2] |
| iLoc-Euk | Eucariotas | Sim | 22 | [28] |
| iLoc-Plant | Plantas | Sim | 12 | [41] |
| iLoc-Virus | Vírus | Sim | 6 | [29] |
| iLoc-Gpos | Bact. Gram+[a] | Sim | 4 | [48] |
| iLoc-Gneg | Bact. Gram−[b] | Sim | 8 | [51] |

[a] bactéria gram-positiva;
[b] bactéria gram-negativa.

No entanto, ao contrário dos predictores Cell-PLoc [35] e Cell-PLoc 2.0 [36], os predictores iLoc-Cell utilizam uma abordagem diferente para construção dos vectores de referência das proteínas. Em vez de atribuírem os valores 0 e 1 aos elementos do vector (presença ou ausência de termos GO), estes utilizam a percentagem de proteínas homólogas que estão associadas a cada termo GO [2]. Por outro lado, em vez de usarem tanto as informações sobre a evolução sequencial (secção 2.1.1.5) e sobre o domínio funcional (secção 2.1.1.4), o iLoc-Cell [2, 29, 41, 48, 51] utilizam apenas a primeira

abordagem como complemento da GO (secção 2.1.2).

Do ponto de vista de classificação, os predictores iLoc-Cell [2, 28, 29, 41, 48, 51] utilizam o classificador AL-KNN (Accumulation-label K-nearest neighbor)[1] para prever a localização subcelular das proteínas (Tabela 2.2).

Posto isto, o processo de previsão através dos predictores iLoc-Cell [2] pode ser dividido em duas situações (Figura 2.2):

1. Se a proteína em estudo puder ser representada através da abordagem GO (secção 2.1.2), então $P_{GO}$ será utilizado para determinar as localizações subcelulares das proteínas.

2. Se a proteína em estudo não tiver homologias significativas com nenhuma proteína na base de dados Swiss-Prot ou se não tiver nenhum termo GO a ela associada, então $P_{SeqEvo}$ será utilizado para a previsão (secção 2.1.1.5).

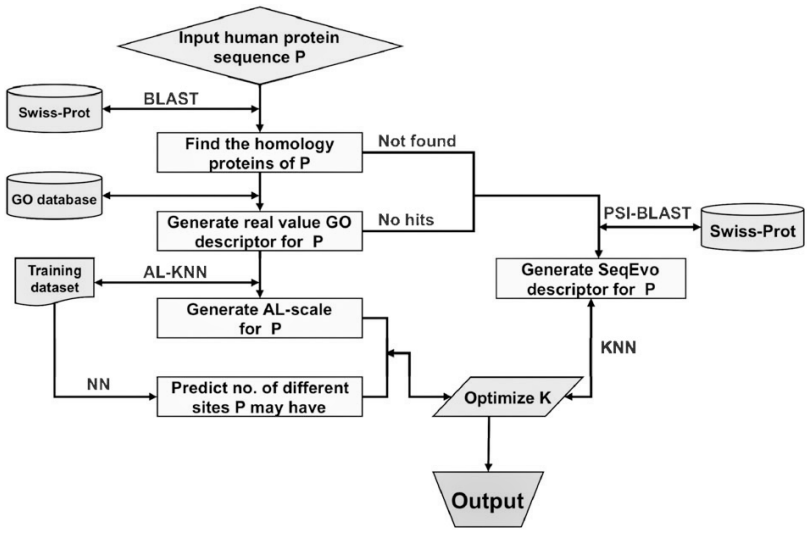


Figura 2.2: Representação do processo de previsão dos predictores iLoc-Cell (iLoc-Hum, iLoc-Euk, iLoc-Plant, iLoc-Gpos e iLoc-Gneg) (Figura extraída de Chou et al. [2]).

### 2.6.3 PolyU-Loc

PolyU-Loc é um conjunto de sete predictores capazes de prever a localização de proteínas em diferentes organismos (humanos, eucariotas, plantas, vírus): GOASVM[2] [38, 39],

[1] formulação matemática detalhada em [2]
[2] http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/GOASVM.html

mGOASVM[1] [42], HybridGO-Loc[2] [43], R3P-Loc[3] [12], mPLR-Loc[4] [11], **mLASSO**[5] [3, 10] e **mEN**[6] [10]. Com excepção do primeiro, todos conseguem prever a localização de proteínas que estão localizadas em, ou que se movem entre, duas ou mais regiões subcelulares (Tabela 2.5).

Tabela 2.5: PolyU: predictores para previsão da localização subcelular de proteínas em seis organismos/espécies

| Predictor | Organismo/Espécie | Localização Múltipla | Número de Localizações | Referências |
|---|---|---|---|---|
| GOASVM | Humanos<br>Eucariotas | Não<br>Não | 12<br>16 | [38, 39] |
| mGOASVM | Plantas<br>Vírus | Sim<br>Sim | 12<br>6 | [42] |
| HybridGO-Loc | Plantas<br>Vírus | Sim<br>Sim | 12<br>6 | [43] |
| R3P-Loc | Eucariotas<br>Plantas | Sim<br>Sim | 22<br>12 | [12] |
| mPLR-Loc | Eucariotas<br>Plantas | Sim<br>Sim | 22<br>12 | [11] |
| **mLASSO** | **Humanos** | **Sim** | **14** | [3, 10] |
| **mEN** | **Humanos** | **Sim** | **14** | [10] |

Muitos predictores *multi-label*, como o Cell-PLoc 2.0 (secção 2.6.1) [36], iLoc-Cell (secção 2.6.2) [2, 28, 29, 41, 48, 51], mGOASVM [42], HybridGO-Loc [43], R3P-Loc [12] e mPLR-Loc [11], usam a informação GO como representação das proteínas e aplicam vários classificadores diferentes para a previsão. No entanto, de acordo com Wan et al. [10], devido à elevada dimensão dos vectores, esses predictores têm duas grandes desvantagens. A primeira é a **falta de interpretabilidade**, ou seja, os predictores dão informações sobre a localização subcelular das proteinas, mas não fornecem razões biológicas para tal. Este é, possivelmente, um problema comum para a maioria das abordagens baseadas na aprendizagem automática, pois, geralmente, é difícil correlacionar características estatísticas de dados biológicas com fenómenos biológicos. A segunda é a **susceptibilidade a *overfitting***, isto é, o número de termos GO extraídos a partir

[1] http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/mGOASVM.html
[2] http://bioinfo.eie.polyu.edu.hk/HybridGoServer/
[3] http://bioinfo.eie.polyu.edu.hk/R3PLocServer/
[4] http://bioinfo.eie.polyu.edu.hk/mPLRLocServer/
[5] http://bioinfo.eie.polyu.edu.hk/SpaPredictorServer/
[6] http://bioinfo.eie.polyu.edu.hk/SpaPredictorServer/

de bases de dados de conhecimento (e.g. GOA) é consideravelmente maior do que o número de proteínas de interesse. A maioria dos predictores constroem vectores de termos GO de elevada dimensão, onde é provável que hajam muitos que sejam irrelevantes ou redundantes [10].

Para contornar a falta de interpretabilidade e a susceptibilidade a *overfitting*, foram desenvolvidos dois servidores web, mLASSO [3, 10] e mEN [10], os quais são capazes de prever e interpretar a localização subcelular de proteínas humanas localizadas em, ou que se movem entre, duas ou mais regiões subcelulares.

Como referido anteriormente, para uma dada proteína, o predictor deve conseguir lidar com dois tipos de situações: (1) o número de acesso da proteína é conhecido e (2) apenas a sequência proteica é conhecida. Para o primeiro caso, os respectivos termos GO são extraídos directamente da base de dados GOA usando os números de acesso como chave para a pesquisa. Para o segundo caso, o BLAST [127] é utilizado para encontrar os números de acesso das proteínas homólogas, os quais serão utilizados para extrair os termos GO [12]. Enquanto que a base de dados GOA permite associar o número de acesso da proteína a um conjunto de termos GO, para algumas proteínas (ou mesmo para as suas homólogas) não é possível associar qualquer termo GO. Nesses casos, a maioria dos predictores utiliza métodos alternativos. No entanto, essas estratégias podem levar a um mau desempenho do predictor e a um aumento da complexidade computacional e de armazenamento [12].

Para evitar estes problemas, Wan et al. [12] criaram duas bases de dados pequenas e eficientes: **ProSeq** (base de dados de sequências) e **ProSeq-GO** (base de dados de termos GO). Para isso, todos os números de acesso da base de dados Swiss-Prot e os números de acesso válidos (os que têm pelo menos um termo GO anotado a ele) da base de dados GOA foram extraídos. Em seguida, foram seleccionados os números de acesso em comum - "Valid Swiss-Prot ACs" (ou seja, cada um deles corresponde a pelo menos um termo GO na base de dados GOA). Estes, por sua vez, foram utilizados para extrair as sequências proteicas da base de dados Swiss-Prot e, assim, construir a nova base de dados de sequências (ProSeq). Da mesma forma, usando esses números de acesso, os termos GO foram extraídos da base de dados GOA para a nova base de dados (ProSeq-GO) (Figura 2.3) [12]. Assim, para uma dada proteína, os servidores web mLASSO e mEN podem recorrer às duas bases de dados para extrair e construir os vector de termos GO.

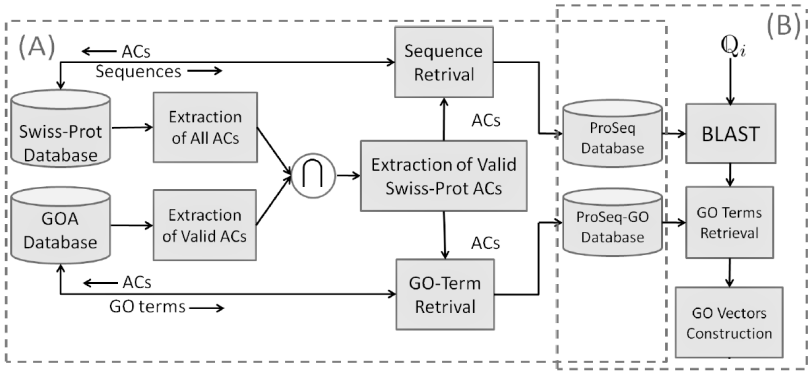


Figura 2.3: Representação dos (a) procedimentos para a criação das bases de dados ProSeq e ProSeq-GO e (b) contrução dos vectores de termos GO para os predictores mLASSO e mEN (Figura adaptada de Wan et al. [3]).

Através da estratégia *one-vs-rest*, mLASSO e mEN identificaram 87 e 429 (de mais de 8000) termos GO, respectivamente, que desempenham um papel essencial na determinação da localização subcelular. De salientar que a maioria dos termos GO seleccionados pelo mEN pertencem às categorias processo biológico e função molecular, o que sugere que os termos GO dessas categorias também desempenham um papel vital na previsão da localização subcelular, como referido anteriormente. Com estes termos GO essenciais, não ficamos a saber apenas onde a proteína está localizada mas, também, o motivo biológico pelo qual isso acontece [10]. Posto isto, os vectores representativos das proteínas serão **87-D (mLASSO)** e **429-D (mEN)**, onde cada elemento corresponde ao número de ocorrências de um dado termo GO. Posteriormente, estes serão classificados através dos classificadores LASSO (Least absolute shrinkage and selection operator)[1] e EN (Elastic net)[1], respectivamente (Tabela 2.2) [10].

Em suma, ambos os predictores mLASSO e mEN são interpretativos e têm um desempenho melhor do que os restantes predictores existentes. Porém, mEN selecciona mais termos GO relevantes do que mLASSO tendo, por isso, melhor taxa de sucesso global (Tabela 2.3) [10].

[1] formulação matemática detalhada em [10]

# Capítulo 3

# Previsão da Localização Subcelular das Proteínas

Como referido anteriormente, o objectivo primordial deste projecto foi prever a localização subcelular das proteínas codificadas por 800 genes humanos envolvidos no tráfego da CFTR. Uma vez que proteínas localizadas em determinados compartimentos intracelulares possuem características em comum, os algoritmos de aprendizagem automática podem ser úteis na previsão da localização subcelular de proteínas. Assim, durante o processo de aprendizagem, um conjunto de dados é utilizado para estimar os parâmetros de classificação e testar o classificador, permitindo a construção de um modelo de previsão. Este, por sua vez, será a base sobre a qual as previsões serão feitas num conjunto de dados com classificação desconhecida [145]. O processo de aprendizagem e previsão foi realizado através do software MEKA [20, 21], uma extensão do WEKA [20] para previsão *multi-label*. Os resultados obtidos pela aplicação dos modelos de previsão foram, então, comparados com as localizações subcelulares extraídas da base de dados UniProtKB/Swiss-Prot [122] e com as localizações subcelulares previstas pelos predictores mLASSO [10] e mEN [10].

## 3.1 Conjunto de dados

### 3.1.1 Conjunto de dados de treino

O conjunto de dados de treino utilizado neste estudo para avaliar o desempenho dos algoritmos foi construído por Shen and Chou [1] e utilizado em diferentes estudos de

previsão da localização subcelular de proteínas humanas [2, 3, 10]. Este conjunto de dados foi elaborado especialmente para proteínas humanas, cobrindo até 14 localizações subcelulares (centrossoma, citoplasma, citoesqueleto, retículo endoplasmático, endossoma, espaço extracelular, complexo de Golgi, lisossoma, microssoma, mitocôndria, núcleo, peroxissoma, membrana plasmática e sinapse). Para além disso, inclui proteínas com múltiplas localizações e nenhuma das proteínas tem mais de 25% de identidade de sequência com outra dentro do mesmo subconjunto (localização subcelular). O mesmo é constituído por 3681 proteínas locativas (3106 proteínas diferentes) distribuídas pelas 14 localizações subcelulares. Das 3106 proteínas, 2580 estão associadas a apenas uma localização subcelular, 480 a duas, 43 a três e 3 a quatro localizações subcelulares [1]. A Tabela 3.1 compara as diferentes versões do conjunto de dados utilizado (*dataset* 3).

Tabela 3.1: Comparação entre três conjuntos de dados utilizados para a prever a localização subcelular de proteínas humanas

| **Localizações** | **Dataset 1[a] Versão 49.3 (31-Mar-2006)** | **Dataset 2[b] Versão 50.7 (19-Sept-2006)** | **Dataset 3[c] Versão 55.3 (29-Apr-2008)** |
|---|---|---|---|
| **centrossoma** | 20 | 39 | 77 |
| **citoplasma** | 155 | 633 | 817 |
| **citoesqueleto** | 12 | 47 | 79 |
| **retículo endoplasmático** | 28 | 157 | 229 |
| **endossoma** | - | 48 | 24 |
| **espaço extracelular** | 140 | 325 | 385 |
| **complexo de Golgi** | 33 | 112 | 161 |
| **lisossoma** | 32 | 63 | 77 |
| **microssoma** | 7 | 15 | 24 |
| **mitocôndria** | 125 | 307 | 364 |
| **núcleo** | 196 | 877 | 1021 |
| **peroxissoma** | 18 | 46 | 47 |
| **membrana plasmática** | 153 | 455 | 354 |
| **sinápse** | - | 10 | 22 |
| **TOTAL** | | | |
| **Proteínas locativas** | **919** | **3134** | **3681** |
| **Proteínas diferentes** | **919[d]** | **2750[e]** | **3106[f]** |

[a] Utilizado pelo predictor Hum-PLoc [37];
[b] Utilizado pelo predictor Hum-mPLoc [9, 35];
[c] Utilizado pelos predictores Hum-mPLoc 2.0 [1, 36]; iLoc-Hum [2]; mLASSO-Hum [3, 10]; mLASSO e mEN [10];
[d] Todas as proteínas pertencem a uma e só uma localização;
[e] Das 2750 proteínas diferentes, 2396 pertencem apenas a uma localização, 325 a duas localizações, 28 a três localizações e 1 a quatro localizações;
[f] Das 3016 proteínas diferentes, 2580 pertencem apenas a uma localização, 480 a duas localizações, 43 a três localizações e 3 a quatro localizações.

### 3.1.2 Conjunto de dados para previsão

Para prever a localização subcelular das proteínas codificadas por 800 genes humanos, os Ensembl IDs dos mesmos foram convertidos em UniProtKB/Swiss-Prot IDs (proteínas). Posto isto, determinou-se que, desses 800 genes, 7 (ENSG00000144596, ENSG00000168970, ENSG00000234616, ENSG00000249209, ENSG00000249624, ENSG00000261408, ENSG00000266028) não codificam nenhuma proteína anotada na base de dados UniProtKB/Swiss-Prot. Por outro lado, 790 genes codificam apenas uma proteína, enquanto que os genes ENSG00000020256, ENSG00000206503 e ENSG00000087460, codificam, respectivamente, duas, três e quatro proteínas anotadas na base de dados UniProtKB/Swiss-Prot. Assim, apenas foi possível prever a localização de 799 proteínas (codificadas por 793 genes).

### 3.1.3 Construção dos vectores representativos das proteínas

O processo de construção dos vectores representativos das proteínas envolve duas etapas: (1) extracção dos termos GO e (2) construção dos vectores de termos GO.

Na primeira etapa, os termos GO das proteínas em estudo foram extraídos da base de dados QuickGO[1] [146]. No entanto, algumas proteínas não têm termos GO associados e, por isso, o vector representativo das mesmas é nulo. Como tal, para estes casos, as sequências proteicas das mesmas foram utilizadas para extrair, através do BLAST (com e-value = 0.001), os números de acesso das suas proteínas homólogas. Estes, por sua vez, foram sucessivamente utilizados, por ordem decrescente de resultado, até que pelo menos um termo GO fosse encontrado.

Da perspectiva de construção dos vectores de termos GO há dois factores a ter em conta: (1) dimensão do vector e (2) elementos do vector. Para explorar várias possibilidades, foram utilizadas diferentes abordagens de construção dos vectores. Assim, foram criados seis conjuntos de dados de treino diferentes:

- **Conjunto de dados de treino T-A1 (10165-D, 0-1)**: vectores com dimensão 10165-D (todos os termos GO distintos para as proteínas em estudo), onde cada elemento corresponde a 0 ou 1 (Valor 1-0);

[1] https://www.ebi.ac.uk/QuickGO/

- **Conjunto de dados de treino T-A2 (10165-D, TF)**: vectores com dimensão 10165-D (todos os termos GO distintos para as proteínas em estudo), onde cada elemento corresponde à frequência do termo GO (TF);
- **Conjunto de dados de treino T-B1 (429-D, 0-1)**: vectores com dimensão 429-D (termos GO essenciais obtidos por Wan et al. [10] através do classificador mEN), onde cada elemento corresponde a 0 ou 1 (Valor 1-0);
- **Conjunto de dados de treino T-B2 (429-D, TF)**: vectores com dimensão 429-D (termos GO essenciais obtidos por Wan et al. [10] através do classificador mEN), onde cada elemento corresponde à frequência do termo GO (TF);
- **Conjunto de dados de treino T-C1 (87-D, 0-1)**: vectores com dimensão 87-D (termos GO essenciais obtidos por Wan et al. [10] através do classificador mLASSO), onde cada elemento corresponde a 0 ou 1 (Valor 1-0);
- **Conjunto de dados de treino T-C2 (87-D, TF)**: vectores com dimensão 87-D (termos GO essenciais obtidos por Wan et al. [10] através do classificador mLASSO), onde cada elemento corresponde à frequência do termo GO (TF).

A mesma abordagem foi aplicada aos conjuntos de dados para previsão, originando seis conjuntos de dados diferentes: P-A1 (10165-D, 1-0), P-A2 (10165-D, TF), P-B1 (429-D, 1-0), P-B2 (429-D, TF), P-C1 (87-D, 1-0) e P-C2 (87-D, TF).

## 3.2 Processo de Aprendizagem e Previsão

O processo de aprendizagem para previsão da localização subcelular de proteínas foi realizado através do software MEKA [20, 21]. Este software foi criado para aplicar e avaliar a classificação *multi-label* através de diferentes métodos de transformação do problema (secção 2.3), os quais, por sua vez, utilizam métodos *single-label* como classificadores base [21].

Posto isto, os modelos de previsão foram gerados pelos algoritmos BR, LC e CC, juntamente com os classificadores base *single-label* SMO, PART e J48 (secção 2.3). Assim, cada um dos seis conjuntos de dados de treino criados anteriormente (T-A1, T-A2, T-B1, T-B2, T-C1 e T-C2) foi submetido no software MEKA [20, 21] e classificado pelos conjuntos de classificadores BR-SMO, BR-PART, BR-J48, CC-SMO, CC-PART,

CC-J48, LC-SMO, LC-PART e LC-J48, que foram avaliados através do método *10-fold cross-validation* (secção 2.4).

Posteriormente, os conjuntos de dados e classificadores com melhor desempenho foram submetidos novamente no software MEKA [20, 21] e avaliados através do método LOOCV (secção 2.4). De salientar que este método é computacionalmente mais exigente que o *10-fold cross-validation*, no entanto, é mais eficiente, sendo, por isso, utilizado pelos principais predictores de localização subcelular de proteínas humanas (secção 2.6). Em seguida, os modelos de previsão gerados foram aplicados aos respectivos conjuntos de dados para previsão das localizações subcelulares.

Complementarmente, as localizações subcelulares das 799 proteínas em estudo foram, também, previstas pelos predictores mLASSO e mEN [10][1] (secção 2.6.3).

Por último, os resultados obtidos nos processos anteriores foram comparados com as localizações subcelulares presentes na base de dados UniProtKB/Swiss-Prot [122] para cada proteína, as quais foram extraídas seguindo dois critérios: (1) extracção de todas as localizações subcelulares associadas a cada proteína e (2) extração das localizações subcelulares associadas a cada proteína e que foram determinadas experimentalmente. Como referido anteriormente, apenas foram consideradas 14 localizações subcelulares: *Centrosome [SL-0048], Cytoplasm [SL-0086], Cytoskeleton [SL-0090], Endoplasmic reticulum [SL-0095], Endosome [SL-0101], Extracellular space [SL-0112], Golgi apparatus [SL-0132], Lysosome [SL-0158], Microsome [SL-0166], Mitochondrion [SL-0173], Nucleus [SL-0191], Peroxisome [SL-0204], Cell membrane [SL-0039] e Synapse [SL-0258].*

## 3.3 Resultados

### 3.3.1 Avaliação dos classificadores através do método *10-fold cross validation*

As Tabelas 3.2 e 3.3 apresentam os resultados da aplicação dos diferentes classificadores aos seis conjuntos de dados de treino. Em termos gerais, é possível observar que a abordagem que recorre ao número de vezes que um termo GO surge anotado a uma dada proteína (TF) apresenta uma taxa de sucesso global média superior (OAA = 0,686, OLA = 0,785) à abordagem que utiliza apenas a presença ou ausência de um termo GO (1-0) (OAA = 0,632, OLA = 0,756). Admitindo que uma proteína possa ser

[1] http://bioinfo.eie.polyu.edu.hk/SpaPredictorServer/

anotada por diferentes grupos de investigação a diferentes ou até contraditórios termos GO, existe a possibilidade de ocorrer inconsistências nas anotações, o que pode afectar negativamente o desempenho dos métodos baseados em aprendizagem automática [7]. Por outro lado, o mesmo termo GO pode aparecer anotado mais do que uma vez à mesma proteína, sendo que cada anotação está associada a uma evidência ou base de dados diferente. Isto significa que quanto maior for a frequência que um termo GO é usado para anotar uma dada proteína, mais vezes a anotação foi confirmada por diferentes grupos de investigação e mais credível será essa anotação [7]. Assim, a utilização da abordagem TF realça os termos GO anotados com mais frequência. De salientar que as restantes métricas de desempenho (*accuracy*, *precision*, *recall*, *F1-score*, *hamming loss*, *under-prediction*, *equal-prediction* e *over-prediction*) também apresentaram melhores resultados médios quando utilizada a abordagem TF.

Por sua vez, de acordo com as métricas de desempenho OAA, *accuracy*, *precision*, *F1-score*, *under-*, *equal-* e *over-prediction*, os métodos LC e CC são, em média, superiores ao método BR. Em concreto, os métodos LC, CC e BR atingiram uma OAA média de 0,673, 0,662 e 0,642, respectivamente. Por outro lado, o método BR obteve uma OLA média superior (0,786) relativamente aos métodos LC (0,745) e CC (0,780), o que pode ser justificado por uma *over-prediction* relativamente aos restantes métodos. Para além disso, é importante salientar que, como referido anteriormente, ao contrário do método BR, os métodos LC e CC consideram a relação entre *labels* (localizações subcelulares). Quanto aos classificadores base *single-label*, o SMO obteve, em média, um desempenho superior (OAA = 0,668) do que os classificadores J48 (OAA = 0,659) e PART (OAA = 0,650).

Por último, a utilização dos termos GO mais relevantes, seleccionados pelos predictores mLASSO e mEN [10], apresentaram um desempenho médio superior quando comparado com a abordagem que extraiu todos os termos GO associados a cada proteína. Por outro lado, os conjuntos de dados T-C2 (OAA = 0,691, OLA = 0,786) e T-B2 (OAA = 0,691, OLA = 0,791) apresentaram, em média, melhores resultados que os restantes. Desta forma, o conjunto de dados de treino T-B2, quando classificado pelos algoritmos CC-SMO, CC-J48 e LC-SMO, e conjunto de dados T-C2, quando classificado pelos algoritmos CC-SMO, CC-J48, LC-SMO e LC-J48, apresentaram uma OAA superior a 69,2%.

Tabela 3.2: Resultados da aplicação dos classificadores BR (SMO, PART e J48), CC (SMO, PART e J48) e LC (SMO, PART e J48) aos conjuntos de dados (A) T-A1 (10165-D, 1-0), (B) T-B1 (429-D, 1-0) e (C) T-C1 (87-D, 1-0) para previsão das localizações subcelulares de proteínas humanas. O método usado para avaliar os classificadores foi o *10-fold cross-validation*.
*OAA: overall actual accuracy; OLA: overall locative accuracy; F1: F1-score; HL: hamming loss; UP: under-prediction; EP: equal-prediction; OP: over-prediciton.*

**(A)** **T-A1 (10165-D, 1-0)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,618 | 0,577 | 0,588 | 0,630 | 0,603 | 0,602 | 0,683 | 0,621 | 0,639 |
| **OLA** | 0,781 | 0,747 | 0,807 | 0,777 | 0,733 | 0,780 | 0,751 | 0,705 | 0,697 |
| **Accuracy** | 0,734 | 0,690 | 0,721 | 0,739 | 0,705 | 0,720 | 0,763 | 0,706 | 0,717 |
| **Precision** | 0,771 | 0,724 | 0,749 | 0,777 | 0,750 | 0,754 | 0,818 | 0,756 | 0,775 |
| **Recall** | 0,815 | 0,772 | 0,829 | 0,809 | 0,763 | 0,805 | 0,793 | 0,743 | 0,739 |
| **F1** | 0,773 | 0,729 | 0,766 | 0,776 | 0,739 | 0,760 | 0,791 | 0,735 | 0,743 |
| **HL** | 0,039 | 0,044 | 0,042 | 0,040 | 0,043 | 0,042 | 0,037 | 0,048 | 0,045 |
| **UP** | 0,171 | 0,177 | 0,228 | 0,151 | 0,126 | 0,186 | 0,064 | 0,098 | 0,056 |
| **EP** | 0,703 | 0,659 | 0,657 | 0,738 | 0,719 | 0,700 | 0,812 | 0,780 | 0,808 |
| **OP** | 0,127 | 0,164 | 0,115 | 0,110 | 0,155 | 0,115 | 0,124 | 0,122 | 0,136 |

**(B)** **T-B1 (429-D, 1-0)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,627 | 0,606 | 0,620 | 0,643 | 0,627 | 0,645 | 0,677 | 0,631 | 0,650 |
| **OLA** | 0,797 | 0,758 | 0,783 | 0,787 | 0,746 | 0,755 | 0,751 | 0,724 | 0,725 |
| **Accuracy** | 0,745 | 0,715 | 0,733 | 0,750 | 0,724 | 0,737 | 0,760 | 0,721 | 0,729 |
| **Precision** | 0,778 | 0,753 | 0,769 | 0,790 | 0,770 | 0,782 | 0,816 | 0,773 | 0,777 |
| **Recall** | 0,829 | 0,791 | 0,813 | 0,819 | 0,778 | 0,786 | 0,791 | 0,762 | 0,763 |
| **F1** | 0,784 | 0,753 | 0,771 | 0,787 | 0,758 | 0,768 | 0,789 | 0,753 | 0,757 |
| **HL** | 0,038 | 0,041 | 0,039 | 0,038 | 0,041 | 0,039 | 0,037 | 0,044 | 0,043 |
| **UP** | 0,175 | 0,159 | 0,167 | 0,145 | 0,116 | 0,106 | 0,066 | 0,093 | 0,077 |
| **EP** | 0,704 | 0,682 | 0,694 | 0,750 | 0,745 | 0,759 | 0,811 | 0,788 | 0,808 |
| **OP** | 0,121 | 0,159 | 0,139 | 0,106 | 0,138 | 0,135 | 0,124 | 0,119 | 0,116 |

**(C)** **T-C1 (87-D, 1-0)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,631 | 0,615 | 0,632 | 0,653 | 0,636 | 0,652 | 0,679 | 0,637 | 0,642 |
| **OLA** | 0,794 | 0,753 | 0,780 | 0,786 | 0,747 | 0,757 | 0,744 | 0,724 | 0,729 |
| **Accuracy** | 0,744 | 0,717 | 0,739 | 0,757 | 0,731 | 0,744 | 0,760 | 0,725 | 0,729 |
| **Precision** | 0,777 | 0,755 | 0,777 | 0,799 | 0,778 | 0,793 | 0,818 | 0,779 | 0,783 |
| **Recall** | 0,827 | 0,785 | 0,811 | 0,820 | 0,781 | 0,793 | 0,785 | 0,762 | 0,767 |
| **F1** | 0,783 | 0,752 | 0,776 | 0,792 | 0,764 | 0,777 | 0,787 | 0,755 | 0,759 |
| **HL** | 0,037 | 0,041 | 0,038 | 0,037 | 0,041 | 0,039 | 0,037 | 0,044 | 0,043 |
| **UP** | 0,169 | 0,146 | 0,151 | 0,131 | 0,111 | 0,101 | 0,057 | 0,089 | 0,083 |
| **EP** | 0,705 | 0,707 | 0,719 | 0,765 | 0,762 | 0,776 | 0,813 | 0,789 | 0,796 |
| **OP** | 0,126 | 0,147 | 0,130 | 0,104 | 0,127 | 0,122 | 0,130 | 0,121 | 0,121 |

Tabela 3.3: Resultados da aplicação dos classificadores BR (SMO, PART e J48), CC (SMO, PART e J48) e LC (SMO, PART e J48) aos conjuntos de dados (A) T-A2 (10165-D, TF), (B) T-B2 (429-D, TF) e (C) T-C2 (87-D, TF) para previsão das localizações subcelulares de proteínas humanas. O método usado para avaliar os classificadores foi o *10-fold cross-validation*.
*OAA: overall actual accuracy; OLA: overall locative accuracy; F1: F1-score; HL: hamming loss; UP: under-prediction; EP: equal-prediction; OP: over-prediciton.*

**(A)** **T-A2 (10165-D, TF)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,629 | 0,678 | 0,681 | 0,659 | 0,685 | 0,693 | 0,683 | 0,682 | 0,690 |
| **OLA** | 0,738 | 0,806 | 0,825 | 0,751 | 0,797 | 0,813 | 0,733 | 0,752 | 0,776 |
| **Accuracy** | 0,714 | 0,771 | 0,779 | 0,742 | 0,777 | 0,787 | 0,753 | 0,755 | 0,768 |
| **Precision** | 0,750 | 0,803 | 0,809 | 0,786 | 0,819 | 0,827 | 0,809 | 0,799 | 0,809 |
| **Recall** | 0,765 | 0,834 | 0,848 | 0,782 | 0,829 | 0,842 | 0,770 | 0,789 | 0,808 |
| **F1** | 0,743 | 0,803 | 0,812 | 0,770 | 0,809 | 0,819 | 0,777 | 0,781 | 0,795 |
| **HL** | 0,038 | 0,032 | 0,031 | 0,040 | 0,033 | 0,032 | 0,039 | 0,040 | 0,037 |
| **UP** | 0,110 | 0,131 | 0,143 | 0,091 | 0,109 | 0,117 | 0,043 | 0,080 | 0,087 |
| **EP** | 0,704 | 0,741 | 0,740 | 0,796 | 0,774 | 0,772 | 0,831 | 0,818 | 0,820 |
| **OP** | 0,185 | 0,128 | 0,117 | 0,113 | 0,117 | 0,111 | 0,126 | 0,102 | 0,093 |

**(B)** **T-B2 (429-D, TF)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,684 | 0,682 | 0,679 | 0,712 | 0,685 | 0,699 | 0,717 | 0,683 | 0,682 |
| **OLA** | 0,774 | 0,812 | 0,826 | 0,790 | 0,809 | 0,821 | 0,760 | 0,758 | 0,772 |
| **Accuracy** | 0,760 | 0,778 | 0,781 | 0,787 | 0,783 | 0,792 | 0,786 | 0,763 | 0,766 |
| **Precision** | 0,795 | 0,813 | 0,813 | 0,829 | 0,825 | 0,832 | 0,844 | 0,810 | 0,812 |
| **Recall** | 0,802 | 0,841 | 0,852 | 0,821 | 0,839 | 0,847 | 0,801 | 0,799 | 0,806 |
| **F1** | 0,786 | 0,811 | 0,815 | 0,812 | 0,816 | 0,824 | 0,810 | 0,791 | 0,795 |
| **HL** | 0,031 | 0,032 | 0,032 | 0,033 | 0,033 | 0,031 | 0,033 | 0,039 | 0,037 |
| **UP** | 0,088 | 0,131 | 0,148 | 0,072 | 0,117 | 0,115 | 0,032 | 0,084 | 0,090 |
| **EP** | 0,742 | 0,743 | 0,733 | 0,830 | 0,771 | 0,778 | 0,840 | 0,805 | 0,807 |
| **OP** | 0,170 | 0,126 | 0,119 | 0,097 | 0,112 | 0,107 | 0,128 | 0,111 | 0,103 |

**(C)** **T-C2 (87-D, TF)**

| | BR | | | CC | | | LC | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** | **SMO** | **PART** | **J48** |
| **OAA** | 0,665 | 0,676 | 0,673 | 0,712 | 0,690 | 0,692 | 0,722 | 0,691 | 0,702 |
| **OLA** | 0,747 | 0,802 | 0,825 | 0,784 | 0,801 | 0,810 | 0,758 | 0,763 | 0,787 |
| **Accuracy** | 0,739 | 0,769 | 0,776 | 0,786 | 0,780 | 0,785 | 0,790 | 0,770 | 0,783 |
| **Precision** | 0,775 | 0,801 | 0,809 | 0,832 | 0,823 | 0,824 | 0,849 | 0,821 | 0,827 |
| **Recall** | 0,776 | 0,833 | 0,849 | 0,817 | 0,830 | 0,840 | 0,802 | 0,800 | 0,823 |
| **F1** | 0,763 | 0,801 | 0,811 | 0,812 | 0,811 | 0,817 | 0,813 | 0,797 | 0,811 |
| **HL** | 0,032 | 0,034 | 0,033 | 0,033 | 0,033 | 0,033 | 0,032 | 0,037 | 0,035 |
| **UP** | 0,077 | 0,132 | 0,148 | 0,066 | 0,108 | 0,116 | 0,026 | 0,072 | 0,084 |
| **EP** | 0,722 | 0,746 | 0,737 | 0,832 | 0,778 | 0,779 | 0,840 | 0,817 | 0,818 |
| **OP** | 0,200 | 0,122 | 0,115 | 0,102 | 0,114 | 0,106 | 0,134 | 0,111 | 0,098 |

### 3.3.2 Avaliação dos classificadores através do método *Leave-one-out cross-validation*

Os conjuntos de dados e classificadores com melhor desempenho foram submetidos novamente no software MEKA [20, 21] e re-avaliados através do método LOOCV. Assim, os classificadores CC-SMO, CC-J48 e LC-SMO foram aplicados ao conjunto de dados de treino T-B2 (429-D, TF) e os classificadores CC-SMO, CC-J48, LC-SMO e LC-J48 ao conjunto de dados de treino T-C2 (87-D, TF).

Os modelos de previsão gerados pelo MEKA [20, 21] atingiram uma taxa global de sucesso entre 69,2 e 72,3% (OAA) ou entre 76,1 e 80,3% (OLA). Comparativamente com os principais predictores existentes para previsão da localização subcelular de proteínas humanas (secção 2.6), os sete modelos de previsão gerados pelo MEKA [20, 21] apresentaram um desempenho superior aos predictores Hum-mPLoc 2.0[1] (OLA = 0,627) [1] e iLoc-Hum (OAA = 0,682 e OLA = 0,763) [2] e inferior aos predictores mLASSO (OAA = 0,729 e OLA = 0,820) [10] e mEN (OAA = 0,743 e OLA = 0,836) [10]. Em particular, o classificador LC-SMO obteve uma OAA igual a 0,723, quando aplicado ao conjunto de dados TC-2, e 0,720, quando aplicado ao conjunto de dados TB-2, ou seja, apenas 1 e 2,5% inferior aos predictores mLASSO e mEN [10], respectivamente. Por outro lado, o classificador CC-J48 obteve uma OLA apenas 1,7 e 3,3% inferior aos predictores mLASSO e mEN [10], respectivamente.

Como se pode constatar pela Tabela 3.4, o mEN [10] apresentou, relativamente aos restantes predictores, um desempenho superior na previsão das localizações centrossoma (0,779) e lisossoma (0,961). Por sua vez, o predictor mLASSO [10] teve um resultado superior na previsão das localizações citoplasma (0,856) e retículo endoplasmático (0,847), enquanto que o predictor T-C2/LC-J48 nas localizações endossoma (0,583), espaço extracelular (0,878) e membrana plasmática (0,797). Tanto o mEN como o mLASSO [10] previram correctamente a localização de 92,3% das proteínas locativas localizadas na mitocôndria e 90,4% das proteínas locativas localizadas no núcleo; e tanto o mEN [10] como o T-C2/LC-J48 59,1% das proteínas locativas localizadas na sinapse. Por outro lado, o predictor T-B2/CC-J48 previu correctamente 80,7 e 87,2% das proteínas locativas localizadas no complexo de Golgi e no peroxissoma, assim como, juntamente com o predictor T-C2/CC-J48, 75% das proteínas locativas localizadas no microssoma.

[1] OLA não disponível

Tabela 3.4: Resultados da aplicação dos classificadores CC-SMO, CC-J48 e LC-SMO ao conjunto de dados T-B2 e dos classificadores CC-SMO, CC-J48, LC-SMO e LC-J48 ao conjunto de dados T-C2 e comparação com os predictores mLASSO e mEN. O método usado para avaliar os classificadores foi o *Leave-one-out cross-validation*.
*1. centrossoma; 2. citoplasma; 3. citoesqueleto; 4. retículo endoplasmático; 5. endossoma; 6. espaço extracelular; 7. complexo de Golgi; 8. lisossoma; 9. microssoma; 10. mitocôndria; 11. núcleo; 12. peroxissoma; 13. membrana plasmática e 14. sinapse; OAA: overall actual accuracy; OLA: overall locative accuracy; F1: F1-score; HL: hamming loss; UP: under-prediction; EP: equal-prediction; OP: over-prediciton.*

| | T-B2 | | | T-C2 | | | | mLASSO | mEN |
|---|---|---|---|---|---|---|---|---|---|
| | **CC-SMO** | **CC-J48** | **LC-SMO** | **CC-SMO** | **CC-J48** | **LC-SMO** | **LC-J48** | | |
| **1** | $\frac{47}{77} = 0{,}610$ | $\frac{53}{77} = 0{,}688$ | $\frac{48}{77} = 0{,}623$ | $\frac{45}{77} = 0{,}584$ | $\frac{55}{77} = 0{,}714$ | $\frac{52}{77} = 0{,}675$ | $\frac{51}{77} = 0{,}662$ | $\frac{42}{77} = 0{,}545$ | $\frac{60}{77} = 0{,}779$ |
| **2** | $\frac{674}{817} = 0{,}825$ | $\frac{609}{817} = 0{,}745$ | $\frac{604}{817} = 0{,}739$ | $\frac{667}{817} = 0{,}816$ | $\frac{627}{817} = 0{,}767$ | $\frac{591}{817} = 0{,}723$ | $\frac{592}{817} = 0{,}725$ | $\frac{699}{817} = 0{,}856$ | $\frac{683}{817} = 0{,}836$ |
| **3** | $\frac{34}{79} = 0{,}430$ | $\frac{35}{79} = 0{,}443$ | $\frac{28}{79} = 0{,}354$ | $\frac{29}{79} = 0{,}367$ | $\frac{31}{79} = 0{,}392$ | $\frac{42}{79} = 0{,}532$ | $\frac{31}{79} = 0{,}392$ | $\frac{29}{79} = 0{,}367$ | $\frac{32}{79} = 0{,}405$ |
| **4** | $\frac{176}{229} = 0{,}769$ | $\frac{189}{229} = 0{,}825$ | $\frac{175}{229} = 0{,}764$ | $\frac{174}{229} = 0{,}760$ | $\frac{182}{229} = 0{,}795$ | $\frac{182}{229} = 0{,}795$ | $\frac{173}{229} = 0{,}755$ | $\frac{194}{229} = 0{,}847$ | $\frac{190}{229} = 0{,}830$ |
| **5** | $\frac{2}{24} = 0{,}083$ | $\frac{8}{24} = 0{,}333$ | $\frac{0}{24} = 0{,}000$ | $\frac{0}{24} = 0{,}000$ | $\frac{9}{24} = 0{,}375$ | $\frac{0}{24} = 0{,}000$ | $\frac{14}{24} = 0{,}583$ | $\frac{1}{24} = 0{,}042$ | $\frac{5}{24} = 0{,}208$ |
| **6** | $\frac{286}{385} = 0{,}743$ | $\frac{329}{385} = 0{,}855$ | $\frac{309}{385} = 0{,}803$ | $\frac{270}{385} = 0{,}701$ | $\frac{337}{385} = 0{,}875$ | $\frac{306}{385} = 0{,}795$ | $\frac{338}{385} = 0{,}878$ | $\frac{311}{385} = 0{,}808$ | $\frac{314}{385} = 0{,}816$ |
| **7** | $\frac{114}{161} = 0{,}708$ | $\frac{130}{161} = 0{,}807$ | $\frac{98}{161} = 0{,}609$ | $\frac{118}{161} = 0{,}733$ | $\frac{111}{161} = 0{,}689$ | $\frac{105}{161} = 0{,}652$ | $\frac{112}{161} = 0{,}696$ | $\frac{118}{161} = 0{,}733$ | $\frac{128}{161} = 0{,}795$ |
| **8** | $\frac{61}{77} = 0{,}792$ | $\frac{60}{77} = 0{,}779$ | $\frac{60}{77} = 0{,}779$ | $\frac{59}{77} = 0{,}766$ | $\frac{61}{77} = 0{,}792$ | $\frac{61}{77} = 0{,}792$ | $\frac{73}{77} = 0{,}948$ | $\frac{62}{77} = 0{,}805$ | $\frac{74}{77} = 0{,}961$ |
| **9** | $\frac{16}{24} = 0{,}667$ | $\frac{18}{24} = 0{,}750$ | $\frac{16}{24} = 0{,}667$ | $\frac{1}{24} = 0{,}042$ | $\frac{18}{24} = 0{,}750$ | $\frac{1}{24} = 0{,}042$ | $\frac{15}{24} = 0{,}625$ | $\frac{1}{24} = 0{,}042$ | $\frac{14}{24} = 0{,}583$ |
| **10** | $\frac{321}{364} = 0{,}882$ | $\frac{327}{364} = 0{,}898$ | $\frac{316}{364} = 0{,}868$ | $\frac{323}{364} = 0{,}887$ | $\frac{327}{364} = 0{,}898$ | $\frac{315}{364} = 0{,}865$ | $\frac{320}{364} = 0{,}879$ | $\frac{336}{364} = 0{,}923$ | $\frac{336}{364} = 0{,}923$ |
| **11** | $\frac{882}{1021} = 0{,}864$ | $\frac{877}{1021} = 0{,}859$ | $\frac{862}{1021} = 0{,}844$ | $\frac{899}{1021} = 0{,}881$ | $\frac{889}{1021} = 0{,}871$ | $\frac{847}{1021} = 0{,}830$ | $\frac{873}{1021} = 0{,}855$ | $\frac{922}{1021} = 0{,}903$ | $\frac{923}{1021} = 0{,}904$ |
| **12** | $\frac{39}{47} = 0{,}830$ | $\frac{41}{47} = 0{,}872$ | $\frac{39}{47} = 0{,}830$ | $\frac{39}{47} = 0{,}830$ | $\frac{39}{47} = 0{,}830$ | $\frac{39}{47} = 0{,}830$ | $\frac{38}{47} = 0{,}809$ | $\frac{34}{47} = 0{,}723$ | $\frac{39}{47} = 0{,}830$ |
| **13** | $\frac{257}{354} = 0{,}726$ | $\frac{273}{354} = 0{,}771$ | $\frac{245}{354} = 0{,}692$ | $\frac{266}{354} = 0{,}751$ | $\frac{266}{354} = 0{,}751$ | $\frac{250}{354} = 0{,}706$ | $\frac{282}{354} = 0{,}797$ | $\frac{267}{354} = 0{,}754$ | $\frac{266}{354} = 0{,}751$ |
| **14** | $\frac{9}{22} = 0{,}409$ | $\frac{7}{22} = 0{,}318$ | $\frac{8}{22} = 0{,}364$ | $\frac{6}{22} = 0{,}273$ | $\frac{5}{22} = 0{,}227$ | $\frac{9}{22} = 0{,}409$ | $\frac{13}{22} = 0{,}591$ | $\frac{3}{22} = 0{,}136$ | $\frac{13}{22} = 0{,}591$ |
| **OAA** | $\frac{2217}{3106} = 0{,}714$ | $\frac{2158}{3106} = 0{,}695$ | $\frac{2236}{3106} = 0{,}720$ | $\frac{2221}{3106} = 0{,}715$ | $\frac{2149}{3106} = 0{,}692$ | $\frac{2246}{3106} = 0{,}723$ | $\frac{2197}{3106} = 0{,}707$ | $\frac{2265}{3106} = 0{,}729$ | $\frac{2307}{3106} = 0{,}743$ |
| **OLA** | $\frac{2918}{3681} = 0{,}793$ | $\frac{2956}{3681} = 0{,}803$ | $\frac{2808}{3681} = 0{,}763$ | $\frac{2896}{3681} = 0{,}787$ | $\frac{2957}{3681} = 0{,}803$ | $\frac{2800}{3681} = 0{,}761$ | $\frac{2925}{3681} = 0{,}795$ | $\frac{3019}{3681} = 0{,}820$ | $\frac{3077}{3681} = 0{,}836$ |
| **Accuracy** | 0,789 | 0,783 | 0,788 | 0,789 | 0,782 | 0,790 | 0,789 | 0,814 | 0,827 |
| **Precision** | 0,832 | 0,825 | 0,844 | 0,834 | 0,822 | 0,847 | 0,833 | 0,859 | 0,869 |
| **Recall** | 0,823 | 0,833 | 0,804 | 0,819 | 0,834 | 0,803 | 0,831 | 0,857 | 0,870 |
| **F1** | 0,815 | 0,814 | 0,812 | 0,814 | 0,813 | 0,813 | 0,818 | 0,843 | 0,855 |
| **HL** | 0,033 | 0,033 | 0,033 | 0,033 | 0,034 | 0,032 | 0,034 | 0,029 | 0,028 |
| **UP** | 0,071 | 0,109 | 0,033 | 0,063 | 0,115 | 0,027 | 0,086 | - | - |
| **EP** | 0,833 | 0,779 | 0,841 | 0,835 | 0,777 | 0,844 | 0,820 | - | - |
| **OP** | 0,096 | 0,111 | 0,125 | 0,102 | 0,108 | 0,129 | 0,093 | - | - |

De salientar que a mitocôndia e o núcleo foram as localizações que, em média, foram mais fáceis de prever, com uma taxa de sucesso entre 87 e 92% e entre 83 e 90%, respectivamente, seguindo-se o lisossoma (77 a 96%), o peroxisoma (72 a 87%), o espaço extracelular (70 a 88%), o retículo endoplasmático (76 a 85%), o citoplasma (72 a 86%), a membrana plasmática (71 a 80%) e o complexo de Golgi (61 a 81%). Pelo contrário, o citoesqueleto (35 a 53%), o endossoma (0 a 58%) e a sinapse (14 a 59%) foram as localizações que apresentaram, em média, uma taxa de sucesso inferior a 50%. Um dos motivos para tal pode ser o facto do conjunto de dados de treino ser desproporcional quanto número de proteínas associadas a cada localização. Por exemplo, as localizações núcleo, citoplasma, expaço extracelular, mitocôndria e membrana plasmática estão associadas a 1021, 817, 385, 364 e 354 proteínas, respectivamente, entre as 3106 que pertencem ao conjunto de dados e treino. Por outro lado, as localizações citoesqueleto, endossoma, microssoma e sinapse estão associadas a apenas 79, 24, 24 e 22 proteínas, respectivamente, do conjunto de dados de treino, o que pode dificultar a tarefa de aprendizagem.

#### 3.3.2.1 Localização subcelular das proteínas do conjunto de dados para previsão

Após a criação dos modelos de previsão, os modelos T-B2/CC-SMO, T-B2/CC-J48 e T-B2/LC-SMO foram aplicados ao conjunto de dados para previsão P-B2 (429-D, TF) e os modelos T-C2/CC-SMO, T-C2/CC-J48, T-C2/LC-SMO e T-C2/LC-J48 ao conjunto de dados para previsão P-C2 (87-D, TF). A Tabela 3.5 apresenta o número de proteínas locativas associadas a cada localização, de acordo com a aplicação dos modelos de previsão gerados pelo MEKA [20, 21] e dos predictores mLASSO e mEN [10], assim como das localizações extraídas da base de dados UniProtKB/Swiss-Prot [122]. De relembrar que as localizações foram extraídas da base de dados UniProtKB/Swiss-Prot [122] seguindo dois critérios: (1) extracção de todas as localizações associadas a cada proteína e (2) extracção das localizações associadas a cada proteína e que foram determinadas experimentalmente. De forma geral, o citoplasma, o núcleo e a membrana celular foram as localizações subcelulares previstas para o maior número de proteínas. Pelo contrário, as localizações subcelulares citoesqueleto, o endossoma e o peroxissoma foram atribuídas a menos de 10 proteínas. Por outro lado, os predictores T-B2/CC-SMO, T-B2/LC-SMO, T-C2/CC-SMO e T-C2/LC-SMO não associaram nenhuma das

799 proteínas em estudo ao endossoma e os predictores T-C2/CC-SMO e T-C2/LC-SMO previram que nenhuma proteína está localizada no microssoma.

De acordo com as Tabelas 3.5 e 3.6, todos os predictores foram capazes de associar pelo menos uma localização subcelular a cada uma das 799 proteínas, excepto os predictores T-B2/CC-J48 e o T-C2/CC-J48, que apenas conseguiram prever a localização subcelular de 695 e 697 proteínas, respectivamente, entre as 799 em estudo. Por sua vez, recorrendo à base de dados UniProtKB/Swiss-Prot [122], apenas foi possível extrair a localização subcelular de 620 proteínas, isto se considerarmos todas as localizações subcelulares extraídas, ou 217 proteínas, se considerarmos apenas as localizações subcelulares determinadas experimentalmente. No entanto, a base de dados UniProtKB/Swiss-Prot [122] está em permanente actualização, o que sugere que, no futuro, mais informações sobre a localização subcelular das proteínas estarão disponíveis.

Tabela 3.5: Número de proteínas locativas associadas a cada localização subcelular, de acordo com os predictores T-B2/CC-SMO, T-B2/CC-J48 e T-B2/LC-SMO, quando aplicados ao conjunto de dados P-B2, os predictores T-C2/CC-SMO, T-C2/CC-J48, T-C2/LC-SMO e T-C2/LC-J48, quando aplicados ao conjunto de dados P-C2, os predictores mLASSO e mEN e a base de dados UniProtKB/Swiss-Prot. *1. centrossoma; 2. citoplasma; 3. citoesqueleto; 4. retículo endoplasmático; 5. endossoma; 6. espaço extracelular; 7. complexo de Golgi; 8. lissossoma; 9. microssoma; 10. mitocôndria; 11. núcleo; 12. peroxissoma; 13. membrana plasmática e 14. sinapse.*

| | **P-B2** | | | **P-C2** | | | | | | **Swiss-Prot** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CC** | | **LC** | **CC** | | **LC** | | **mLASSO** | **mEN** | | |
| | **SMO** | **J48** | **SMO** | **SMO** | **J48** | **SMO** | **J48** | | | **(A)** | **(B)** |
| **1** | 2 | 5 | 3 | 3 | 7 | 6 | 10 | 6 | 11 | 11 | 5 |
| **2** | 142 | 135 | 161 | 145 | 145 | 159 | 179 | 153 | 153 | 188 | 66 |
| **3** | 8 | 9 | 7 | 6 | 8 | 7 | 8 | 3 | 6 | 31 | 11 |
| **4** | 39 | 44 | 44 | 39 | 51 | 46 | 41 | 43 | 45 | 55 | 17 |
| **5** | 0 | 6 | 0 | 0 | 4 | 0 | 5 | 2 | 4 | 19 | 12 |
| **6** | 56 | 76 | 81 | 55 | 79 | 89 | 77 | 72 | 78 | 20 | 2 |
| **7** | 28 | 32 | 30 | 30 | 33 | 28 | 41 | 29 | 32 | 38 | 20 |
| **8** | 12 | 19 | 11 | 12 | 18 | 11 | 20 | 14 | 17 | 17 | 9 |
| **9** | 14 | 18 | 15 | 0 | 16 | 0 | 15 | 4 | 16 | 18 | 1 |
| **10** | 56 | 60 | 55 | 52 | 63 | 53 | 52 | 64 | 74 | 67 | 12 |
| **11** | 197 | 185 | 188 | 197 | 196 | 177 | 194 | 212 | 199 | 210 | 73 |
| **12** | 7 | 6 | 7 | 5 | 6 | 5 | 6 | 8 | 10 | 8 | 2 |
| **13** | 299 | 195 | 237 | 303 | 185 | 239 | 236 | 253 | 248 | 194 | 70 |
| **14** | 11 | 8 | 12 | 13 | 11 | 12 | 13 | 11 | 18 | 17 | 2 |
| **T Loc**[c] | 871 | 798 | 851 | 860 | 822 | 832 | 897 | 874 | 911 | 893 | 302 |
| **T Dif**[d] | 799 | 695 | 799 | 799 | 697 | 799 | 799 | 799 | 799 | 620 | 217 |

[a] Swiss-Prot (A): extracção de todas as localizações subcelulares associadas a cada proteína;
[b] Swiss-Prot (B): extracção das localizações associadas a cada proteína determinadas experimentalmente;
[c] Número total de proteínas locativas;
[d] Número total de proteínas diferentes.

Tabela 3.6: Número de proteínas que foram associadas a 0, 1, 2, ..., 14 localizações subcelulares de acordo com os diferentes predictores e base de dados UniProtKB/Swiss-Prot.

| | P-B2 | | | P-C2 | | | | | | Swiss-Prot | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CC | | LC | CC | | LC | | mLASSO | mEN | | |
| | SMO | J48 | SMO | SMO | J48 | SMO | J48 | | | (A) | (B) |
| **0** | 0 | 104 | 0 | 0 | 102 | 0 | 0 | 0 | 0 | 179 | 582 |
| **1** | 729 | 599 | 752 | 738 | 583 | 766 | 704 | 735 | 709 | 426 | 152 |
| **2** | 68 | 91 | 44 | 61 | 106 | 33 | 92 | 62 | 86 | 146 | 51 |
| **3** | 2 | 3 | 1 | 0 | 5 | 0 | 3 | 1 | 1 | 26 | 11 |
| **4** | 0 | 2 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 15 | 1 |
| **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 1 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **10** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **11** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| **13** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **TOTAL** | 799 | 695 | 799 | 799 | 697 | 799 | 799 | 799 | 799 | 620 | 217 |

A Tabela 3.6 indica que os predictores atribuíram à maioria das proteínas apenas uma localização subcelular, sendo que nenhum dos modelos de previsão gerados pelo MEKA [20, 21] atribuiu mais do que 4 localizações subcelulares a nenhuma das 799 proteínas em estudo. Na verdade, segundo os mesmos, no máximo 8 proteínas estão localizadas em mais do que duas localizações subcelulares. Por outro lado, o mEN associou 5 localizações subcelulares à proteína Q8TF05, 10 à proteína P22455 e, tal como o mLASSO, 12 à proteína P30154. Por sua vez, de acordo com as localizações subcelulares determinadas experimentalmente extraídas da base de dados UniProtKB/Swiss-Prot [122], 152 proteínas estão localizadas em apenas uma localização subcelular, enquanto que 65 proteínas estão localizadas em, ou movem-se entre, duas ou mais regiões subcelulares.

Por último, as Tabelas 3.7 e 3.8 fazem a comparação entre as localizações previstas pelos predictores e as localizações extraídas da base de dados UniProtKB/Swiss-Prot [122], com base em todas as localizações subcelulares extraídas ou apenas localizações determinadas experimentalmente, respectivamente. Desta forma, apresentam, para cada localização subcelular, a proporção de proteínas locativas determinadas pelos predictores que correspondem às extraídas da base de dados UniProtKB/Swiss-Prot [122],

de acordo com a fórmula:

$$\frac{\sum |Loc_{Predictor} \cap Loc_{Swiss-Prot}|}{\sum |Loc_{Swiss-Prot}|} \tag{3.1}$$

Posto isto, tendo em conta todas as localizações subcelulares extraídas da base de dados UniProtKB/Swiss-Prot [122], é possível verificar que todos os modelos de previsão gerados pelo MEKA [20, 21] foram capazes de prever correctamente a localização subcelular de pelo menos 70,3% (628 em 893) das proteínas locativas. Em particular, os predictores T-C2/CC-J48 e T-B2/CC-J48 atingiram uma taxa de sucesso global de 76,7% (685 em 893) e 75,1% (671 em 893), respectivamente, enquanto que o mEN previu correctamente 75,3% (672 em 893) das proteínas locativas e o mLASSO 72,2% (645 em 893) das proteínas locativas (Tabela 3.7).

Por outro lado, considerando apenas as localizações subcelulares extraídas da base de dados UniProtKB/Swiss-Prot [122] que foram determinadas experimentalmente, verificámos que todos os predictores foram capazes de prever correctamente a localização subcelular de pelo menos 67,9% (205 em 302) das proteínas locativas. Em concreto, os predictores T-B2/CC-J48, T-C2/CC-J48 e T-C2/LC-J48 atingiram uma taxa de sucesso global de 73,8% (223 em 302), 72,5% (219 em 302) e 71,5% (216 em 302), respectivamente, enquanto que o mEN previu correctamente 70,9% (214 em 302) das proteínas locativas e o mLASSO 69,9% (211 em 302) das proteínas locativas (Tabela 3.8).

É importante salientar que as bases de dados biológicas, incluindo a UniProtKB/Swiss-Prot [122], estão em constante actualização e, como tal, as informações biológicas, nomeadamente a localização subcelular das proteínas, estão, geralmente, susceptíveis a alterações. Desta forma, os falsos positivos, ou seja, quando o predictor prevê que uma dada proteína está localizada num determinado compartimento subcelular, mas essa associação não surge no conjunto de dados de treino ou na base de dados utilizada, podem, no futuro, ser considerados verdadeiros positivos, caso se verifique que a proteína está, de facto, localizada no tal compartimento subcelular. De forma semelhante, os verdadeiros negativos, isto é, quando nem o predictor nem o conjunto de dados de treino ou a base de dados utilizada associam uma localização subcelular a determinada proteína, podem, na verdade, ser falsos negativos, caso se verifique, no futuro, que essa proteína está localizada nesse compartimento subcelular.

Tabela 3.7: Comparação entre as localizações subcelulares previstas pelos classificadores T-B2/CC-SMO, T-B2/CC-J48 e T-B2/LC-SMO, quando aplicados ao conjunto de dados P-B2, pelos classificadores T-C2/CC-SMO, T-C2/CC-J48, T-C2/LC-SMO e T-C2/LC-J48, quando aplicados ao conjunto de dados P-C2, e pelos predictores mLASSO e mEN com todas as localizações subcelulares associadas a cada proteína extraídas da base de dados UniProtKB/Swiss-Prot.
*1. centrossoma; 2. citoplasma; 3. citoesqueleto; 4. retículo endoplasmático; 5. endossoma; 6. espaço extracelular; 7. complexo de Golgi; 8. lisossoma; 9. microssoma; 10. mitocôndria; 11. núcleo; 12. peroxissoma; 13. membrana plasmática e 14. sinapse.*

| | **P-B2** | | | **P-C2** | | | | **mLASSO** | **mEN** |
|---|---|---|---|---|---|---|---|---|---|
| | **CC-SMO** | **CC-J48** | **LC-SMO** | **CC-SMO** | **CC-J48** | **LC-SMO** | **LC-J48** | | |
| **1** | $\frac{1}{11} = 0{,}091$ | $\frac{5}{11} = 0{,}455$ | $\frac{2}{11} = 0{,}182$ | $\frac{3}{11} = 0{,}273$ | $\frac{6}{11} = 0{,}545$ | $\frac{6}{11} = 0{,}545$ | $\frac{7}{11} = 0{,}636$ | $\frac{4}{11} = 0{,}364$ | $\frac{7}{11} = 0{,}636$ |
| **2** | $\frac{132}{188} = 0{,}702$ | $\frac{124}{188} = 0{,}660$ | $\frac{126}{188} = 0{,}670$ | $\frac{136}{188} = 0{,}723$ | $\frac{133}{188} = 0{,}707$ | $\frac{129}{188} = 0{,}686$ | $\frac{121}{188} = 0{,}644$ | $\frac{130}{188} = 0{,}691$ | $\frac{125}{188} = 0{,}665$ |
| **3** | $\frac{8}{31} = 0{,}258$ | $\frac{9}{31} = 0{,}290$ | $\frac{7}{31} = 0{,}226$ | $\frac{6}{31} = 0{,}194$ | $\frac{8}{31} = 0{,}258$ | $\frac{7}{31} = 0{,}226$ | $\frac{7}{31} = 0{,}226$ | $\frac{2}{31} = 0{,}065$ | $\frac{4}{31} = 0{,}129$ |
| **4** | $\frac{37}{55} = 0{,}673$ | $\frac{40}{55} = 0{,}727$ | $\frac{39}{55} = 0{,}709$ | $\frac{38}{55} = 0{,}691$ | $\frac{47}{55} = 0{,}855$ | $\frac{43}{55} = 0{,}782$ | $\frac{37}{55} = 0{,}673$ | $\frac{40}{55} = 0{,}727$ | $\frac{39}{55} = 0{,}709$ |
| **5** | $\frac{0}{19} = 0{,}000$ | $\frac{5}{19} = 0{,}263$ | $\frac{0}{19} = 0{,}000$ | $\frac{0}{19} = 0{,}000$ | $\frac{4}{19} = 0{,}211$ | $\frac{0}{19} = 0{,}000$ | $\frac{4}{19} = 0{,}211$ | $\frac{1}{19} = 0{,}053$ | $\frac{2}{19} = 0{,}105$ |
| **6** | $\frac{18}{20} = 0{,}900$ | $\frac{19}{20} = 0{,}950$ | $\frac{17}{20} = 0{,}850$ | $\frac{20}{20} = 1{,}000$ | $\frac{20}{20} = 1{,}000$ | $\frac{18}{20} = 0{,}900$ | $\frac{19}{20} = 0{,}950$ | $\frac{20}{20} = 1{,}000$ | $\frac{19}{20} = 0{,}950$ |
| **7** | $\frac{27}{38} = 0{,}711$ | $\frac{31}{38} = 0{,}816$ | $\frac{28}{38} = 0{,}737$ | $\frac{28}{38} = 0{,}737$ | $\frac{28}{38} = 0{,}737$ | $\frac{26}{38} = 0{,}684$ | $\frac{29}{38} = 0{,}763$ | $\frac{26}{38} = 0{,}684$ | $\frac{30}{38} = 0{,}789$ |
| **8** | $\frac{12}{17} = 0{,}706$ | $\frac{15}{17} = 0{,}882$ | $\frac{11}{17} = 0{,}647$ | $\frac{12}{17} = 0{,}706$ | $\frac{14}{17} = 0{,}824$ | $\frac{11}{17} = 0{,}647$ | $\frac{16}{17} = 0{,}941$ | $\frac{12}{17} = 0{,}706$ | $\frac{13}{17} = 0{,}765$ |
| **9** | $\frac{13}{18} = 0{,}722$ | $\frac{16}{18} = 0{,}889$ | $\frac{13}{18} = 0{,}722$ | $\frac{0}{18} = 0{,}000$ | $\frac{13}{18} = 0{,}722$ | $\frac{0}{18} = 0{,}000$ | $\frac{13}{18} = 0{,}722$ | $\frac{0}{18} = 0{,}000$ | $\frac{12}{18} = 0{,}667$ |
| **10** | $\frac{54}{67} = 0{,}806$ | $\frac{56}{67} = 0{,}836$ | $\frac{53}{67} = 0{,}791$ | $\frac{51}{67} = 0{,}761$ | $\frac{58}{67} = 0{,}866$ | $\frac{51}{67} = 0{,}761$ | $\frac{50}{67} = 0{,}746$ | $\frac{56}{67} = 0{,}836$ | $\frac{57}{67} = 0{,}851$ |
| **11** | $\frac{178}{210} = 0{,}848$ | $\frac{176}{210} = 0{,}838$ | $\frac{174}{210} = 0{,}829$ | $\frac{177}{210} = 0{,}843$ | $\frac{184}{210} = 0{,}876$ | $\frac{166}{210} = 0{,}790$ | $\frac{179}{210} = 0{,}852$ | $\frac{184}{210} = 0{,}876$ | $\frac{186}{210} = 0{,}886$ |
| **12** | $\frac{7}{8} = 0{,}875$ | $\frac{6}{8} = 0{,}750$ | $\frac{7}{8} = 0{,}875$ | $\frac{5}{8} = 0{,}625$ | $\frac{6}{8} = 0{,}750$ | $\frac{5}{8} = 0{,}625$ | $\frac{6}{8} = 0{,}750$ | $\frac{7}{8} = 0{,}875$ | $\frac{7}{8} = 0{,}875$ |
| **13** | $\frac{160}{194} = 0{,}825$ | $\frac{161}{194} = 0{,}830$ | $\frac{152}{194} = 0{,}784$ | $\frac{156}{194} = 0{,}804$ | $\frac{153}{194} = 0{,}789$ | $\frac{154}{194} = 0{,}794$ | $\frac{158}{194} = 0{,}814$ | $\frac{161}{194} = 0{,}830$ | $\frac{160}{194} = 0{,}825$ |
| **14** | $\frac{11}{17} = 0{,}647$ | $\frac{8}{17} = 0{,}471$ | $\frac{12}{17} = 0{,}706$ | $\frac{13}{17} = 0{,}765$ | $\frac{11}{17} = 0{,}647$ | $\frac{12}{17} = 0{,}706$ | $\frac{12}{17} = 0{,}706$ | $\frac{2}{17} = 0{,}118$ | $\frac{11}{17} = 0{,}647$ |
| **TOTAL** | $\frac{658}{893} = 0{,}737$ | $\frac{671}{893} = 0{,}751$ | $\frac{641}{893} = 0{,}718$ | $\frac{645}{893} = 0{,}722$ | $\frac{685}{893} = 0{,}767$ | $\frac{628}{893} = 0{,}703$ | $\frac{658}{893} = 0{,}737$ | $\frac{645}{893} = 0{,}722$ | $\frac{672}{893} = 0{,}753$ |

Tabela 3.8: Comparação entre as localizações subcelulares previstas pelos classificadores T-B2/CC-SMO, T-B2/CC-J48 e T-B2/LC-SMO, quando aplicados ao conjunto de dados P-B2, pelos classificadores T-C2/CC-SMO, T-C2/CC-J48, T-C2/LC-SMO e T-C2/LC-J48, quando aplicados ao conjunto de dados P-C2, e pelos predictores mLASSO e mEN com as localizações subcelulares determinadas experimentalmente extraídas da base de dados UniProtKB/Swiss-Prot.
*1. centrossoma; 2. citoplasma; 3. citoesqueleto; 4. retículo endoplasmático; 5. endossoma; 6. espaço extracelular; 7. complexo de Golgi; 8. lisossoma; 9. microssoma; 10. mitocôndria; 11. núcleo; 12. peroxissoma; 13. membrana plasmática e 14. sinapse.*

| | **P-B2** | | | **P-C2** | | | | **mLASSO** | **mEN** |
|---|---|---|---|---|---|---|---|---|---|
| | **CC-SMO** | **CC-J48** | **LC-SMO** | **CC-SMO** | **CC-J48** | **LC-SMO** | **LC-J48** | | |
| **1** | $\frac{0}{5} = 0{,}000$ | $\frac{4}{5} = 0{,}800$ | $\frac{1}{5} = 0{,}200$ | $\frac{1}{5} = 0{,}200$ | $\frac{4}{5} = 0{,}800$ | $\frac{2}{5} = 0{,}400$ | $\frac{5}{5} = 1{,}000$ | $\frac{3}{5} = 0{,}600$ | $\frac{4}{5} = 0{,}800$ |
| **2** | $\frac{44}{66} = 0{,}667$ | $\frac{40}{66} = 0{,}606$ | $\frac{40}{66} = 0{,}606$ | $\frac{49}{66} = 0{,}742$ | $\frac{43}{66} = 0{,}652$ | $\frac{40}{66} = 0{,}606$ | $\frac{41}{66} = 0{,}621$ | $\frac{41}{66} = 0{,}621$ | $\frac{38}{66} = 0{,}576$ |
| **3** | $\frac{4}{11} = 0{,}364$ | $\frac{3}{11} = 0{,}273$ | $\frac{4}{11} = 0{,}364$ | $\frac{1}{11} = 0{,}091$ | $\frac{2}{11} = 0{,}182$ | $\frac{4}{11} = 0{,}364$ | $\frac{2}{11} = 0{,}182$ | $\frac{2}{11} = 0{,}182$ | $\frac{2}{11} = 0{,}182$ |
| **4** | $\frac{13}{17} = 0{,}765$ | $\frac{16}{17} = 0{,}941$ | $\frac{13}{17} = 0{,}765$ | $\frac{13}{17} = 0{,}765$ | $\frac{16}{17} = 0{,}941$ | $\frac{14}{17} = 0{,}824$ | $\frac{12}{17} = 0{,}706$ | $\frac{11}{17} = 0{,}647$ | $\frac{10}{17} = 0{,}588$ |
| **5** | $\frac{0}{12} = 0{,}000$ | $\frac{4}{12} = 0{,}333$ | $\frac{0}{12} = 0{,}000$ | $\frac{0}{12} = 0{,}000$ | $\frac{2}{12} = 0{,}167$ | $\frac{0}{12} = 0{,}000$ | $\frac{2}{12} = 0{,}167$ | $\frac{1}{12} = 0{,}083$ | $\frac{1}{12} = 0{,}083$ |
| **6** | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ |
| **7** | $\frac{14}{20} = 0{,}700$ | $\frac{16}{20} = 0{,}800$ | $\frac{14}{20} = 0{,}700$ | $\frac{16}{20} = 0{,}800$ | $\frac{15}{20} = 0{,}750$ | $\frac{13}{20} = 0{,}650$ | $\frac{14}{20} = 0{,}700$ | $\frac{12}{20} = 0{,}600$ | $\frac{14}{20} = 0{,}700$ |
| **8** | $\frac{6}{9} = 0{,}667$ | $\frac{8}{9} = 0{,}889$ | $\frac{6}{9} = 0{,}667$ | $\frac{6}{9} = 0{,}667$ | $\frac{8}{9} = 0{,}889$ | $\frac{5}{9} = 0{,}556$ | $\frac{8}{9} = 0{,}889$ | $\frac{6}{9} = 0{,}667$ | $\frac{6}{9} = 0{,}667$ |
| **9** | $\frac{1}{1} = 1{,}000$ | $\frac{1}{1} = 1{,}000$ | $\frac{1}{1} = 1{,}000$ | $\frac{0}{1} = 0{,}000$ | $\frac{1}{1} = 1{,}000$ | $\frac{0}{1} = 0{,}000$ | $\frac{1}{1} = 1{,}000$ | $\frac{0}{1} = 0{,}000$ | $\frac{1}{1} = 1{,}000$ |
| **10** | $\frac{10}{12} = 0{,}833$ | $\frac{9}{12} = 0{,}750$ | $\frac{9}{12} = 0{,}750$ | $\frac{10}{12} = 0{,}833$ | $\frac{9}{12} = 0{,}750$ | $\frac{9}{12} = 0{,}750$ | $\frac{9}{12} = 0{,}750$ | $\frac{10}{12} = 0{,}833$ | $\frac{10}{12} = 0{,}833$ |
| **11** | $\frac{63}{73} = 0{,}863$ | $\frac{60}{73} = 0{,}822$ | $\frac{62}{73} = 0{,}849$ | $\frac{61}{73} = 0{,}836$ | $\frac{66}{73} = 0{,}904$ | $\frac{62}{73} = 0{,}849$ | $\frac{61}{73} = 0{,}836$ | $\frac{65}{73} = 0{,}890$ | $\frac{65}{73} = 0{,}890$ |
| **12** | $\frac{2}{2} = 1{,}000$ | $\frac{1}{2} = 0{,}500$ | $\frac{2}{2} = 1{,}000$ | $\frac{1}{2} = 0{,}500$ | $\frac{1}{2} = 0{,}500$ | $\frac{1}{2} = 0{,}500$ | $\frac{1}{2} = 0{,}500$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ |
| **13** | $\frac{53}{70} = 0{,}757$ | $\frac{58}{70} = 0{,}829$ | $\frac{49}{70} = 0{,}700$ | $\frac{51}{70} = 0{,}729$ | $\frac{49}{70} = 0{,}700$ | $\frac{53}{70} = 0{,}757$ | $\frac{57}{70} = 0{,}814$ | $\frac{55}{70} = 0{,}786$ | $\frac{58}{70} = 0{,}829$ |
| **14** | $\frac{2}{2} = 1{,}000$ | $\frac{1}{2} = 0{,}500$ | $\frac{2}{2} = 1{,}000$ | $\frac{2}{2} = 1{,}000$ | $\frac{1}{2} = 0{,}500$ | $\frac{2}{2} = 1{,}000$ | $\frac{1}{2} = 0{,}500$ | $\frac{1}{2} = 0{,}500$ | $\frac{1}{2} = 0{,}500$ |
| **TOTAL** | $\frac{214}{302} = 0{,}709$ | $\frac{223}{302} = 0{,}738$ | $\frac{205}{302} = 0{,}679$ | $\frac{213}{302} = 0{,}705$ | $\frac{219}{302} = 0{,}725$ | $\frac{207}{302} = 0{,}685$ | $\frac{216}{302} = 0{,}715$ | $\frac{211}{302} = 0{,}699$ | $\frac{214}{302} = 0{,}709$ |

## 3.4 Material Suplementar

O material suplementar a este documento contém os seguintes ficheiros:

- **Conjuntos de dados:** conjuntos de dados utilizados no estudo: T-A1, T-A2, T-B1, T-B2, T-C1, T-C2, P-A1, P-A2, P-B1, P-B2, P-C1 e P-C2;
- **Modelos de previsão:** modelos de previsão gerados pelo MEKA e resultados estatísticos: T-B2/CC-J48, T-B2/CC-SMO, T-B2/LC-SMO, T-C2/CC-J48, T-C2/CC-SMO, T-C2/LC-J48 e T-C2/LC-SMO;
- **Previsões:** previsões das localizações subcelulares das 799 proteínas em estudo, previstas pelos modelos de previsão gerados pelo MEKA, assim como pelos predictores mLASSO e mEN. Complementarmente, fornece, também, as localizações extraídas das bases de dados UniProtKB/Swiss-Prot, para o mesmo conjunto de proteínas.

# Capítulo 4

# Conclusão

Conhecer a localização subcelular de uma dada proteína é particularmente importante para a anotação funcional da mesma, pois o seu papel na célula está intimamente correlacionado com o compartimento ou organelo em que esta reside.

Neste estudo foram apresentados vários predictores *multi-label* disponíveis no MEKA, com o objectivo de prever a localização subcelular de proteínas codificadas por 800 genes humanos. Complementarmente, foram exploradas e comparadas diferentes abordagens para construção dos vectores representativos das proteínas.

Numa primeira fase, através do método de avaliação *10-fold cross-validation*, foi possível perceber que a abordagem que recorre à frequência do termo GO é superior à que utiliza apenas a presença ou ausência do mesmo, uma vez que realça os termos GO anotados com mais frequência. Por outro lado, os métodos de transformação do problema LC e CC apresentaram melhores resultados que o BR, pois consideram que existe dependência entre *labels* (localizações subcelulares). Já os classificadores base *single-label* SMO e J48 tiveram um desempenho superior ao PART. Por último, a utilização dos termos GO mais relevantes, isto é, que desempenham um papel mais evidente na determinação da localização subcelular, mostrou ser mais eficiente que a utilização de todos os termos GO para cada proteína. Posto isto, os sete predictores que se destacaram, ao apresentarem uma taxa de sucesso global entre 69,2 e 72,2% (OAA) e entre 75,8 e 82,1% (OLA), foram o T-B2/CC-SMO, o T-B2/CC-J48, o T-B2/LC-SMO, o T-C2/CC-SMO, o T-C2/CC-J48, o T-C2/LC-SMO e o T-C2/LC-J48.

Numa segunda fase, através do método de avaliação *Leave-one-out cross-validation*, os sete predictores referidos anteriormente atingiram uma taxa de sucesso entre 69,2

e 72,3% (OAA) e entre 76,1 e 80,3% (OLA). Assim, apresentaram um desempenho superior aos predictores Hum-mPLoc 2.0 e iLoc-Hum e inferior aos predictores mLASSO e mEN. No entanto, o predictor T-C2/LC-J48 foi superior na previsão das localizações endossoma, espaço extracelular e membrana plasmática e o predictor T-B2/CC-J48 das localizações complexo de Golgi, peroxissoma e, juntamente com o predictor T-C2/CC-J48, microssoma.

De acordo com os predictores utilizados, a maioria das 799 proteínas está associada a apenas uma localização subcelular, enquanto que no máximo 8 proteínas estão localizadas, ou movem-se entre, duas localizações subcelulares. O citoplasma, o núcleo e a membrana celular foram as localizações subcelulares previstas para o maior número de proteínas. Por outro lado, as localizações centrossoma, citoesqueleto, endossoma, peroxissoma e sinapse foram associadas a menos de 15 proteínas.

Por fim, existem algumas limitações nos métodos/abordagens utilizadas, especialmente no que respeita às proteínas recém descobertas ou ao conjunto de dados de treino utilizado. Assim, no futuro, seria interessante, por exemplo, recorrer ao *text mining* para encontrar anotações [147], incluir a informação relativa aos códigos de evidência do *Gene Ontology* e utilizar um conjunto de dados de treino mais homogéneo, quanto ao número de proteínas locativas por localização subcelular.

# Referências bibliográficas


[1] Hong-Bin Shen and Kuo-Chen Chou. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical biochemistry*, 394(2):269–274, 2009.

[2] Kuo-Chen Chou, Zhi-Cheng Wu, and Xuan Xiao. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, 8(2):629–641, 2012.

[3] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization predictor. *Journal of theoretical biology*, 382:223–234, 2015.

[4] Tim Radford. Metaphors and dreams: the paradox of the DNA revolution is that it shows us a shining future without telling us how to get there.(Double Helix Jubilee). *The Scientist*, 17(1): 24–27, 2003.

[5] H-B Shen, Jie Yang, and K-C Chou. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino acids*, 33(1):57–67, 2007.

[6] GM. Cooper. *The Cell: A Molecular Approach*. Sinauer Associates Inc, 2 edition, 2000.

[7] Shibiao Wan and Man-Wai Mak. *Machine Learning for protein subcellular localization prediction*. Walter de Gruyter GmbH & Co KG, 1 edition, 2015. ISBN 9781501501500.

[8] Kuo-Chen Chou and Hong-Bin Shen. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of proteome research*, 6(5):1728–1734, 2007.

[9] Hong-Bin Shen and Kuo-Chen Chou. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and biophysical research communications*, 355(4):1006–1011, 2007.

[10] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins. *BMC bioinformatics*, 17(1):1, 2016.

[11] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Analytical biochemistry*, 473:14–27, 2015.

[12] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. R3P-Loc: A compact multi-label predictor using ridge regression and random projection for protein subcellular localization. *Journal of theoretical biology*, 360:34–45, 2014.

[13] Michael D Kaytor and Stephen T Warren. Aberrant protein deposition and neurological disease. *Journal of Biological Chemistry*, 274(53):37507–37510, 1999.

[14] Mien-Chie Hung and Wolfgang Link. Protein localization in disease and therapy. *J Cell Sci*, 124 (20):3381–3392, 2011.

[15] V Krutovskikh, G Mazzoleni, N Mironov, Y Omori, A-M Aguelon, M Mesnil, F Berger, C Partensky, and H Yamasaki. Altered homologous and heterologous gap-junctional intercellular communication in primary human liver tumors associated with aberrant protein localization but not gene mutation of connexin 32. *International journal of cancer*, 56(1):87–94, 1994.

[16] Yumay Chen, Chi-Fen Chen, Daniel J Riley, D Craig Allred, et al. Aberrant subcellular localization of BRCA1 in breast cancer. *Science*, 270(5237):789, 1995.

[17] X Lee, JC Keith, N Stumm, I Moutsatsos, JM McCoy, CP Crum, D Genest, D Chin, C Ehrenfels, Robert Pijnenborg, et al. Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia. *Placenta*, 22(10):808–812, 2001.

[18] Atsushi Hayama, Tatemitsu Rai, Sei Sasaki, and Shinichi Uchida. Molecular mechanisms of Bartter syndrome caused by mutations in the BSND gene. *Histochemistry and cell biology*, 119 (6):485–493, 2003.

[19] S Demolombe, I Baro, M Laurent, AS Hongre, A Pavirani, and D Escande. Abnormal subcellular localization of mutated cftr protein in a cystic fibrosis epithelial cell line. *European journal of cell biology*, 65(1):214–219, 1994.

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://doi.acm.org/10.1145/1656274.1656278`.

[21] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016. URL `http://jmlr.org/papers/v17/12-164.html`.

[22] Olesya V Stepanenko, Vladislav V Verkhusha, Irina M Kuznetsova, Vladimir N Uversky, and KK Turoverov. Fluorescent proteins as biomarkers and biosensors: throwing color lights on molecular and cellular processes. *Current Protein and Peptide Science*, 9(4):338–369, 2008.

[23] Rafael Yuste. Fluorescence microscopy today. *Nature methods*, 2(12):902–904, 2005.

[24] John Baynes and Marek H Dominiczak. *Medical biochemistry*. Elsevier Health Sciences, 2014.

[25] Terry M Mayhew and John M Lucocq. Developments in cell biology for quantitative immunoelectron microscopy based on thin sections: a review. *Histochemistry and cell biology*, 130(2): 299–313, 2008.

[26] William Margolin. Green fluorescent protein as a reporter for macromolecular localization in bacterial cells. *Methods*, 20(1):62–72, 2000.

[27] Kuo-Chen Chou and Hong-Bin Shen. Large-scale plant protein subcellular location prediction. *Journal of cellular biochemistry*, 100(3):665–678, 2007.

[28] Kuo-Chen Chou, Zhi-Cheng Wu, and Xuan Xiao. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PloS one*, 6(3):e18258, 2011.

[29] Xuan Xiao, Zhi-Cheng Wu, and Kuo-Chen Chou. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology*, 284(1):42–51, 2011.

[30] Hiroshi Nakashima and Ken Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of molecular biology*, 238(1): 54–61, 1994.

[31] Juan Cedano, Patrick Aloy, Josep A Perez-Pons, and Enrique Querol. Relation between amino acid composition and cellular location of proteins. *Journal of molecular biology*, 266(3):594–600, 1997.

[32] Kuo-Chen Chou and David W Elrod. Protein subcellular location prediction. *Protein engineering*, 12(2):107–118, 1999.

[33] Keun-Joon Park and Minoru Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13): 1656–1663, 2003.

[34] Kuo-Chen Chou and Hong-Bin Shen. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS One*, 5(4):e9931, 2010.

[35] Kuo-Chen Chou and Hong-Bin Shen. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols*, 3(2):153–162, 2008.

[36] Kuo-Chen Chou, Hong-Bin Shen, et al. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science*, 2(10):1090, 2010.

[37] Kuo-Chen Chou and Hong-Bin Shen. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochemical and biophysical research communications*, 347(1): 150–157, 2006.

[38] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *Journal of Theoretical Biology*, 323:40–48, 2013.

[39] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. GOASVM: Protein subcellular localization prediction based on gene ontology annotation and SVM. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2229–2232. IEEE, 2012.

[40] Kuo-Chen Chou and Hong-Bin Shen. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PloS one*, 5(6):e11335, 2010.

[41] Zhi-Cheng Wu, Xuan Xiao, and Kuo-Chen Chou. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Molecular BioSystems*, 7(12):3287–3297, 2011.

[42] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC bioinformatics*, 13(1):1, 2012.

[43] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Hybridgo-loc: Mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. *PLoS One*, 9(3): e89545, 2014.

[44] Hong-Bin Shen and Kuo-Chen Chou. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, 85(3):233–240, 2007.

[45] Hong-Bin Shen and Kuo-Chen Chou. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Journal of Biomolecular Structure and Dynamics*, 28(2):175–186, 2010.

[46] Hong-Bin Shen and Kuo-Chen Chou. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Engineering Design and Selection*, 20 (1):39–46, 2007.

[47] Hong-Bin Shen and Kuo-Chen Chou. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein and peptide letters*, 16(12):1478–1484, 2009.

[48] Zhi-Cheng Wu, Xuan Xiao, and Kuo-Chen Chou. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. *Protein and peptide letters*, 19(1):4–14, 2012.

[49] Kuo-Chen Chou and Hong-Bin Shen. Large-scale predictions of Gram-negative bacterial protein subcellular locations. *Journal of proteome research*, 5(12):3420–3428, 2006.

[50] Hong-Bin Shen and Kuo-Chen Chou. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *Journal of Theoretical Biology*, 264(2):326–333, 2010.

[51] Xuan Xiao, Zhi-Cheng Wu, and Kuo-Chen Chou. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PloS one*, 6(6):e20592, 2011.

[52] Guo-Ping Zhou and Kutbuddin Doctor. Subcellular location prediction of apoptosis proteins. *Proteins: Structure, Function, and Bioinformatics*, 50(1):44–48, 2003.

[53] Guo-Liang Fan and Qian-Zhong Li. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *Journal of theoretical biology*, 304:88–95, 2012.

[54] Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

[55] Monalisa Mandal, Anirban Mukhopadhyay, and Ujjwal Maulik. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. *Medical & biological engineering & computing*, 53(4):331–344, 2015.

[56] Xiao Wang, Weiwei Zhang, Qiuwen Zhang, and Guo-Zheng Li. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics*, 31(16):2639–2645, 2015.

[57] Kuo-Chen Chou and David W Elrod. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochemical and Biophysical Research Communications*, 252(1): 63–68, 1998.

[58] Man-Wai Mak, Jian Guo, and Sun-Yuan Kung. PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(3):416–422, 2008.

[59] Richard Mott, Jörg Schultz, Peer Bork, and Chris P Ponting. Predicting protein cellular localization using a domain projection method. *Genome research*, 12(8):1168–1174, 2002.

[60] Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, and Abdul Sattar. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou s general PseAAC. *Journal of theoretical biology*, 364: 284–294, 2015.

[61] Zhiyong Lu, Duane Szafron, Russell Greiner, Paul Lu, David S Wishart, Brett Poulin, John Anvik, Cam Macdonell, and Roman Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.

[62] Shao-Wu Zhang, Yun-Long Zhang, Hui-Fang Yang, Chun-Hui Zhao, and Quan Pan. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids*, 34(4):565–572, 2008.

[63] Kenta Nakai and Minoru Kanehisa. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Bioinformatics*, 11(2):95–110, 1991.

[64] Olof Emanuelsson, Henrik Nielsen, Søren Brunak, and Gunnar Von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016, 2000.

[65] Henrik Nielsen, Jacob Engelbrecht, Søren Brunak, and Gunnar Von Heijne. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *International journal of neural systems*, 8(05n06):581–599, 1997.

[66] Kuo-Chen Chou and Yu-Dong Cai. Predicting protein localization in budding yeast. *Bioinformatics*, 21(7):944–950, 2005.

[67] KiYoung Lee, Dae-Won Kim, DoKyun Na, Kwang H Lee, and Doheon Lee. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic acids research*, 34(17):4655–4666, 2006.

[68] Kuo-Chen Chou and Yu-Dong Cai. Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition. *Journal of cellular biochemistry*, 90(6):1250–1260, 2003.

[69] Kuo-Chen Chou and Yu-Dong Cai. Prediction of protein subcellular locations by GO–FunD–PseAA predictor. *Biochemical and Biophysical Research Communications*, 320(4):1236–1239, 2004.

[70] Hong-Bin Shen and Kuo-Chen Chou. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry*, 373(2):386–388, 2008.

[71] Kuo-Chen Chou. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, 6(4):262–274, 2009.

[72] Astrid Reinhardt and Tim Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic acids research*, 26(9):2230–2236, 1998.

[73] Aarti Garg, Manoj Bhasin, and Gajendra PS Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of biological Chemistry*, 280(15):14427–14432, 2005.

[74] Kuo-Chen Chou and Hong-Bin Shen. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of proteome research*, 5(8): 1888–1897, 2006.

[75] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Protein subcellular localization prediction based on profile alignment and Gene Ontology. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2011.

[76] Rajesh Nair and Burkhard Rost. Sequence conserved for subcellular localization. *Protein Science*, 11(12):2836–2847, 2002.

[77] Michelle S Scott, David Y Thomas, and Michael T Hallett. Predicting subcellular localization via protein motif co-occurrence. *Genome research*, 14(10a):1957–1966, 2004.

[78] Frank Eisenhaber and Peer Bork. Wanted: subcellular localization of proteins based on sequence. *Trends in cell biology*, 8(4):169–170, 1998.

[79] Sébastien Rey, Michael Acab, Jennifer L Gardy, Matthew R Laird, Christophe Lambert, Fiona SL Brinkman, et al. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic acids research*, 33(suppl 1):D164–D168, 2005.

[80] Manoj Bhasin and GPS Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic acids research*, 32 (suppl 2):W414–W419, 2004.

[81] Paul Horton, Keun-Joon Park, Takeshi Obayashi, and Kenta Nakai. Protein Subcellular Localisation Prediction with WoLF PSORT. In *APBC*, volume 39. Citeseer, 2006.

[82] Kenta Nakai. Protein sorting signals and prediction of subcellular localization. *Advances in protein chemistry*, 54:277–344, 2000.

[83] Lila M Gierasch. Signal sequences. *Biochemistry*, 28(3):923–930, 1989.

[84] Laura J Mauro and Jack E Dixon. Zip codes' direct intracellular protein tyrosine phosphatases to the correct cellular 'address. *Trends in biochemical sciences*, 19(4):151–155, 1994.

[85] MO Lively. Signal peptidases in protein biosynthesis and intracellular transport. *Current opinion in cell biology*, 1(6):1188–1193, 1989.

[86] Ross E Dalbey and Gunnar von Heijne. Signal peptidases in prokaryotes and eukaryotes-a new protease family. *Trends in biochemical sciences*, 17(11):474–478, 1992.

[87] Bruno Martoglio and Bernhard Dobberstein. Signal sequences: more than just greasy peptides. *Trends in cell biology*, 8(10):410–415, 1998.

[88] Gunnar HEIJNE. Patterns of amino acids near signal-sequence cleavage sites. *European journal of biochemistry*, 133(1):17–21, 1983.

[89] Gunnar Heijne. Signal peptides. *eLS*, 1990.

[90] Yukiko Fujiwara and Minoru Asogawa. Prediction of subcellular localizations using amino acid composition and order. *Genome Informatics*, 12:103–112, 2001.

[91] Karsten Hiller, Andreas Grote, Maurice Scheer, Richard Münch, and Dieter Jahn. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic acids research*, 32(suppl 2): W375–W379, 2004.

[92] Jannick Dyrløv Bendtsen, Henrik Nielsen, Gunnar von Heijne, and Søren Brunak. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology*, 340(4):783–795, 2004.

[93] Henrik Nielsen and Anders Krogh. Prediction of signal peptides and signal anchors by a hidden Markov model. In *Ismb*, volume 6, pages 122–130, 1998.

[94] Kuo-Chen Chou and Yu-Dong Cai. Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, 277(48): 45765–45769, 2002.

[95] Yu-Dong Cai, Guo-Ping Zhou, and Kuo-Chen Chou. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical journal*, 84(5): 3257–3263, 2003.

[96] Yu-Dong Cai and Kuo-Chen Chou. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. *Journal of theoretical biology*, 238(2):395–400, 2006.

[97] Hong-Bin Shen and Kuo-Chen Chou. EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. *Biochemical and biophysical research communications*, 364(1): 53–59, 2007.

[98] Kuo-Chen Chou and Yu-Dong Cai. Predicting protein structural class by functional domain composition. *Biochemical and biophysical research communications*, 321(4):1007–1009, 2004.

[99] Kuo-Chen Chou and Hong-Bin Shen. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochemical and Biophysical Research Communications*, 376(2):321–325, 2008.

[100] Hong-Bin Shen and Kuo-Chen Chou. Identification of proteases and their types. *Analytical biochemistry*, 385(1):153–160, 2009.

[101] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, et al. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):1, 2003.

[102] Robert D Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, et al. Pfam: clans, web tools and services. *Nucleic acids research*, 34(suppl 1):D247–D251, 2006.

[103] Ivica Letunic, Richard R Copley, Birgit Pils, Stefan Pinkert, Jörg Schultz, and Peer Bork. SMART 5: domains in the context of genomes and networks. *Nucleic acids research*, 34(suppl 1):D257–D260, 2006.

[104] Aron Marchler-Bauer, John B Anderson, Myra K Derbyshire, Carol DeWeese-Scott, Noreen R Gonzales, Marc Gwadz, Luning Hao, Siqian He, David I Hurwitz, John D Jackson, et al. CDD: a conserved domain database for interactive domain family analysis. *Nucleic acids research*, 35 (suppl 1):D237–D240, 2007.

[105] Alejandro A Schäffer, L Aravind, Thomas L Madden, Sergei Shavirin, John L Spouge, Yuri I Wolf, Eugene V Koonin, and Stephen F Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, 29(14):2994–3005, 2001.

[106] Kuo-Chen Chou and Hong-Bin Shen. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochemical and biophysical research communications*, 360(2):339–345, 2007.

[107] Shao-Wu Zhang, Yan-Fang Liu, Yong Yu, Ting-He Zhang, and Xiao-Nan Fan. MSLoc-DT: A new method for predicting the protein subcellular location of multispecies based on decision templates. *Analytical biochemistry*, 449:164–171, 2014.

[108] Suyu Mei, Wang Fei, and Shuigeng Zhou. Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics*, 12(1):1, 2011.

[109] Yang Yang and Bao-Liang Lu. Protein subcellular multi-localization prediction using a min-max modular support vector machine. *International Journal of Neural Systems*, 20(01):13–28, 2010.

[110] Lili Liu, Zijun Zhang, Qian Mei, and Ming Chen. PSI: a comprehensive and integrative approach for accurate plant subcellular localization prediction. *PloS one*, 8(10):e75826, 2013.

[111] Wei-Zhong Lin, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems*, 9(4):634–644, 2013.

[112] Suyu Mei. Multi-label multi-kernel transfer learning for human protein subcellular localization. *PLoS One*, 7(6):e37716, 2012.

[113] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Adaptive thresholding for multi-label SVM classification with application to protein subcellular localization prediction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3547–3551. IEEE, 2013.

[114] Shibiao Wan, Man-Wai Mak, Sun-Yuan Kung, et al. Semantic similarity over gene ontology for multi-label protein subcellular localization. *Engineering*, 5(10):68, 2013.

[115] Shibiao Wan, Man-Wai Mak, Bai Zhang, Yue Wang, and Sun-Yuan Kung. Ensemble random projection for multi-label classification with application to protein subcellular localization. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5999–6003. IEEE, 2014.

[116] Alona Fyshe, Yifeng Liu, Duane Szafron, Russ Greiner, and Paul Lu. Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics*, 24(21): 2512–2517, 2008.

[117] Scott Brady and Hagit Shatkay. EpiLoc: a (working) text-based system for predicting protein subcellular location. In *Pacific Symposium on Biocomputing*, volume 13, pages 604–615, 2008.

[118] Wen-Lin Huang, Chun-Wei Tung, Shih-Wen Ho, Shiow-Fen Hwang, and Shinn-Ying Ho. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC bioinformatics*, 9(1):1, 2008.

[119] Sang-Mun Chi and Dougu Nam. WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms. *Bioinformatics*, 28(7):1028–1030, 2012.

[120] Zhiyong Lu and Lawrence Hunter. GO molecular function terms are predictive of subcellular localization. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 151. NIH Public Access, 2005.

[121] Evelyn Camon, Michele Magrane, Daniel Barrell, David Binns, Wolfgang Fleischmann, Paul Kersey, Nicola Mulder, Tom Oinn, John Maslen, Anthony Cox, et al. The gene ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome research*, 13(4):662–672, 2003.

[122] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

[123] Evgeni M Zdobnov and Rolf Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848, 2001.

[124] Torsten Blum, Sebastian Briesemeister, and Oliver Kohlbacher. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, 10(1):1, 2009.

[125] Jianjun He, Hong Gu, and Wenqi Liu. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PloS one*, 7 (6):e37155, 2012.

[126] Thai Quang Tung and Doheon Lee. A method to improve protein subcellular localization prediction by integrating various biological data sources. *BMC bioinformatics*, 10(1):1, 2009.

[127] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[128] Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.

[129] Kuo-Chen Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21(1):10–19, 2005.

[130] Francisco M Couto and H Sofia Pinto. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of bioinformatics and computational biology*, 11(05):1371001, 2013.

[131] Jennifer L Gardy, Cory Spencer, Ke Wang, Martin Ester, Gabor E Tusnady, István Simon, Sujun Hua, Christophe Lambert, Kenta Nakai, Fiona SL Brinkman, et al. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic acids research*, 31 (13):3613–3617, 2003.

[132] Annette Höglund, Pierre Dönnes, Torsten Blum, Hans-Werner Adolph, and Oliver Kohlbacher. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, 22(10):1158–1165, 2006.

[133] Piyushkumar Mundra, Madhan Kumar, K Krishna Kumar, Valadi K Jayaraman, and Bhaskar D Kulkarni. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognition Letters*, 28(13):1610–1615, 2007.

[134] E Tantoso and Kuo-Bin Li. AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. *Amino Acids*, 35(2):345–353, 2008.

[135] A Harvey Millar, Chris Carrie, Barry Pogson, and James Whelan. Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *The Plant Cell*, 21(6):1625–1631, 2009.

[136] Leonard J Foster, Carmen L de Hoog, Yanling Zhang, Yong Zhang, Xiaohui Xie, Vamsi K Mootha, and Matthias Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, 2006.

[137] Robert F Murphy. Communicating subcellular distributions. *Cytometry Part A*, 77(7):686–692, 2010.

[138] Song Zhang, Xuefeng Xia, Jincheng Shen, Yun Zhou, and Zhirong Sun. DBMLoc: a Database of proteins with multiple subcellular localizations. *BMC bioinformatics*, 9(1):1, 2008.

[139] C Smith. Subcellular targeting of proteins and drugs. *http://wwwbiocomparecom/Articles/TechnologySpotlight/976/Subcellular-Targeting-Of-Proteins-And-Drugshtml.*, 2008.

[140] Estelle Glory and Robert F Murphy. Automated subcellular location determination and high-throughput microscopy. *Developmental cell*, 12(1):7–16, 2007.

[141] Jack H Wong and Tzi Bun Ng. Studies on an antifungal protein and a chromatographically and structurally related protein isolated from the culture broth of bacillus amyloliquefaciens. *Protein and peptide letters*, 16(11):1399–1406, 2009.

[142] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[143] Kenta Nakai and Paul Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in biochemical sciences*, 24(1):34–35, 1999.

[144] Kuo-Chen Chou and Hong-Bin Shen. Recent progress in protein subcellular location prediction. *Analytical biochemistry*, 370(1):1–16, 2007.

[145] George K Acquaah-Mensah, Sonia M Leach, and Chittibabu Guda. Predicting the subcellular localization of human proteins using machine learning and exploratory data analysis. *Genomics, proteomics & bioinformatics*, 4(2):120–133, 2006.

[146] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'donovan, and Rolf Apweiler. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

[147] Francisco M Couto, Mário J Silva, Vivian Lee, Emily Dimmer, Evelyn Camon, Rolf Apweiler, Harald Kirsch, and Dietrich Rebholz-Schuhmann. GOAnnotator: linking protein GO annotations to evidence text. *Journal of biomedical discovery and collaboration*, 1(1):1, 2006.

# Anexos

# A - Previsão da Localização Subcelular das Proteínas através do software MEKA

Resultados da Previsão da Localização Subcelular das Proteínas através do software MEKA.
*Os valores entre 1 e 14 correspondem às localizações subcelulares: 1 - centrossoma; 2 - citoplasma; 3 - citoesqueleto; 4 - retículo endoplasmático; 5 - endossoma; 6 - espaço extracelular; 7 - complexo de Golgi; 8 - lisossoma; 9 - microssoma; 10 - mitocôndria; 11 - núcleo; 12 - peroxissoma; 13 - membrana plasmática e 14 - sinapse. O valores entre 1 (vermelho) e 7 (verde) correspondem ao número de classificadores que associaram determinada proteína a uma localização subcelular. Os valores a negrito correspondem às localizações extraídas da base de dados UniProtKB/Swiss-Prot para determinada proteína.*

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ENSG00000000938/P09769 | 1 | **0** | **6** | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | **0** | 0 |
| 2 | ENSG00000002587/O14792 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 3 | ENSG00000005001/Q9GZN4 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 4 | ENSG00000005007/Q92900 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 5 | ENSG00000005206/Q8TCT7 | 0 | 0 | 0 | 0 | **0** | 0 | **2** | **3** | 0 | 0 | 0 | 0 | **7** | 0 |
| 6 | ENSG00000005381/P05164 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | ENSG00000005882/Q15119 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 8 | ENSG00000006062/Q99558 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | ENSG00000006534/P43353 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 |
| 10 | ENSG00000007168/P43034 | **4** | **2** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 11 | ENSG00000007237/O60861 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | ENSG00000007264/P42679 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | ENSG00000007933/P31513 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 |
| 14 | ENSG00000008018/P20618 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 15 | ENSG00000008118/Q96NX5 | 0 | **6** | 0 | 0 | 1 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | **0** | 0 |
| 16 | ENSG00000008277/Q9P0K1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 17 | ENSG00000011052/P15531 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 18 | ENSG00000012983/Q9Y4K4 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | ENSG00000014216/P07384 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 |
| 20 | ENSG00000015171/Q15326 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 21 | ENSG00000019991/P14210 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | ENSG00000020256/Q9NPA5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 23 | ENSG00000020256/Q9NTW7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 24 | ENSG00000023330/P13196 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |

## . A - PREVISÃO DA LOCALIZAÇÃO SUBCELULAR DAS PROTEÍNAS ATRAVÉS DO SOFTWARE MEKA

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | ENSG00000025708/P19971 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 26 | ENSG00000026751/Q9NQ25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 27 | ENSG00000027644/P14616 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 28 | ENSG00000033627/Q93050 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 29 | ENSG00000034053/Q99767 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 |
| 30 | ENSG00000039139/Q8TE73 | 0 | **3** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 31 | ENSG00000039319/Q7Z3T8 | 0 | **7** | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | ENSG00000039987/Q8NFU1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 33 | ENSG00000042286/Q9BRQ8 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 |
| 34 | ENSG00000047315/P30876 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 35 | ENSG00000047936/P08922 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 36 | ENSG00000049089/Q14055 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 37 | ENSG00000049192/Q9UKP5 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 38 | ENSG00000049768/Q9BZS1 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 39 | ENSG00000055118/Q12809 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 40 | ENSG00000055609/Q8NEZ4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 41 | ENSG00000058063/Q9Y2G3 | 0 | 0 | 0 | **6** | **0** | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 42 | ENSG00000058729/Q9BVS4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 43 | ENSG00000059573/P54886 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 44 | ENSG00000062282/Q96PD7 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | ENSG00000063245/Q9Y6I3 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | **1** | 0 |
| 46 | ENSG00000063438/A9YTQ3 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 47 | ENSG00000064012/Q14790 | 1 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | ENSG00000064225/Q9Y274 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | ENSG00000064490/O14593 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 50 | ENSG00000064989/Q16602 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | **5** | 0 |
| 51 | ENSG00000065361/P21860 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 52 | ENSG00000066230/P48764 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 53 | ENSG00000066322/Q9BW60 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 54 | ENSG00000066827/Q9P243 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 55 | ENSG00000067992/Q15120 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 56 | ENSG00000068745/Q9UHH9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 57 | ENSG00000069431/O60706 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 58 | ENSG00000069535/P27338 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 59 | ENSG00000069667/P35398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 60 | ENSG00000069696/P21917 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 61 | ENSG00000070614/P52848 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 62 | ENSG00000070729/Q14028 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 63 | ENSG00000070785/Q9NR50 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 64 | ENSG00000070808/Q9UQM7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | **7** | **0** |
| 65 | ENSG00000070886/P29322 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 66 | ENSG00000070961/P20020 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 67 | ENSG00000071203/Q9NXJ0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 68 | ENSG00000071575/Q92519 | 0 | **4** | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 69 | ENSG00000072062/P17612 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **2** | 0 | **2** | 0 |
| 70 | ENSG00000072210/P51648 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 71 | ENSG00000072274/P02786 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 72 | ENSG00000072694/P31994 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 73 | ENSG00000072778/P49748 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 74 | ENSG00000072832/Q14194 | **3** | **4** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75 | ENSG00000073803/O43283 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 76 | ENSG00000073969/P46459 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 77 | ENSG00000074201/P54105 | 0 | **2** | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 78 | ENSG00000074219/Q15562 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 79 | ENSG00000074771/Q9HBY0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 80 | ENSG00000075292/Q14966 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 81 | ENSG00000075914/Q15024 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 82 | ENSG00000075975/Q9H000 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 83 | ENSG00000076248/P13051 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | **6** | 0 | 1 | 0 |
| 84 | ENSG00000076258/P31512 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 |
| 85 | ENSG00000077009/Q9NPI5 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 86 | ENSG00000077044/Q16760 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 87 | ENSG00000077238/P24394 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 88 | ENSG00000077616/Q9Y3Q0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 89 | ENSG00000078061/P10398 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 90 | ENSG00000078549/P41586 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 91 | ENSG00000078618/O43847 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 92 | ENSG00000079112/Q12864 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 93 | ENSG00000079335/Q9UNH5 | **4** | **3** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 94 | ENSG00000079337/O95398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 95 | ENSG00000079459/P37268 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 96 | ENSG00000079482/O60890 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | **1** |
| 97 | ENSG00000079805/P50570 | 0 | **1** | **1** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | **2** | **5** |
| 98 | ENSG00000081181/P78540 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 99 | ENSG00000081377/O60729 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 100 | ENSG00000081842/Q9UN73 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 101 | ENSG00000082014/Q6STE5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 102 | ENSG00000083454/Q93086 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 103 | ENSG00000084734/Q14397 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 104 | ENSG00000086061/P31689 | 0 | **3** | 0 | **0** | 0 | 0 | 0 | 0 | **2** | **1** | **2** | 0 | 1 | 0 |
| 105 | ENSG00000086159/Q13520 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 106 | ENSG00000086300/Q9Y5X0 | **2** | **3** | **0** | 4 | **1** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 107 | ENSG00000086475/P49903 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | **0** | 0 |
| 108 | ENSG00000086506/P09105 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 109 | ENSG00000087085/P22303 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | **0** | 0 | **1** | **4** |
| 110 | ENSG00000087116/O95450 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 111 | ENSG00000087191/P62195 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 112 | ENSG00000087460/O95467 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| 113 | ENSG00000087460/P63092 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 114 | ENSG00000087460/P84996 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 115 | ENSG00000087460/Q5JWF2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 116 | ENSG00000087903/P48378 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 117 | ENSG00000088002/O00204 | 0 | **5** | 0 | **0** | 0 | 0 | 0 | 0 | **3** | 0 | **0** | 0 | 0 | 0 |
| 118 | ENSG00000088451/O95455 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 119 | ENSG00000088826/Q9NWM0 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 120 | ENSG00000089820/P98171 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 121 | ENSG00000090020/P19634 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 122 | ENSG00000090674/Q9GZU1 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | **3** | 0 | 0 | 0 | 0 | **7** | 0 |
| 123 | ENSG00000091137/O43511 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 124 | ENSG00000091262/O95255 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 125 | ENSG00000091583/P02749 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 126 | ENSG00000091844/Q9UGC6 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | **0** |
| 127 | ENSG00000093010/P21964 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 1 |
| 128 | ENSG00000093134/Q9NY84 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 129 | ENSG00000094755/O00591 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** |
| 130 | ENSG00000094804/Q99741 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 131 | ENSG00000094963/Q99518 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 |
| 132 | ENSG00000095321/P43155 | 0 | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | **7** | 0 | **0** | 0 | 0 |
| 133 | ENSG00000095587/Q9Y6L7 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 134 | ENSG00000095627/Q9BXT4 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 135 | ENSG00000096384/P08238 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 136 | ENSG00000099337/Q9Y257 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 137 | ENSG00000099783/P52272 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 138 | ENSG00000100024/Q9UBR1 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 139 | ENSG00000100253/Q9UGB7 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 140 | ENSG00000100804/P28074 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 141 | ENSG00000101188/P30989 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 142 | ENSG00000101190/Q9UL49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 143 | ENSG00000101213/Q13882 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 144 | ENSG00000101266/P68400 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 145 | ENSG00000101292/Q8NFJ6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 146 | ENSG00000101361/O00567 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 147 | ENSG00000101425/P17213 | 0 | 5 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 148 | ENSG00000101445/Q96T49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | **3** | 0 |
| 149 | ENSG00000101473/O14734 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 |
| 150 | ENSG00000101557/P54578 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 151 | ENSG00000101638/O15466 | 0 | 0 | 0 | 0 | 0 | 1 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 152 | ENSG00000101665/O15105 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 153 | ENSG00000101695/Q96EQ8 | 0 | 1 | 0 | 0 | 0 | 1 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 154 | ENSG00000101782/O14730 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 155 | ENSG00000101892/Q9UN42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 2 | 0 |
| 156 | ENSG00000101958/P23416 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** |
| 157 | ENSG00000101974/Q8NB49 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 158 | ENSG00000102078/O95258 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 159 | ENSG00000102226/P51784 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 160 | ENSG00000102452/Q8IZF0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 161 | ENSG00000102554/Q13887 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 162 | ENSG00000102974/P49711 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 163 | ENSG00000103355/Q8NF86 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 164 | ENSG00000103569/O43315 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 7 | 0 |
| 165 | ENSG00000103994/Q9H2Y7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 166 | ENSG00000104142/Q9P253 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 167 | ENSG00000104154/O14863 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 5 | 0 |
| 168 | ENSG00000104213/Q15198 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 169 | ENSG00000104321/O75762 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 170 | ENSG00000104432/P13232 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 171 | ENSG00000104490/P61601 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 172 | ENSG00000104537/P27216 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 |
| 173 | ENSG00000104687/P00390 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 |
| 174 | ENSG00000104755/Q99965 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 175 | ENSG00000104808/Q9UQ10 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 176 | ENSG00000104825/Q15653 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 177 | ENSG00000104936/Q09013 | 0 | **0** | 0 | **6** | 0 | 0 | 0 | 0 | 0 | **3** | **1** | 0 | **0** | 0 |
| 178 | ENSG00000104973/Q71SY5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 179 | ENSG00000105520/Q96GM1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 180 | ENSG00000105707/P05981 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 181 | ENSG00000105726/Q9HD20 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 182 | ENSG00000105855/P26012 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 183 | ENSG00000106069/P52757 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 184 | ENSG00000106070/Q13322 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 185 | ENSG00000106105/P41250 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 186 | ENSG00000106113/Q13324 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 187 | ENSG00000106128/Q02643 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 188 | ENSG00000106211/P04792 | 0 | **7** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 189 | ENSG00000106367/P61966 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 190 | ENSG00000106628/P49005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 191 | ENSG00000106633/P35557 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 192 | ENSG00000106648/Q7Z4T8 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 193 | ENSG00000106976/Q05193 | 0 | **2** | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 194 | ENSG00000107201/O95786 | 0 | **6** | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 |
| 195 | ENSG00000107537/O14832 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 |
| 196 | ENSG00000107736/Q9H251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 197 | ENSG00000107789/Q9UNW1 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 198 | ENSG00000108010/O76003 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 |
| 199 | ENSG00000108018/Q8WY21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 200 | ENSG00000108176/Q9UKB3 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 |
| 201 | ENSG00000108262/Q9Y2X7 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 202 | ENSG00000108370/O75916 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 203 | ENSG00000108423/Q9UJT1 | **0** | **6** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 204 | ENSG00000108669/Q15438 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| 205 | ENSG00000108828/Q99536 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 |
| 206 | ENSG00000108861/P51452 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 207 | ENSG00000109103/Q13432 | **5** | **0** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 208 | ENSG00000109163/P30968 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 209 | ENSG00000109323/O00462 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 210 | ENSG00000109424/P25874 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 211 | ENSG00000109576/Q8N5Z0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 212 | ENSG00000109762/Q9H3E2 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 213 | ENSG00000110169/P02790 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 214 | ENSG00000110243/Q6Q788 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 215 | ENSG00000110455/Q96QU6 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 216 | ENSG00000110619/P49589 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 217 | ENSG00000110955/P06576 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 218 | ENSG00000110975/Q6XYQ8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 219 | ENSG00000111144/P09960 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 220 | ENSG00000111206/Q08050 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 221 | ENSG00000111218/Q9NR22 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 222 | ENSG00000111331/Q9Y6K5 | 0 | **6** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | **2** | 0 | 0 | 0 |
| 223 | ENSG00000111335/P29728 | 0 | **2** | 0 | **2** | 0 | 0 | 0 | 0 | **4** | **4** | **2** | 0 | 0 | 0 |
| 224 | ENSG00000111641/P46087 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 225 | ENSG00000111667/P45974 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 0 |
| 226 | ENSG00000111700/Q9NPD5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 227 | ENSG00000111799/Q99715 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 228 | ENSG00000111885/P33908 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 229 | ENSG00000112304/Q9NPJ3 | 0 | **0** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | **3** | **5** | 0 | 0 | 0 |
| 230 | ENSG00000112365/O43167 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 231 | ENSG00000112541/Q9Y233 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 232 | ENSG00000112695/P14406 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | **5** | 0 | 0 | 1 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 233 | ENSG00000112769/Q16363 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 234 | ENSG00000112818/Q16819 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 235 | ENSG00000113212/Q9Y5E2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 236 | ENSG00000113263/Q08881 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 237 | ENSG00000113356/O15318 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 238 | ENSG00000113441/Q9UIQ6 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 239 | ENSG00000113492/Q9BYV1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 240 | ENSG00000113494/P16471 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 241 | ENSG00000113525/P05113 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 242 | ENSG00000113638/Q6PID6 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 243 | ENSG00000113916/P41182 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 244 | ENSG00000114124/Q8WTQ7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 245 | ENSG00000114378/Q12794 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 246 | ENSG00000114867/Q04637 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 247 | ENSG00000115484/P50991 | **4** | **3** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 248 | ENSG00000115525/Q9UNP4 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 249 | ENSG00000115593/Q8NB12 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 250 | ENSG00000115661/O75716 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 251 | ENSG00000116016/Q99814 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 252 | ENSG00000116039/P15313 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 253 | ENSG00000116120/Q9NSD9 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 254 | ENSG00000116649/P19623 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 255 | ENSG00000116745/Q16518 | 0 | **2** | 0 | **0** | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | **2** | 0 |
| 256 | ENSG00000116783/Q59H18 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 |
| 257 | ENSG00000116830/Q9UNY4 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 258 | ENSG00000116996/Q12836 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 |
| 259 | ENSG00000117009/O15229 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 260 | ENSG00000117013/P56696 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 261 | ENSG00000117114/O95490 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 262 | ENSG00000117118/P21912 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 263 | ENSG00000117222/Q15291 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 264 | ENSG00000117228/P32455 | 0 | **6** | 0 | 0 | 0 | 1 | **1** | 0 | 0 | 0 | 0 | 0 | **0** | 0 |
| 265 | ENSG00000117335/P15529 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 266 | ENSG00000117448/P14550 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 267 | ENSG00000117528/P28288 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | **4** | 0 | 0 |
| 268 | ENSG00000117620/Q9Y2D2 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 269 | ENSG00000117713/O14497 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 270 | ENSG00000117751/Q12972 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 271 | ENSG00000118160/Q9UPR5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 272 | ENSG00000118946/O14917 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 273 | ENSG00000119121/Q9BX84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 274 | ENSG00000119125/Q9Y2T3 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 275 | ENSG00000119231/Q96HI0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 276 | ENSG00000119684/Q9UHC1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 277 | ENSG00000119725/Q86VK4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 278 | ENSG00000119900/Q5TC84 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 279 | ENSG00000119953/O75940 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 280 | ENSG00000120659/O14788 | 0 | **0** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 281 | ENSG00000120875/Q13115 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 282 | ENSG00000120889/O14763 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 283 | ENSG00000120907/P35348 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | **7** | 0 |
| 284 | ENSG00000121361/Q15842 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 285 | ENSG00000121417/Q13398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 286 | ENSG00000121691/P04040 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | **7** | 0 | 0 |
| 287 | ENSG00000121988/Q5FWF4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 288 | ENSG00000121989/P27037 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 289 | ENSG00000122126/Q01968 | 0 | 1 | 0 | 0 | **2** | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 290 | ENSG00000122375/Q9UHM6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 291 | ENSG00000122420/P43088 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 292 | ENSG00000122482/Q9H582 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 293 | ENSG00000122566/P22626 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 294 | ENSG00000122705/P09496 | 0 | **0** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 295 | ENSG00000122861/P00749 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 296 | ENSG00000123411/Q9H2S9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 297 | ENSG00000123500/Q03692 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 298 | ENSG00000123737/Q06265 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 299 | ENSG00000123815/Q96D53 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | **5** | 0 |
| 300 | ENSG00000123989/Q8IZ52 | 0 | **0** | 0 | 0 | 0 | 0 | **6** | 0 | 0 | **3** | 0 | 0 | 0 | 0 |
| 301 | ENSG00000124140/Q9H2X9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 302 | ENSG00000124215/Q8IXH8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 303 | ENSG00000124383/O00566 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 304 | ENSG00000124406/Q9Y2Q0 | 0 | 0 | 0 | **5** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **3** | 0 |
| 305 | ENSG00000124731/Q9NP99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 306 | ENSG00000124789/P49790 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 307 | ENSG00000125384/P43116 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 308 | ENSG00000125447/Q9NZ52 | 0 | 0 | 0 | 0 | **0** | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 309 | ENSG00000125482/Q15361 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 310 | ENSG00000125676/Q8NI27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 311 | ENSG00000125810/Q9NPY3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 312 | ENSG00000125821/Q8TEA8 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 313 | ENSG00000125851/P16519 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 314 | ENSG00000126067/P49721 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 315 | ENSG00000126091/Q11203 | 0 | 0 | 0 | 0 | 0 | 4 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 316 | ENSG00000126602/Q12931 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 317 | ENSG00000127080/Q9H8X2 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 318 | ENSG00000127220/Q96I13 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 319 | ENSG00000127328/Q96QF0 | 0 | **5** | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 320 | ENSG00000127533/Q96RI0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 321 | ENSG00000127540/O14957 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 322 | ENSG00000127554/P55789 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 |
| 323 | ENSG00000127920/P61952 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 324 | ENSG00000128313/Q9BWW9 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 325 | ENSG00000128342/P15018 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 326 | ENSG00000128510/Q9UI42 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 327 | ENSG00000128595/O43852 | 0 | 0 | 0 | **4** | 0 | 2 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 328 | ENSG00000128829/Q9P2K8 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 329 | ENSG00000128918/O94788 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 330 | ENSG00000128951/P33316 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | **3** | 0 | 0 | 0 |
| 331 | ENSG00000129128/P61009 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 |
| 332 | ENSG00000129292/A8MW92 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 333 | ENSG00000129673/Q16613 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 334 | ENSG00000129744/P52961 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 335 | ENSG00000130055/Q9HCC8 | 0 | **2** | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 |
| 336 | ENSG00000130489/O43819 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |

## . A - PREVISÃO DA LOCALIZAÇÃO SUBCELULAR DAS PROTEÍNAS ATRAVÉS DO SOFTWARE MEKA

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 337 | ENSG00000130517/Q9NXJ5 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 338 | ENSG00000130803/Q96PQ6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 339 | ENSG00000130821/P48029 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 340 | ENSG00000130948/P37058 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 341 | ENSG00000130958/Q76EJ3 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 342 | ENSG00000130997/Q7Z5Q5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 343 | ENSG00000131269/O75027 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 344 | ENSG00000131355/Q9BY15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 345 | ENSG00000131375/Q9Y6W3 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 |
| 346 | ENSG00000131471/Q16853 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 347 | ENSG00000131477/O60895 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 5 | 0 |
| 348 | ENSG00000131773/O75525 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 349 | ENSG00000131910/Q15466 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 350 | ENSG00000132005/P22670 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 351 | ENSG00000132164/P48066 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 352 | ENSG00000132329/O60894 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 353 | ENSG00000132423/Q9NZJ6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 354 | ENSG00000132535/P78352 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | **7** |
| 355 | ENSG00000132703/P02743 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 356 | ENSG00000132915/P16499 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 357 | ENSG00000133710/Q9NQ38 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 358 | ENSG00000134363/P19883 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 359 | ENSG00000134516/Q92608 | 0 | **0** | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 360 | ENSG00000134569/O75096 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 361 | ENSG00000134817/P35414 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 362 | ENSG00000134882/Q8NBM4 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 363 | ENSG00000135341/O43318 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 364 | ENSG00000135677/P15586 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 365 | ENSG00000135776/Q9NRK6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 366 | ENSG00000136444/Q9HA92 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 367 | ENSG00000136451/Q14119 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 368 | ENSG00000136824/O95347 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 369 | ENSG00000136856/Q9NY64 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 370 | ENSG00000136868/O15431 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 371 | ENSG00000137338/Q96JS3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 372 | ENSG00000137409/Q9NZJ7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 373 | ENSG00000137563/Q92820 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 |
| 374 | ENSG00000137573/Q8IWU6 | 0 | 0 | 0 | **5** | 0 | 4 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 375 | ENSG00000137601/Q96PY6 | **3** | **1** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 376 | ENSG00000137713/P30154 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 377 | ENSG00000137752/P29466 | 0 | **6** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 378 | ENSG00000137776/Q9NWH9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 379 | ENSG00000137825/P23677 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 380 | ENSG00000137845/O14672 | 0 | 0 | 0 | 1 | 0 | 0 | **4** | 0 | 0 | 0 | 2 | 0 | **3** | 0 |
| 381 | ENSG00000137871/Q6N043 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 382 | ENSG00000137975/Q9UQC9 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 |
| 383 | ENSG00000137976/Q8WZ79 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 |
| 384 | ENSG00000137992/P11182 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 385 | ENSG00000138028/Q99674 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| 386 | ENSG00000138074/Q9Y289 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 387 | ENSG00000138075/Q9H222 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 388 | ENSG00000138095/P42704 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 389 | ENSG00000138135/O95992 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 390 | ENSG00000138271/Q9BY21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 391 | ENSG00000138346/P51530 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | **6** | 0 | 0 | 0 |
| 392 | ENSG00000138395/Q96Q40 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 393 | ENSG00000138668/Q14103 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 394 | ENSG00000138796/Q16836 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 395 | ENSG00000138942/Q96GF1 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 |
| 396 | ENSG00000139352/P50553 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 397 | ENSG00000139880/Q86UP0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 398 | ENSG00000140015/Q8NCM2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 399 | ENSG00000140400/Q9NTJ4 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 400 | ENSG00000140470/Q8TE56 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 401 | ENSG00000140479/P29122 | 0 | 0 | 0 | **3** | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 402 | ENSG00000140548/Q8N1W2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 403 | ENSG00000140835/Q8NCG5 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 404 | ENSG00000141562/Q9UHQ1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 405 | ENSG00000141639/P31152 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 406 | ENSG00000141644/Q9UIS9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 407 | ENSG00000141956/P57071 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 408 | ENSG00000142166/P17181 | 1 | 0 | 0 | 0 | **0** | 0 | 0 | **4** | 0 | 0 | 0 | 0 | **3** | 0 |
| 409 | ENSG00000142186/Q96KG9 | **3** | **7** | **0** | 0 | 0 | 0 | **1** | 0 | 0 | 0 | **2** | 0 | 0 | 0 |
| 410 | ENSG00000142507/P28072 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 411 | ENSG00000142619/Q9ULW8 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 412 | ENSG00000142623/Q9ULC6 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 413 | ENSG00000142789/P09093 | 0 | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 414 | ENSG00000142973/P13584 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 |
| 415 | ENSG00000143067/Q5TEC3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 1 | 0 |
| 416 | ENSG00000143106/P28066 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 417 | ENSG00000143147/Q8N6U8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 418 | ENSG00000143156/Q9Y5B8 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 419 | ENSG00000143179/Q9BZX2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 420 | ENSG00000143196/Q07507 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 421 | ENSG00000143217/Q96NY8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 422 | ENSG00000143226/P12318 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 423 | ENSG00000143248/O15539 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 424 | ENSG00000143469/Q8NB59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 425 | ENSG00000143515/P98198 | 0 | 0 | 0 | **4** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | **4** | 0 |
| 426 | ENSG00000143627/P30613 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 427 | ENSG00000143933/P62158 | 2 | **4** | **0** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 428 | ENSG00000144028/O75643 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 429 | ENSG00000144481/Q7Z2W7 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 430 | ENSG00000144792/Q6AZW8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 431 | ENSG00000144843/P54922 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 432 | ENSG00000144935/P48995 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 433 | ENSG00000145244/Q9Y5Q5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 434 | ENSG00000145246/Q9P241 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 435 | ENSG00000145283/Q3KNW5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 436 | ENSG00000145632/Q9NYY3 | **2** | **3** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 437 | ENSG00000145730/P19021 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 438 | ENSG00000145808/Q8TE59 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 439 | ENSG00000145863/Q16445 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **7** |
| 440 | ENSG00000145888/P23415 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 441 | ENSG00000145936/Q16558 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 442 | ENSG00000146039/Q9Y2C5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 443 | ENSG00000146828/Q9BXP2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 444 | ENSG00000147041/Q8TDW5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 445 | ENSG00000147099/Q9BY41 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 446 | ENSG00000147432/Q05901 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **7** |
| 447 | ENSG00000147576/Q8IWW8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 448 | ENSG00000147613/Q96QS6 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 449 | ENSG00000148339/Q6KCM7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 450 | ENSG00000148358/Q5VW38 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 451 | ENSG00000148516/P37275 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 452 | ENSG00000148600/Q96JP9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 453 | ENSG00000148832/Q6QHF9 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 |
| 454 | ENSG00000148834/P78417 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 455 | ENSG00000149295/P14416 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 456 | ENSG00000149968/P08254 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 457 | ENSG00000150054/Q5T2T1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 458 | ENSG00000150471/Q9HAR2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 459 | ENSG00000150687/O95084 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| 460 | ENSG00000151005/Q9H0I9 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 461 | ENSG00000151062/Q7Z3S7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 462 | ENSG00000151067/Q13936 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 463 | ENSG00000151079/P17658 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 464 | ENSG00000151208/Q8TDM6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 465 | ENSG00000151422/P16591 | 0 | **4** | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 | **2** | 0 |
| 466 | ENSG00000151577/P35462 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 467 | ENSG00000151617/P25101 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 468 | ENSG00000152034/Q969V1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 469 | ENSG00000152332/Q8TAS1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 470 | ENSG00000152463/Q9NV23 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 471 | ENSG00000152782/Q8TE04 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 472 | ENSG00000152822/Q13255 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 473 | ENSG00000152944/Q13503 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 474 | ENSG00000153294/Q8IZF3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 475 | ENSG00000153487/Q9UK53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 476 | ENSG00000153936/Q7LGA3 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 477 | ENSG00000154263/Q8WWZ4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 478 | ENSG00000154518/P48201 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 479 | ENSG00000154646/P98073 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 480 | ENSG00000154736/Q9UNA0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 481 | ENSG00000154767/Q01831 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 482 | ENSG00000154783/Q6ZNL6 | 0 | **1** | **4** | **3** | **1** | 0 | **4** | 0 | 0 | 0 | 0 | 0 | **0** | 0 |
| 483 | ENSG00000154845/Q8TF05 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 484 | ENSG00000154889/Q53F39 | 0 | 1 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 485 | ENSG00000155506/Q6PKG0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 486 | ENSG00000156006/P11245 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 487 | ENSG00000156096/P06133 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 1 | 0 |
| 488 | ENSG00000156222/O00337 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 489 | ENSG00000156453/Q08174 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 490 | ENSG00000156642/Q9Y639 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 491 | ENSG00000156735/O95429 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 492 | ENSG00000156925/O60481 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 493 | ENSG00000156958/Q01415 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 494 | ENSG00000156983/P55201 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 495 | ENSG00000157110/Q93062 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 496 | ENSG00000157500/Q9UKG1 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 497 | ENSG00000157800/Q8NCC5 | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 498 | ENSG00000158014/Q9BRI3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 4 | 0 |
| 499 | ENSG00000158445/Q14721 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **4** |
| 500 | ENSG00000158516/P48052 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 501 | ENSG00000158604/Q7Z7H5 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 502 | ENSG00000158691/O43309 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 503 | ENSG00000158828/Q9BXM7 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 504 | ENSG00000158864/O75306 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 505 | ENSG00000159199/P05496 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 506 | ENSG00000160211/P11413 | 1 | 5 | 0 | 2 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| 507 | ENSG00000160307/P04271 | 0 | **6** | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 508 | ENSG00000160326/Q9UGQ3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 509 | ENSG00000160447/Q6P5Z2 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 510 | ENSG00000160683/P32302 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 511 | ENSG00000160867/P22455 | 0 | **2** | 0 | **3** | **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 512 | ENSG00000160961/Q96JL9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 513 | ENSG00000161031/Q96PD5 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 514 | ENSG00000161405/Q9UKT9 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 515 | ENSG00000161509/Q14957 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **7** |
| 516 | ENSG00000161905/P16050 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| 517 | ENSG00000161921/Q9H2A7 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 518 | ENSG00000162390/Q8WXI4 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 519 | ENSG00000162409/P54646 | 0 | **3** | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 |
| 520 | ENSG00000162444/Q96R05 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 521 | ENSG00000162461/Q6PIV7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 522 | ENSG00000162551/P05186 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 523 | ENSG00000162694/Q9UBQ6 | 0 | 0 | 0 | **4** | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 524 | ENSG00000162892/Q13007 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 525 | ENSG00000162909/P17655 | 0 | **5** | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | **1** | 0 |
| 526 | ENSG00000163378/Q5NDL2 | 0 | 0 | 0 | **4** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 527 | ENSG00000163394/P32238 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 528 | ENSG00000163430/Q12841 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 529 | ENSG00000163545/Q9H093 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 |
| 530 | ENSG00000163581/P11168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 531 | ENSG00000163586/P07148 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| 532 | ENSG00000163624/Q92903 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 533 | ENSG00000163882/P52434 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 534 | ENSG00000163904/Q9HC62 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 |
| 535 | ENSG00000163932/Q05655 | 0 | **5** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | **1** | 0 |
| 536 | ENSG00000164078/Q04912 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 537 | ENSG00000164089/Q8TBG4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 538 | ENSG00000164199/Q8WXG9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 539 | ENSG00000164219/P53609 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 540 | ENSG00000164344/P03952 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 541 | ENSG00000164651/Q8IXZ3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 542 | ENSG00000164715/Q8IWU2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 543 | ENSG00000164733/P07858 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | **6** | 0 | 0 | 1 | 0 | 0 | 0 |
| 544 | ENSG00000164867/P29474 | 0 | **1** | **5** | 0 | 0 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | **4** | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 545 | ENSG00000164879/P07451 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 546 | ENSG00000165030/Q16649 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 547 | ENSG00000165078/Q8N4T0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 548 | ENSG00000165140/P09467 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 549 | ENSG00000165194/Q8TAB3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 550 | ENSG00000165449/Q7RTY1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 551 | ENSG00000165458/O15357 | 0 | **2** | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 552 | ENSG00000165671/Q96L73 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 553 | ENSG00000165731/P07949 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 554 | ENSG00000165752/Q86UX6 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 555 | ENSG00000166006/Q96PR1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** |
| 556 | ENSG00000166261/O95125 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 557 | ENSG00000166311/P17405 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 |
| 558 | ENSG00000166340/O14773 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 559 | ENSG00000166347/P00167 | 0 | **2** | 0 | **5** | 0 | 0 | 0 | 0 | **5** | 2 | 0 | 0 | 0 | 0 |
| 560 | ENSG00000166446/Q8N8U2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 561 | ENSG00000166482/P55083 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 562 | ENSG00000166484/Q13164 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 563 | ENSG00000166526/P17036 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 564 | ENSG00000166548/O00142 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 565 | ENSG00000166704/Q8WXB4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 566 | ENSG00000166747/O43747 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 567 | ENSG00000166796/P07864 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 568 | ENSG00000166930/Q9H3V2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 569 | ENSG00000166959/Q9BY19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 570 | ENSG00000167363/Q9H479 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 571 | ENSG00000167434/P22748 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 572 | ENSG00000167491/Q86YP4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 573 | ENSG00000167600/Q96SQ9 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 |
| 574 | ENSG00000167685/Q8N0Y2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 575 | ENSG00000167700/Q96ES6 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 576 | ENSG00000167754/Q9Y337 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 577 | ENSG00000167755/Q92876 | 0 | **5** | 0 | **0** | 0 | 2 | 0 | 0 | **0** | **0** | **1** | 0 | 0 | 0 |
| 578 | ENSG00000167769/Q8TDN7 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 579 | ENSG00000167791/Q9NPB3 | 0 | **0** | 1 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 580 | ENSG00000167967/Q66K89 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 581 | ENSG00000168038/Q96C45 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 582 | ENSG00000168065/Q9NSA0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 583 | ENSG00000168412/P48039 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 584 | ENSG00000168615/Q13443 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 585 | ENSG00000169118/Q9HCP0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 586 | ENSG00000169302/Q8WU08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 587 | ENSG00000169372/P78560 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 588 | ENSG00000169427/Q9NPC2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 589 | ENSG00000169432/Q15858 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 590 | ENSG00000169504/Q9Y696 | **0** | **4** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | **3** | **2** | 0 | **0** | 0 |
| 591 | ENSG00000169692/O15120 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 592 | ENSG00000169733/Q9Y644 | 0 | 0 | 0 | 0 | 0 | 3 | **4** | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 593 | ENSG00000169814/P43251 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 594 | ENSG00000169826/Q8N6G5 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 595 | ENSG00000169877/Q9NZD4 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 596 | ENSG00000169885/Q8TD86 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 597 | ENSG00000170049/O43448 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 598 | ENSG00000170255/Q96LB2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 599 | ENSG00000170374/Q8TDD2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | 0 | 0 | 0 |
| 600 | ENSG00000170448/Q6ZNB6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 3 | 0 |
| 601 | ENSG00000170899/O15217 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 602 | ENSG00000170955/Q969G5 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 603 | ENSG00000171016/Q9Y3Y4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 604 | ENSG00000171094/Q9UM73 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 605 | ENSG00000171105/P06213 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 606 | ENSG00000171314/P18669 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 607 | ENSG00000171320/Q56NI9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 608 | ENSG00000171435/Q6VAB6 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 609 | ENSG00000171469/Q8N587 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 610 | ENSG00000172115/P99999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 611 | ENSG00000172183/Q96AZ6 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 612 | ENSG00000172380/Q9UBI6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 613 | ENSG00000172404/Q7Z6W7 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 |
| 614 | ENSG00000172466/P17028 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 615 | ENSG00000172497/Q8WYK0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 616 | ENSG00000172680/P00540 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 617 | ENSG00000172818/O14753 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 618 | ENSG00000172878/Q6UB28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 619 | ENSG00000172935/Q96AM1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 620 | ENSG00000172939/O95747 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 621 | ENSG00000173137/Q3MIX3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 622 | ENSG00000173198/Q9Y271 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 623 | ENSG00000173208/Q9UBJ2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 |
| 624 | ENSG00000173391/P78380 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 625 | ENSG00000173432/P0DJI8 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 626 | ENSG00000173531/P26927 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 627 | ENSG00000173535/O14798 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 628 | ENSG00000173894/Q14781 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 629 | ENSG00000173917/P14652 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 630 | ENSG00000174243/Q9BUQ8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 631 | ENSG00000174255/P51504 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 632 | ENSG00000174332/Q8NBF1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 633 | ENSG00000174437/P16615 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 634 | ENSG00000174684/O43505 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 635 | ENSG00000174938/Q6UXD5 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 636 | ENSG00000174990/P35218 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 637 | ENSG00000175336/Q13790 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 638 | ENSG00000175514/Q8TDT2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 639 | ENSG00000175538/Q9Y6H6 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 640 | ENSG00000175564/P55916 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 641 | ENSG00000175567/P55851 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 642 | ENSG00000175697/Q8NFN8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 643 | ENSG00000176083/Q8IZ20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 644 | ENSG00000176393/Q9H4A4 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 645 | ENSG00000176402/Q8NFK1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 646 | ENSG00000177302/Q13472 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| 647 | ENSG00000177464/P46093 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 648 | ENSG00000177613/Q9H0L4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 649 | ENSG00000177628/P04062 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 650 | ENSG00000177879/Q92572 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 651 | ENSG00000177888/Q5SVQ8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 652 | ENSG00000178015/Q8NGU9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 653 | ENSG00000178172/Q6UWN8 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 654 | ENSG00000178394/P08908 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 655 | ENSG00000179097/P30939 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 656 | ENSG00000179546/P28221 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 657 | ENSG00000179603/O00222 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 658 | ENSG00000179750/Q9UH17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 659 | ENSG00000179761/Q9P0Z9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 |
| 660 | ENSG00000179930/Q5T619 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 661 | ENSG00000180370/Q13177 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 |
| 662 | ENSG00000180532/Q8NAM6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 663 | ENSG00000180616/P30874 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 664 | ENSG00000180914/P30559 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 665 | ENSG00000181085/Q8TD08 | 0 | 2 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 666 | ENSG00000181619/Q8IZ08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 667 | ENSG00000181666/P10072 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 668 | ENSG00000181896/Q8IZC7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 669 | ENSG00000182054/P48735 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 670 | ENSG00000182134/Q9Y2W6 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 | 0 |
| 671 | ENSG00000182256/Q99928 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **6** |
| 672 | ENSG00000182473/Q9UPT5 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 673 | ENSG00000182541/P53671 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 | 0 | 0 |
| 674 | ENSG00000182580/P54753 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 675 | ENSG00000182631/Q9NSD7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 676 | ENSG00000183309/O75123 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 677 | ENSG00000183542/O43908 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 678 | ENSG00000183621/Q7Z4V0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 679 | ENSG00000183729/P48145 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 680 | ENSG00000183734/Q99929 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 681 | ENSG00000183778/Q9Y2C3 | 0 | 0 | 0 | 1 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 682 | ENSG00000183856/Q86VI3 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 683 | ENSG00000183862/Q16280 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 |
| 684 | ENSG00000183914/Q9P225 | 0 | **3** | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 685 | ENSG00000184012/O15393 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 686 | ENSG00000184076/Q9UDW1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 687 | ENSG00000184292/P09758 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 4 | 0 |
| 688 | ENSG00000185010/P00451 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 689 | ENSG00000185219/P59923 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 690 | ENSG00000185252/Q16587 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 691 | ENSG00000185760/Q9NR82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 692 | ENSG00000185883/P27449 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 693 | ENSG00000185897/O14843 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 694 | ENSG00000186009/P51164 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 695 | ENSG00000186184/Q9Y2S0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 696 | ENSG00000186204/Q9HCS2 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | 0 |
| 697 | ENSG00000186474/Q9UKR0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 698 | ENSG00000186517/Q7Z6I6 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 699 | ENSG00000186918/Q9H8N7 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 700 | ENSG00000187210/Q02742 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 701 | ENSG00000187258/Q6W5P4 | 0 | **6** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| 702 | ENSG00000187527/Q4VNC0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 703 | ENSG00000187555/Q93009 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 704 | ENSG00000187714/Q16572 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 705 | ENSG00000188290/Q9HCC6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 706 | ENSG00000188342/P13984 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 707 | ENSG00000188386/Q96LZ3 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 708 | ENSG00000188655/P60153 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 709 | ENSG00000188778/P13945 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 710 | ENSG00000188822/P34972 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 711 | ENSG00000188992/Q6XZB0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 712 | ENSG00000196090/O14522 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 713 | ENSG00000196131/Q8NFZ6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 714 | ENSG00000196154/P26447 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| 715 | ENSG00000196371/P22083 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 716 | ENSG00000196588/Q969V6 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 717 | ENSG00000196639/P35367 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 718 | ENSG00000196767/P49335 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 719 | ENSG00000196811/P07510 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | **7** |
| 720 | ENSG00000197106/Q9H1V8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | **0** |
| 721 | ENSG00000197140/Q8TC27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 722 | ENSG00000197324/Q7Z4F1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 723 | ENSG00000197417/Q9UHJ6 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 724 | ENSG00000197566/Q9P2J8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 725 | ENSG00000197580/Q9BYV7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 726 | ENSG00000197714/Q14592 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 727 | ENSG00000197888/O75795 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 1 | 0 |
| 728 | ENSG00000197891/Q96S37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 729 | ENSG00000197961/P58317 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 730 | ENSG00000198077/P20853 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 | 1 | 0 |
| 731 | ENSG00000198160/Q8N108 | 0 | **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 | 0 | 0 |
| 732 | ENSG00000198178/Q8WTT0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 733 | ENSG00000198189/Q8NBQ5 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 734 | ENSG00000198205/P98168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 735 | ENSG00000198650/P17735 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 |
| 736 | ENSG00000198682/O95340 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 737 | ENSG00000198753/Q9ULL4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 738 | ENSG00000198821/P20963 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 739 | ENSG00000198825/Q9Y2H2 | 0 | 3 | 0 | 0 | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 740 | ENSG00000198829/Q9BXA5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 741 | ENSG00000198830/P05204 | 0 | **5** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 742 | ENSG00000198835/Q5T442 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 743 | ENSG00000198929/O75052 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 744 | ENSG00000203837/Q17RR3 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 745 | ENSG00000204099/Q8WWR8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 |
| 746 | ENSG00000204195/Q58HT5 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 747 | ENSG00000204356/P18615 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 748 | ENSG00000204370/O14521 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 749 | ENSG00000204390/P34931 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 |
| 750 | ENSG00000204472/P55008 | 0 | **4** | **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | **0** | 0 |
| 751 | ENSG00000204560/O60231 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 752 | ENSG00000204574/Q8NE71 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 753 | ENSG00000204961/Q9Y5H5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 754 | ENSG00000204983/P07477 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 755 | ENSG00000205143/A6NKF2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 756 | ENSG00000205213/Q9BXB1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 757 | ENSG00000205581/P05114 | 0 | **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 758 | ENSG00000206190/O60312 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 |
| 759 | ENSG00000206503/P04439 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 760 | ENSG00000206503/P13746 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 761 | ENSG00000206503/P16188 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| 762 | ENSG00000213339/Q9BXR0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** | 0 | 0 | 0 | 0 |
| 763 | ENSG00000213347/Q9BW11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 764 | ENSG00000213398/P04180 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 765 | ENSG00000213445/Q96FS4 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 766 | ENSG00000213619/O75489 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 |
| 767 | ENSG00000213780/Q92759 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 768 | ENSG00000213927/Q9Y4X3 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 769 | ENSG00000214022/Q9BWE0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 770 | ENSG00000215644/P47871 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** | 0 |
| 771 | ENSG00000221914/P63151 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 772 | ENSG00000231925/O15533 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 773 | ENSG00000232810/P01375 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **5** | 0 |
| 774 | ENSG00000233276/P07203 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 775 | ENSG00000240857/Q9HBH5 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| 776 | ENSG00000241119/O60656 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 |
| 777 | ENSG00000243279/O60831 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 778 | ENSG00000243477/Q93015 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 779 | ENSG00000243480/P04746 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 780 | ENSG00000244405/P41161 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 781 | ENSG00000244474/P22310 | 0 | 0 | 0 | **7** | 0 | 0 | 0 | 0 | **2** | 0 | 0 | 0 | 0 | 0 |
| 782 | ENSG00000245680/Q52M93 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 783 | ENSG00000248383/Q9H158 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 784 | ENSG00000251369/Q7Z398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 785 | ENSG00000251664/Q9UN75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 786 | ENSG00000253873/Q9Y5H2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 787 | ENSG00000253953/Q9UN71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 788 | ENSG00000254986/Q9NY33 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 789 | ENSG00000257017/P00738 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 790 | ENSG00000258818/P34096 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 791 | ENSG00000261701/P00739 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 792 | ENSG00000263002/Q14588 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 793 | ENSG00000265107/P36382 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 794 | ENSG00000269437/Q9GZY0 | 0 | **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 795 | ENSG00000272674/Q9NRJ7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 796 | ENSG00000273513/A0A087X1G2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 |
| 797 | ENSG00000275111/Q9BSG1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 798 | ENSG00000276644/Q9UI36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |
| 799 | ENSG00000278129/P17098 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **7** | 0 | 0 | 0 |