Pós-Graduação em Ciência da Computação

"COMBINAÇÃO DE KERNELS PARA PREDIÇÃO DE INTERAÇÕES EM REDES BIOLÓGICAS"

Por

André Câmara Alves do Nascimento

Tese de Doutorado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE/2015

Universidade Federal de Pernambuco
Centro de Informática
Pós-graduação em Ciência da Computação

André Câmara Alves do Nascimento

COMBINAÇÃO DE KERNELS PARA PREDIÇÃO DE INTERAÇÕES EM REDES BIOLÓGICAS

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Orientador: *Ricardo Bastos Cavalcante Prudêncio*
Co-Orientador: *Ivan Gesteira Costa Filho*

RECIFE
2015

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

N244c Nascimento, André Câmara Alves do
Combinação de kernels para predição de interações em redes biológicas / André Câmara Alves do Nascimento – Recife: O Autor, 2015.
109 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2015.
Inclui referências e apêndice.

1. Inteligência artificial. 2. Aprendizado de máquina. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2015-182

Tese de Doutorado apresentada por André Câmara Alves do Nascimento à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Combinação de Kernels para Predição de Interações em Redes Biológica**"' orientada pelo **Prof. Ricardo Bastos Cavalcante Prudêncio** e aprovada pela Banca Examinadora formada pelos professores:

Prof. Francisco de Assis Tenório de Carvalho
Centro de Informática/UFPE

Prof. Cleber Zanchettin
Centro de Informática / UFPE

Profa. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE

Profa. Ana Carolina Lorena
Instituto de Ciência e Tecnologia / UNIFESP

Prof. José Luiz de Lima Filho
Departamento de Bioquímica e Biofísica / UFPE

Visto e permitida a impressão.
Recife, 9 de novembro de 2015.

Profa. Edna Natividade da Silva Barros
Coordenadora da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Dedico esta tese aos meus pais,
Roosevelt e Cândida.

Agradecimentos

Eu não poderia concluir este trabalho sem a ajuda e suporte de muitas pessoas, e acima de tudo de Deus, que tão bem guardou a mim e a minha família durante todo este tempo.

Primeiramente, agradeço aos professores Ricardo Prudêncio e Ivan Costa, que com paciência e segurança têm me direcionado por tantos anos de pesquisa. Agradeço ainda pela oportunidade por eles oferecida para realização do estágio doutoral em outro país, que me enriqueceu de experiências que levarei por toda a minha vida acadêmica. Agradeço à CAPES pelo suporte financeiro, e ao IZKF Aachen, por disponilibilizar os recursos computacionais sem os quais não teria sido possível a realização dos experimentos tão importantes para a conclusão deste trabalho.

Aos meus amigos no Cin/UFPE, IFPE, IZKF e DEINFO pelas conversas sobre diversos aspectos da pesquisa, e pelos ótimos momentos de descontração. Ao grande amigo Frederico Duarte, por me apresentar a Biologia e Química Computacional, e pela orientação e apoio na difícil fase inicial do doutorado.

E finalmente, nada disso seria possível sem o suporte da minha família. Sou especialmente grato à minha esposa Izabel, que me apoiou desde o início, e cuidou de tudo sozinha durante o período que fiquei fora; às minhas filhas Clara e Marina por toda alegria que me trazem; à Gizel e Gabriella pelas divertidas interrupções matinais; e a minha eterna companheira e irmã Amélia. Não poderia deixar de mencionar minha gratidão à minha mãe, Cândida, por desenvolver em mim o gosto pela ciência e pelo conhecimento, e ao meu pai, Roosevelt, pelo apoio incondicional em todas as decisões que tomei em minha vida.

Resumo

Redes droga-proteína têm recebido bastante atenção nos últimos anos, dada sua relevância para a inovação farmacêutica e produção de novos fármacos. Muitas abordagens *in silico* distintas para predição de interações droga-proteína têm sido propostas, muitas das quais baseadas em uma classe particular de métodos de aprendizagem de máquina chamada de métodos de kernel. Estes algoritmos de classificação de padrões são capazes de incorporar conhecimento prévio na forma de funções de similaridade, i.e., um kernel, e têm tido sucesso em diversos problemas de aprendizagem supervisionada. A seleção da função de kernel adequada e seus respectivos parâmetros pode ter grande influência no desempenho do classificador construído. Recentemente, a aprendizagem de múltiplos kernels (*Multiple Kernel Learning* - MKL) tem sido introduzida para solucionar este problema, permitindo a utilização de múltiplos kernels, ao invés de considerar apenas um kernel para uma dada tarefa. A principal motivação para tal abordagem é similar a considerada na combinação de múltiplos classificadores: ao invés de restringir-se a um único kernel, é preferível utilizar um conjunto de kernels distintos, e deixar que um algoritmo selecione os melhores, ou sua respectiva combinação. Abordagens MKL também podem ser vistas como uma estratégia de integração de dados. Apesar dos avanços técnicos nos últimos anos, as abordagens propostas anteriormente não são capazes de lidar com os grandes espaços de interação entre drogas e proteínas e integrar múltiplas fontes de informação simultaneamente. Neste trabalho, é proposto um método de aprendizagem de múltiplos kernels para a combinação não esparsa de kernels na predição de interações em redes droga-proteína. O método proposto permite a integração de múltiplas fontes heterogêneas de informação para a identificação de novas interações, e também pode ser aplicado em redes de tamanhos arbitrários. Além disso, o método proposto pode também selecionar automaticamente os kernels mais relevantes, retornando pesos que indiquem a sua importância na predição de interações droga-proteína na rede em questão. A análise empírica em quatro bases de dados, utilizando vinte kernels distintos indicou que o método proposto obteve desempenho comparável ou superior a todos os métodos avaliados. Ademais, os pesos associados aos kernels analisados refletiram a qualidade preditiva obtida por cada kernel em experimentos exaustivos para cada par de kernels, um indicativo do sucesso do método em identificar automaticamente fontes de informação biológica relevantes. Nossas análises demonstraram que a estratégia de integração de dados é capaz de melhorar a qualidade das interações preditas, e pode acelerar a identificação de novas interações, bem como identificar informações relevantes para a tarefa.

Palavras-chave: predição de links, aprendizagem de múltiplos kernels, redes bipartidas, métodos de kernel

Abstract

Drug-target networks are receiving a lot of attention in late years, given its relevance for pharmaceutical innovation and drug lead discovery. Many different in silico approaches for the identification of new drug-target interactions have been proposed, many of them based on a particular class of machine learning algorithms called kernel methods. These pattern classification algorithms are able to incorporate previous knowledge in the form of similarity functions, i.e., a kernel, and it has been successful in a wide range of supervised learning problems. The selection of the right kernel function and its respective parameters can have a large influence on the performance of the classifier. Recently, Multiple Kernel Learning algorithms have been introduced to address this problem, enabling one to use multiple kernels instead of a single one for a given task. The main motivation for such approach is similar to the one considered in ensemble methods: instead of being restricted to only one kernel, it is preferrable to use a set of distinct kernels, and let the algorithm choose the best ones, or its combination. The MKL approach can also be seen as a data integration strategy. Despite technical advances in the latest years, previous approaches are not able to cope with large drug-target interaction spaces and integrate multiple sources of biological information simultaneously. In this work, we propose a new multiple kernel learning algorithm for the non-sparse combination of kernels in bipartite link prediction on drug-target networks. This method allows the integration of multiple heterogeneous information sources for the identification of new interactions, and can also work with networks of arbitrary size. Moreover, our method can also automatically select the more relevant kernels, returning weights indicating their importance in the drug-target prediction at hand. Empirical analysis on four data sets, using twenty distinct kernels indicates that our method has higher or comparable predictive performance than all evaluated methods. Moreover, the predicted weights reflect the predictive quality of each kernel on exhaustive pairwise experiments, which indicates the success of the method to automatically indicate relevant biological information sources. Our analysis show that the proposed data integration strategy is able to improve the quality of the predicted interactions, and can speed up the identification of new drug-target interactions as well as identify relevant information for the task.

Keywords: link prediction, multiple kernel learning, bipartite networks, kernel methods

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AERS	*Adverse Event Reporting System*
ATC	*Anatomical Therapeutic Chemical*
CMAP	*Connectivity Map*
FDA	*U.S. Food and Drug Administration*
GEO	*Gene Expression Omnibus*
GIP	*Gaussian Interaction Profile*
GPCR	*G-protein coupled receptor*
GO	*Gene Ontology*
HTS	*High Throughput Screening*
JAPIC	*Japan Pharmaceutical Information Center*
KFC	*Kernel Fischer Criterion*
KKT	Karush-Kuhn-Tucker
KRM	*Kernel Regression Model*
MKL	*Multiple Kernel Learning*
OMS	Organização Mundial da Saúde
PKM	*Pairwise Kernel Method*
PPI	*Protein-protein interaction*
PSD	Positiva semidefinida
QSAR	*Quantitative Structure Activity Relationship*
RKHS	*Reproducing kernel Hilbert space*
RLS	*Regularized least squares*
SVM	*Support Vector Machines*

Sumário

1

Introdução

O mundo real é repleto de relacionamentos entre as mais diversas entidades. Redes são uma forma flexível e conveniente de representar as relações entre elementos, as quais estão presentes em tudo: pesquisa, comércio, biologia e nas próprias relações de amizade entre pessoas. O estudo das redes em geral se utiliza de estruturas matemáticas chamadas grafos, um formalismo robusto o suficiente para construção de modelos que possam ser estudados sistematicamente. Um grafo consiste em um conjunto de nós, chamados de vértices, e das ligações entre os nós, chamadas de arestas. Exemplos de relações do mundo real que são modeladas como grafos incluem redes sociais, redes de estradas, comunicação, *hyperlinks*, interações químicas, colaboração, dentre outras. É comum que em algumas redes, além da sua própria estrutura de ligações, também existam estruturas adicionais, como por exemplo, informações descritivas acerca dos nós individuais ou características associadas às suas arestas.

Os processos biológicos não são exceção. A maior parte dos componentes celulares exercem suas funções através de intensas interações com outras entidades biológicas, e até mesmo entre tecidos e órgãos diferentes em um organismo (BARABáSI; GULBAHCE; LOSCALZO, 2011). A rede de interações resultante destes processos nos humanos se apresenta como um grande desafio para as ciências biomédicas, uma vez que o interactoma, como é chamado, possui em torno de 25.000 genes codificantes, combinados com algo em torno de 1.000 metabólitos e outros tipos de moléculas, chegando a ter em torno de 100.000 nós (BARABáSI; GULBAHCE; LOSCALZO, 2011).

Entender a dinâmica destas redes de interação representa um grande desafio, dada a heterogeneidade e a alta dimensionalidade dos dados, bem como o nível de complexidade envolvido em tais interações. Entretanto, os benefícios decorrentes de modelos que simulem as relações no interactoma são imensos, especialmente no tratamento de doenças. A utilização de métodos computacionais tem se mostrado fundamental neste sentido, uma vez que viabiliza a simulação e predição *in silico* de interações até então desconhecidas, possibilitando que as análises experimentais sejam concentradas nas interações com maior probabilidade de ocorrência.

A predição de novas interações biológicas é na verdade uma instância de um problema mais genérico, a predição de *links* em redes complexas. Esta é uma área de pesquisa bastante

ativa na Ciência da Computação, que busca estimar a probabilidade de existência de uma ligação (i.e. *link*) entre dois nós em uma rede, baseada nas interações observadas e nos atributos dos nós (Lü; ZHOU, 2011). Uma vasta gama de problemas têm sido abordados como tarefas de predição de *links* nas mais diversas áreas: em marketing, onde ações de cliques de usuários podem ser preditos a partir do histórico das suas atividades; na análise de redes sociais, de modo que usuários podem ser direcionados a conteúdos de interesse baseado em suas relações de amizade e suas preferências pessoais (RAYMOND; KASHIMA, 2010); ou ainda na identificação de prováveis interações entre proteínas em um organismo (BEN-HUR; NOBLE, 2005).

Dada a grande diversidade de contextos e problemas em que pode ser aplicada, a predição de interações tem chamado bastante atenção da comunidade acadêmica, com a evolução de sistemas computacionais capazes de extrapolar o reducionismo imposto pela limitação tecnológica, para um novo modelo sistêmico, no qual entidades não atuam isoladamente, mas como parte de uma complexa rede de interações.

Em geral, métodos de predição de *links* podem ser classificados em termos da informação utilizada para a predição em três categorias (RAYMOND; KASHIMA, 2010): métodos baseados em *links* (LIBEN-NOWELL; KLEINBERG, 2007; HUANG; LI; CHEN, 2005), métodos baseados em informações dos nós (BEN-HUR; NOBLE, 2005; RAYMOND; KASHIMA, 2010), ou ainda métodos híbridos, que utilizam uma combinação de ambos (LI; CHEN, 2013). Com relação à natureza da predição, podemos classificar os métodos em duas categorias (MENON; ELKAN, 2011): estrutural, na qual a entrada é o grafo parcialmente observado, e o que se deseja é predizer a existência ou não de arestas para os pares de nós desconectados; ou temporal, onde se é recebido como entrada uma série de grafos observados por completo em vários momentos distintos, e o objetivo é predizer o estado das arestas no próximo passo.

Um outro ponto importante para definir se um determinado método é aplicável ou não a uma rede em particular, é observar se o grafo subjacente é um grafo unipartido ou bipartido (KUNEGIS; LUCA; ALBAYRAK, 2010). Nos grafos unipartidos, não há restrições sobre as conexões entre os nós (arestas), enquanto nos grafos bipartidos os nós dividem-se em dois conjuntos distintos, e as arestas conectam apenas nós pertencentes a conjuntos diferentes, i.e., não existem ligações entre nós de um mesmo tipo. Redes de autoria e de interação droga-proteína são exemplos de redes que podem ser modeladas como grafos bipartidos.

Métodos de aprendizagem de máquina são uma classe de algoritmos que constroem modelos preditivos baseados em dados, atualmente utilizados nas mais diversas áreas do conhecimento humano (DUDA; HART; STORK, 2012). Tais métodos também têm sido aplicados com sucesso em diferentes contextualizações da predição de *links* (LIBEN-NOWELL; KLEINBERG, 2007; Lü; ZHOU, 2011), por exemplo, em redes sociais (LIBEN-NOWELL; KLEINBERG, 2007), sistemas de recomendação (LI; CHEN, 2013), e redes biológicas (BARABáSI; GULBAHCE; LOSCALZO, 2011). Nos últimos anos, uma classe particular de algoritmos de aprendizagem de máquina, os chamados métodos de kernel, tem se destacado na predição de interações farmacológicas (DING et al., 2013). Métodos de kernel produzem modelos preditivos incorporando

conhecimento prévio na forma de funções de similaridade, chamadas de kernels.

Neste trabalho, o problema da predição de interações em redes biológicas será abordado como um problema de predição de *links* de natureza estrutural, e a estratégia da solução proposta segue a linha híbrida (informações das interações combinadas com informações dos nós), sendo utilizados métodos de aprendizagem de máquina baseados em kernel para estabelecer um *ranking* das interações mais prováveis na rede.

Apesar da grande quantidade de métodos para predição de interações existentes na literatura, estes algoritmos em geral incorporam uma quantidade limitada de informações para a realização das predições. Recentemente, a introdução da aprendizagem de múltiplos kernels (*Multiple Kernel Learning* - MKL) (GöNEN; ALPAYDıN, 2011), propõe uma estratégia eficiente e escalável de incorporar diferentes visões sobre os dados em algoritmos de aprendizagem de máquina baseados em kernel, selecionando automaticamente as informações mais relevantes ao problema em questão. Por outro lado, apesar do recente interesse no desenvolvimento de abordagens capazes de integrar fontes de dados heterogêneas para predição de interações em redes complexas, a maior parte dos trabalhos anteriores utiliza combinações simples e pré-estabelecidas, e muitas vezes são incapazes de selecionar automaticamente as fontes de informação mais relevantes para o problema. Isto se dá principalmente porque técnicas de aprendizagem de múltiplos kernels normalmente não são diretamente aplicáveis ao problema da predição de links, especialmente quando a rede em questão pode ser modelada como um grafo bipartido.

A proposta deste trabalho é explorar as limitações dos trabalhos anteriores, com o desenvolvimento de um método de combinação de aprendizagem de kernels específico para o problema da predição de *links* em redes bipartidas. As redes estudadas serão de contexto biológico, dada a característica heterogênea das mesmas, a grande quantidade de visões diferentes que podem ser extraídas da rede e seus componentes, bem como a disponibilidade de dados abertos sobre estas entidades.

O restante deste capítulo introduz alguns dos temas relacionados a este trabalho, começando por uma breve descrição da predição de interações em redes biológicas na Seção 1.1, e uma introdução da motivação dos métodos de MKL na Seção 1.2. Em seguida, uma definição clara do objetivo do trabalho é apresentada na Seção 1.3. Por fim, a Seção 1.4 descreve como este documento está organizado.

1.1 Predição de interações biológicas

Um dos maiores desafios das ciências da saúde atualmente é o desenvolvimento de novos medicamentos (CSERMELY et al., 2013). Entretanto, apesar dos crescentes investimentos na área, o desenvolvimento rápido e de baixo custo de novos fármacos ainda está muito distante da realidade atual. Em média, são necessários de 12 a 15 anos, e até 1 bilhão de dólares para trazer uma nova droga para o mercado (BUTCHER, 2005; CSERMELY et al., 2013; EKINS

et al., 2011). Apesar da indústria farmacêutica ser uma das maiores investidoras em pesquisa e desenvolvimento no mundo, atualmente existem aproximadamente uma centena de alvos farmacológicos para drogas aprovadas, de um total de mais de 20.000 proteínas não redundantes no proteoma humano (CSERMELY et al., 2013).

Como já foi dito, a heterogeneidade de representações de entidades e processos biológicos impõe uma restrição reducionista às análises de tais dados, e em especial, à predição de interações em redes biológicas. Este é um problema de extrema relevância nos estudos de Biologia de Sistemas (JUNKER; SCHREIBER, 2008; DING et al., 2013). Nesse contexto, redes droga-proteína têm recebido bastante atenção nos últimos anos, dada sua relevância para a inovação farmacêutica e produção de novos fármacos. Este tipo de rede consiste em uma modelagem matemática na qual a existência de interação entre uma droga (composto químico) e uma proteína (alvo farmacológico) é representada através de uma aresta ligando estes nós em um grafo bipartido (Figura 1.1 A). Apesar do aumento da quantidade de dados sobre interações droga-proteína identificadas, ela é apenas uma pequena fração do total de prováveis interações, uma vez que existem em torno de 6.000–8.000 alvos de interesse farmacológico no genoma humano (CHEN; LIU; YAN, 2012). De fato, as dificuldades e custos decorrentes dos procedimentos experimentais de descoberta de tais interações biológicas limitam a descoberta de novas interações.

Uma grande variedade de abordagens computacionais tem sido desenvolvidas para analisar e predizer interações entre compostos químicos e proteínas em um organismo, sendo em geral categorizados em: métodos baseados em ligantes ou métodos baseados em *docking* (YAMANISHI, 2013). Métodos baseados em ligantes, como o QSAR, caracterizam-se por comparar um ligante candidato aos ligantes conhecidos de uma determinada proteína alvo a fim de predizer uma potencial ligação utilizando métodos de aprendizagem de máquina (BUTINA; SEGALL; FRANKCOMBE, 2002; DUDEK; ARODZ; GáLVEZ, 2006; YAMANISHI, 2013). Um problema com este tipo de abordagem, é que normalmente seu desempenho cai a medida que o número de ligantes conhecidos para uma dada proteína diminui (YAMANISHI, 2013).

A abordagem de *docking* é uma abordagem poderosa, e sua aplicação só é possível quando a estrutura tridimensional do alvo é conhecida, geralmente a partir de cristalografia de raios X (*x-ray crystallography*) ou predita por modelagem via homologia (JORGENSEN, 2004; CHEN; LIU; YAN, 2012; MOUSAVIAN; MASOUDI-NEJAD, 2014). Modelos tridimensionais de estruturas químicas são armazenadas em computadores, na posição ideal para o ponto de ligação, sobre a qual é calculado um *score* de atividade potencial. Os compostos com maior *score* podem então ser comprados ou sintetizados , e então submetidos a testes experimentais. A necessidade do conhecimento prévio da estrutura tridimensional da proteína para a realização de simulações de *docking* limita bastante o uso deste método em escala genômica (YAMANISHI, 2013).

Recentemente, com o crescimento de bases de dados públicas (ERTL; JELFS, 2007), um novo conjunto de métodos computacionais com foco na mineração da informação disponível têm sido desenvolvidos. Nesta abordagem, interações conhecidas, drogas e alvos são unificados

em um único espaço, chamado de espaço quimiogenômico, do qual são extraídas diversas informações. Das interações conhecidas é obtida uma matriz de interações Y, na qual uma interação entre uma droga d_i e uma proteína p_j é representada por $Y_{ij} = 1$, e $Y_{ij} = 0$ caso contrário. Em seguida, é aplicada uma medida de similaridade entre as drogas presentes no espaço quimiogenômico, i.e., cada droga é comparada com todas as outras drogas do conjunto. O mesmo é feito para proteínas (Figura 1.1 B). Estas informações são por fim combinadas para construir um modelo preditivo, que é utilizado para inferir novas interações na rede (Figura 1.1 C). A hipótese principal consiste na idéia de que ligantes parecidos tendem a ligar-se a proteínas similares. Dessa forma, o modelo preditivo pode incorporar características topológicas da rede de interações conhecidas, bem como descritores de cada tipo de entidade envolvida. Tais descritores podem ser calculados sobre as estruturas químicas dos compostos, efeitos colaterais, sequências de aminoácidos, perfis de expressão gênica, ou qualquer outra informação relevante para a tarefa em questão.

A construção de tal modelo preditivo pode ser realizada utilizando-se os chamados métodos de kernel, já citados anteriormente neste capítulo. Entretanto, a utilização deste tipo de algoritmo requer que seja definida uma medida de similaridade (kernel) entre as instâncias do problema, que neste caso, correspondem a *pares* de nós (ou simplesmente, arestas), uma vez que o que se deseja ao final do processo é predizer a probabilidade de ocorrência de uma dada aresta na rede. Este detalhe introduz uma série de aspectos que tornam a predição de links baseada em métodos de kernel um problema com características bastante peculiares. A principal limitação a ser superada é o fato de que a ordem de crescimento do consumo de memória durante a fase de treinamento de uma rede bipartida com $n+m$ nós (com duas matrizes de kernel de dimensões $n \times n$ e $m \times m$) pode chegar a $(nm)^2$, uma vez que a matriz de kernel de pares obtida terá dimensões $nm \times nm$ (Figura 1.2).

1.2 Aprendizagem de múltiplos kernels

Métodos de kernel são uma família de algoritmos para classificação de padrões, capazes de incorporar conhecimento prévio na forma de funções de similaridade, ou simplesmente, um kernel. Sua aplicação tem obtido sucesso em diversos problemas de aprendizagem supervisionada (SCHOLKOPF; SMOLA, 2001). Entretanto, a seleção da função de kernel e seus respectivos parâmetros é um aspecto de grande influência no desempenho obtido pelo classificador construído. Em geral, tais funções e parâmetros são selecionados através de um processo de validação cruzada sobre um conjunto de validação diferente do conjunto de treinamento. A aprendizagem de múltiplos kernels (MKL)(GöNEN; ALPAYDıN, 2011), é uma área emergente dentro da aprendizagem de máquina, cuja motivação é similar a considerada na combinação de múltiplos classificadores: ao invés de restringir-se ao uso de um único kernel, é preferível utilizar um conjunto de kernels distintos, e deixar que um algoritmo selecione os melhores, ou sua respectiva combinação. A vantagem em se utilizar kernels diferentes para um mesmo problema, é que

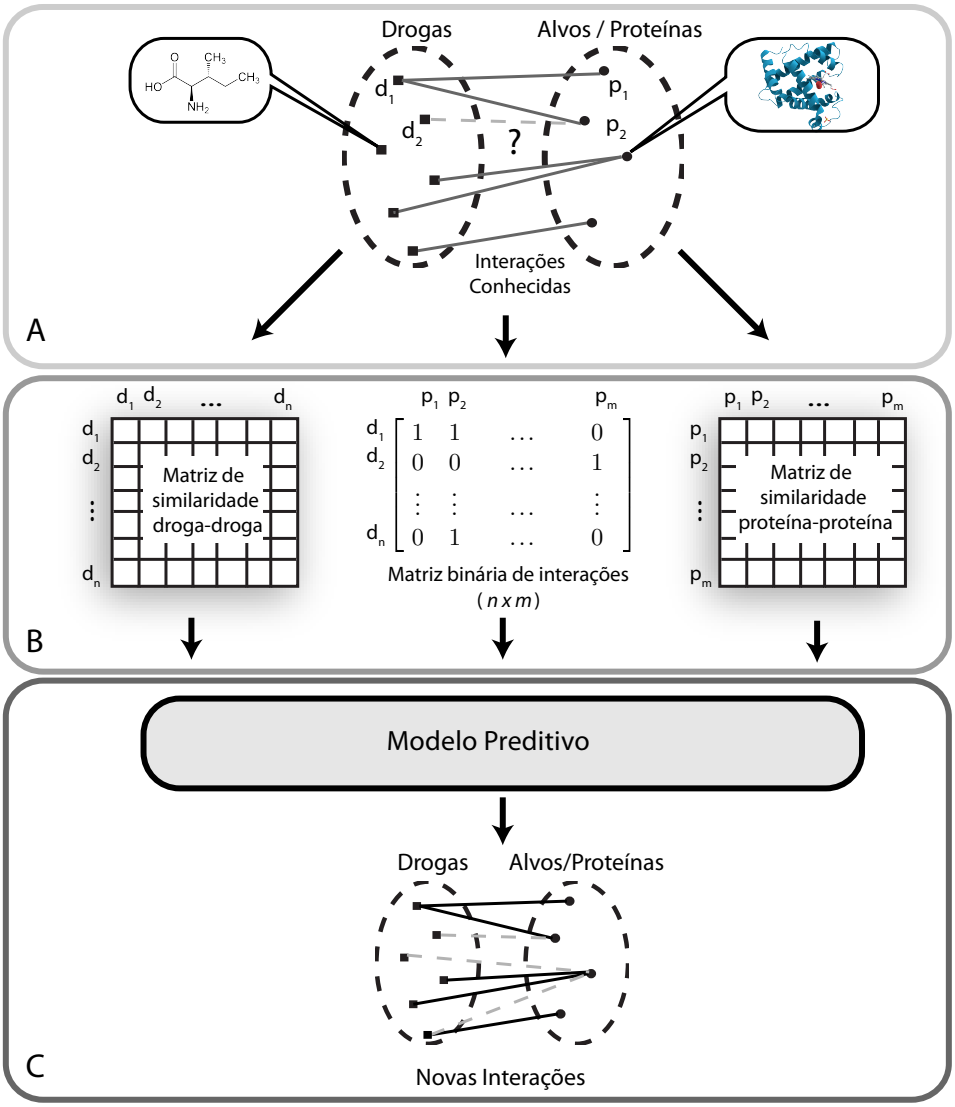

Figura 1.1: A predição de interações droga-proteína baseada em similaridade é composta de três etapas principais: a partir de um conjunto de interações conhecidas (A), são extraídas medidas de similaridade entre cada tipo de entidade envolvida (droga-droga e proteína-proteína) (B). Em seguida estas informações são utilizadas para construção de um modelo preditivo, o qual é posteriormente usado para predizer novas interações (C).

diferentes kernels trazem diferentes noções de similaridade entre as instâncias do problema, e esta possibilidade pode enriquecer o treinamento do classificador com informações relevantes. Uma outra vantagem na utilização de múltiplos kernels é que as noções de similaridade consideradas, embora sejam sobre os mesmos elementos, podem ser extraídas de diferentes representações dos mesmos. Dessa forma, a combinação de kernels pode ser vista como uma estratégia de combinação de diferentes fontes de informação, ou combinação intermediária (Figura 1.3 (c)), de acordo com a classificação proposta por SCHöLKOPF; TSUDA; VERT (2004). As outras duas abordagens seriam a integração prévia (a) e a integração tardia (b).

A integração prévia consiste na concatenação de múltiplas representações vetoriais

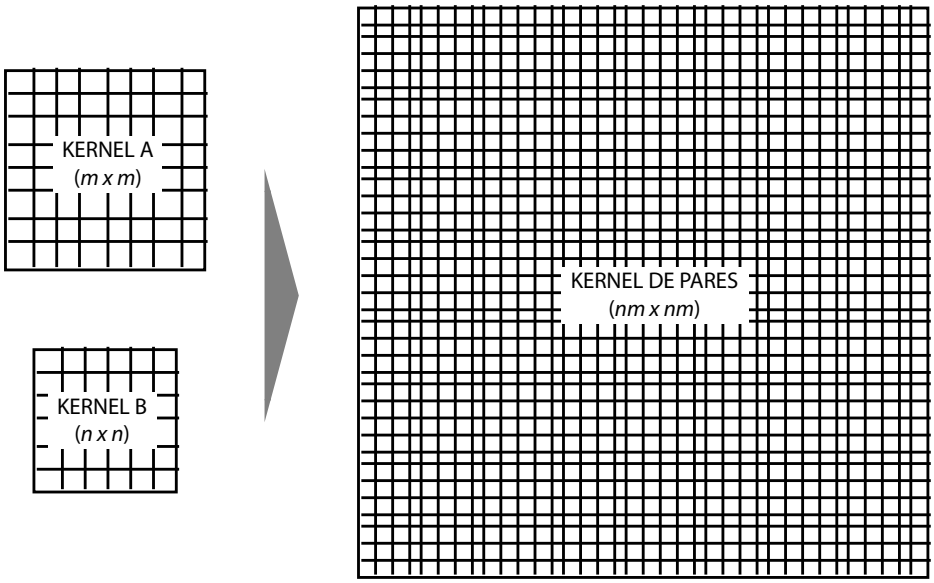

Figura 1.2: O crescimento da matriz de kernel de pares em relação ao tamanho da rede considerada.

das instâncias (*feature vectors*) em um único vetor, o qual é utilizado para treinar um único classificador. Esta estratégia impõe um aumento do vetor resultante obtido, proporcional aos tamanhos das representações base, podendo levar a uma desproporção com relação ao tamanho do conjunto de treinamento, i.e. a "maldição da dimensionalidade"(*curse of dimensionality*) (A.K. Jain; R.P.W. Duin; MAO, 2000). A integração tardia consiste em treinar diferentes classificadores, um para cada representação, e posteriormente combiná-los em um outro classificador final. A desvantagem nesse caso é a necessidade de se treinar diferentes classificadores, para uma única tarefa, podendo elevar consideravelmente os custos para a tarefa de predição. A integração intermediária encontra-se como um meio termo entre as abordagens anteriores, e diferencia-se da integração prévia pelo fato de que não utiliza os descritores das representações originais, e sim uma nova representação, baseada em similaridades entre instâncias (kernels), que por sua vez são fornecidas como entrada para um único classificador (máquina de kernel).

1.3 Definição do problema e método proposto

Sendo assim, podemos definir o problema abordado nesta tese da seguinte maneira: dada uma rede bipartida (e.g. rede droga-proteína) e diferentes visões sobre os nós da rede (i.e., kernels), o objetivo é construir um método capaz de identificar e selecionar os kernels mais relevantes para predizer novas arestas (interações) não existentes na rede original. Podemos definir em termos objetivos a principal hipótese deste trabalho como:

Hipótese: *A qualidade das predições de novos links em uma rede bipartida produzidas por um método de aprendizagem de máquina é melhorada com a integração e seleção automática*

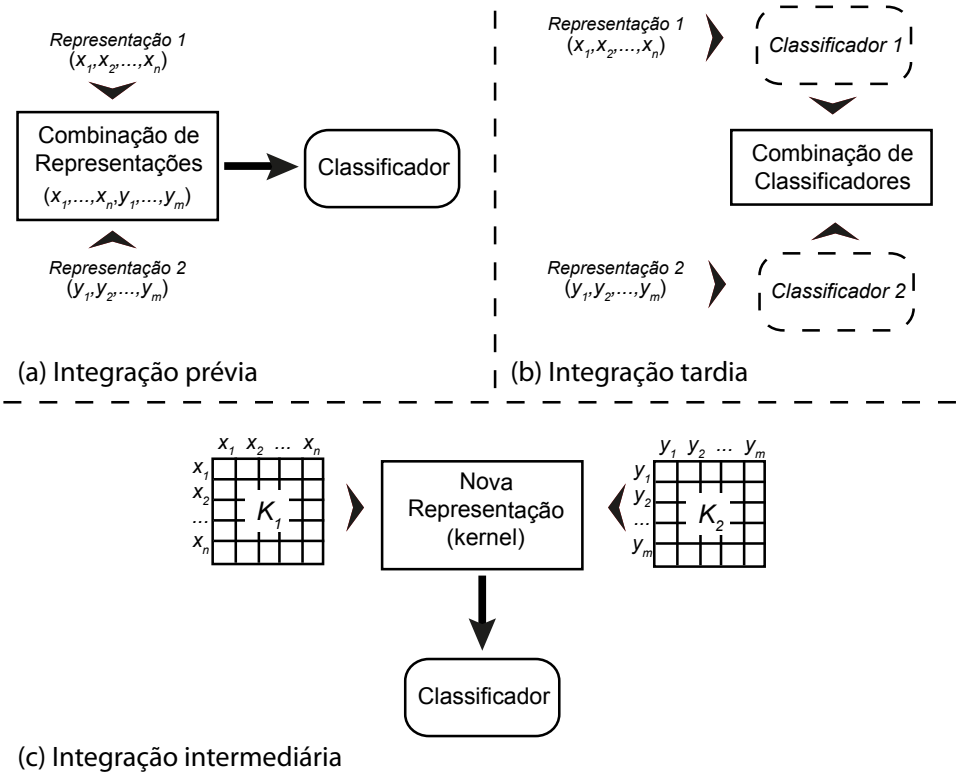

Figura 1.3: Métodos de integração para diferentes representações de características: (a) integração prévia, (b) integração tardia, e (c) integração intermediária.

de múltiplas noções de similaridade construídas sobre os nós da rede subjacente.

Métodos de MKL têm sido aplicados com sucesso em diversos problemas de aprendizagem de máquina, inclusive na predição de interações biológicas (LANCKRIET et al., 2004; BEN-HUR; NOBLE, 2005; TANABE et al., 2008; WANG et al., 2013). Entretanto, a maior parte destes trabalhos se restringe a redes unipartidas, nas quais as relações ocorrem entre entidades do mesmo tipo (e.g. redes droga-droga ou proteína-proteína). Além disso, a maioria dos algoritmos de MKL propostos na literatura é baseado no algoritmo *Support Vector Machines* (SVM) (SCHOLKOPF; SMOLA, 2001; GöNEN; ALPAYDıN, 2011), que por sua vez sofre das limitações de memória discutidas na Seção 1.1 quando aplicado ao problema de predição de interações em redes. Um outro ponto importante é o fato de que a predição de links em redes bipartidas, e.g. droga-proteína, demanda uma abordagem específica (KUNEGIS; LUCA; ALBAYRAK, 2010; LI; CHEN, 2013), uma vez que a regra de formação de arestas neste tipo de rede é diferente da encontrada em rede unipartidas. De maneira que, apesar de algumas iniciativas já envolverem a combinação de medidas de similaridade no contexto de redes bipartidas (e.g. droga-proteína) (WANG et al., 2011; PERLMAN et al., 2011; SAWADA; KOTERA; YAMANISHI, 2014), os estudos sobre a combinação de múltiplas fontes de informação neste domínio ainda são bastante limitados.

Neste trabalho, buscamos explorar três deficiências em trabalhos anteriores:

1. Poucos métodos de predição de interações são capazes de incorporar mais de um tipo de medida de similaridade para cada tipo de entidade em redes bipartidas, muitos dos quais realizam combinações simples de kernels (como a média).

2. Poucos métodos MKL são diretamente aplicáveis a problemas de predição de interações em redes bipartidas. Isto se dá principalmente porque tais algoritmos consideram que os kernels são calculados sobre o mesmo conjunto de objetos, o que não ocorre nas redes bipartidas. Neste tipo de rede, os objetos em conjuntos diferentes não são necessariamente comparáveis entre si (e.g., uma droga e uma proteína).

3. Grande parte dos algoritmos MKL utiliza o algoritmo SVM como aprendiz base (GöNEN; ALPAYDıN, 2011). As limitações impostas pelo algoritmo SVM para matrizes de kernel de grandes dimensões, como é o caso do kernel de pares (KASHIMA et al., 2009), e o custo inerente da otimização no procedimento de aprendizagem dos kernels, demandam uma solução específica para o problema.

Dessa forma, construímos um novo método MKL, específico para o problema de predição de interações em redes bipartidas, o qual chamamos de KronRLS-MKL. A solução proposta reduz o impacto do alto custo do cálculo do kernel de pares, substituindo o SVM pelo algoritmo *Kronecker Regularized Least Squares* (KronRLS) (PAHIKKALA; WAEGEMAN, 2010; PAHIKKALA et al., 2013), que até então era restrito ao uso de um único par de kernels. O algoritmo proposto consiste em um método MKL no qual os coeficientes de combinação são aprendidos em conjunto com o modelo. A otimização dos pesos é obtida por um processo de otimização alternada, em que os parâmetros do modelo e os pesos na combinação dos kernels base são otimizados de forma intercalada. A esparcidade da solução é controlada pela regularização dos pesos dos kernels, uma vez que neste tipo de problema é comum que os diversos kernels possuam informações complementares. A abordagem proposta se distingue das anteriores na combinação de kernels construídos sobre conjuntos disjuntos de objetos tanto em termos de desempenho como também em termos do consumo de memória.

Podemos ainda destacar como contribuição secundária o estudo extensivo da relevância de diferentes medidas de similaridade na tarefa da predição de interações droga proteína, uma vez que foram analisados um total de 10 diferentes descritores de drogas e 10 de proteínas. Recentemente, SAWADA; KOTERA; YAMANISHI (2014) apresentou o primeiro estudo de larga escala comparando o desempenho de um total de 18 descritores de drogas associados a 4 descritores de proteínas na mesma tarefa. Os experimentos realizados nesta tese distinguem-se deste estudo tanto em termos do algoritmo de predição utilizado, bem como em relação aos kernels considerados.

1.4 Estrutura do Documento

Este documento está estruturado da seguinte maneira: o Capítulo 2, define o problema de predição de interações através do uso de métodos de kernel. O Capítulo 3 descreve a aprendizagem de múltiplos kernels e trabalhos relacionados. O Capítulo 4 descreve em maiores detalhes o método proposto, enquanto o Capítulo 5 apresenta a metodologia experimental e os resultados obtidos. E, finalmente, o Capítulo 6 apresenta as conclusões e possíveis trabalhos futuros.

2

Predição de interações farmacológicas baseada em similaridade

Uma rede é uma estrutura composta por um conjunto de nós e um conjunto de relações entre eles. Este tipo de estrutura está presente em diversas situações do mundo real, como, por exemplo, na relação entre páginas Web, nas relações de amizade em uma rede social, ou em rotas utilizadas por companhias aéreas. Uma rede simples consiste apenas de um conjunto de nós e das ligações não direcionadas entre eles, e normalmente é representada matematicamente como um grafo, i.e., uma estrutura formal na qual os nós são chamados de vértices e ligações são chamadas de arestas.

Desse modo, uma rede pode ser representada como um grafo $G = (V,E)$, com $V = \{v_1,...,v_n\}$ vértices e $E = \{(v_i,v_j)|v_i,v_j \in V\}$, com $i,j \in \{1,2,...,n\}$, o conjunto de arestas (interações). Dentre as diversas formas de representação de tal estrutura, destacam-se a forma gráfica, como indicado na Figura 2.1, ou como uma matriz, também chamada de matriz de adjacências, a qual corresponde a uma matriz binária $Y \in \{0,1\}^{n\times n}$, onde $Y_{ij} = 1$ quando $(v_i,v_j) \in E$, e $Y_{ij} = 0$ caso contrário.

Apesar do grande poder de representação desta estrutura de rede simples, existem alguns casos mais particulares, que visam incorporar parâmetros específicos de cada domínio, e distinguem-se basicamente em relação a (KUNEGIS, 2011):

Estrutura das ligações Podem ser não-direcionadas ou direcionadas, unipartidas ou bipartidas. Numa rede não direcionada, dizemos que uma ligação (*link*) entre o nó i ao nó j (v_i,v_j) é equivalente a (v_j,v_i) (Figura 2.1 (a)). Nas redes direcionadas, é levada em consideração a origem e o destino de cada *link*, de modo que $(v_i,v_j) \neq (v_j,v_i)$. Redes bipartidas são uma classe de redes na qual os nós dividem-se em dois conjuntos disjuntos, i.e., $V = V_A \cup V_B$, com ligações apenas entre nós de grupos diferentes, ou seja, $E = \{(v_i,v_j)|v_i \in V_A, v_j \in V_B\}$ (Figura 2.1 (b)).

Tipo das ligações Nesse aspecto, as redes podem ter arestas simples, onde apenas uma ligação entre dois nós é considerada; ou múltiplas, onde diferentes arestas entre os mesmos nós

são permitidas. Informações adicionais podem ser associadas a cada aresta, como o custo de determinado caminho atribuindo-se um valor numérico (peso) a cada aresta.

Metadados Uma vez que todas as redes são criadas por um processo de crescimento, no qual nós e arestas são acrescentados ao longo do tempo, as arestas e nós podem incorporar informações paralelas (por exemplo, o tempo de criação da ligação).

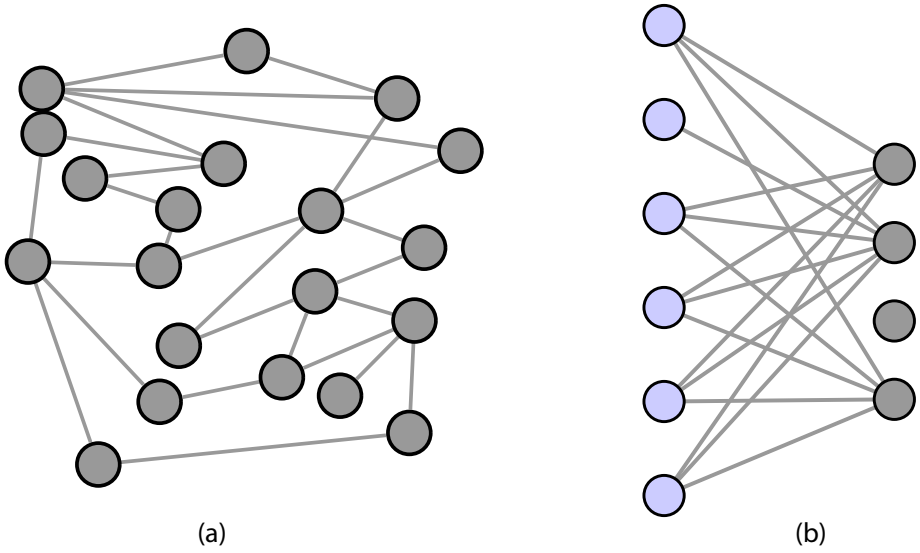

Figura 2.1: (a) Grafo não direcionado unipartido. (b) Grafo bipartido.

Boa parte dos processos biológicos podem ser modelados como uma rede, sob diferentes aspectos. Por exemplo, duas proteínas que atuam dentro de um mesmo processo biológico podem ser representadas como dois nós ligados por uma aresta; ou as reações bioquímicas dentro de uma via metabólica podem ser modeladas como um grafo direcionado, indicando a sequência em que ocorrem as reações. A modelagem do universo biológico como uma rede de interações, sejam entre proteínas, doenças ou fármacos, permite a observação de propriedades e comportamentos que dificilmente seriam observados quando analisados de forma isolada. Associado a isso, a modelagem estatística e estrutural das redes de interação entre entidades biológicas, acrescenta uma outra camada ao complexo mecanismo de ação de drogas, ou a elucidação de causas para o surgimento de doenças. A união destes universos tem viabilizado o desenvolvimento de métodos *in silico* para a identificação, validação e predição de interações entre componentes biológicos, com resultados muitas vezes validados em ensaios *in vitro* (KEISER et al., 2009).

2.1 Análise de redes biológicas

A Biologia de Sistemas é uma área que tem se destacado nas ciências biológicas, e caracteriza-se por utilizar uma abordagem analítica para investigar os relacionamentos entre os componentes de um sistema biológico, com o objetivo de entender suas propriedades emergentes

(JUNKER; SCHREIBER, 2008). Embora suas origens sejam tão antigas quanto a Biologia Molecular, o seu desenvolvimento foi acelerado nas últimas décadas, devido à convergência de uma série de avanços produzidos nas últimas décadas em diversas áreas, dentre os quais se destacam (ARRELL; TERZIC, 2010): (i) o sequenciamento de diversos genomas e a disponibilização em bancos de dados abertos; (ii) o desenvolvimento de plataformas de *High Throughput Screening* (HTS), que permitem analisar sequências de DNA ou RNA em larga escala; (iii) os avanços na ciência da computação, especialmente da Bioinformática; (iv) os avanços nas técnicas de espectrometria de massa, que permitem medir e identificar proteínas e metabólitos; e, por fim, (v) o estabelecimento da internet como principal meio de acesso e disseminação de conhecimento e dados sobre sistemas biológicos.

Abordagens sistêmicas são multidisciplinares por natureza, e neste caso, em especial, combinam múltiplas perspectivas das ciências biológicas, em suas experimentações *in vivo*, *ex vivo* e *in vitro*, bem como da ciência da computação (*in silico*), através da sua modelagem matemática e estatística. Este arcabouço visa incorporar dados de um ou mais níveis de complexidade nos organismos, sejam eles baseados em experimentos com DNA, várias formas de RNA, proteínas e modificações pós-traducionais, complexos protéicos, redes de interação, organelas, células, tecidos, órgãos, indivíduos, populações e ecologias e ambientes (ARRELL; TERZIC, 2010).

A reconstrução e modelagem de redes biológicas é um dos problemas mais desafiadores na área de Biologia de Sistemas, na qual muitos métodos e tecnologias têm sido propostos (MASON; VERWOERD, 2007; TUNCBAG et al., 2009; YAMANISHI, 2013). A elucidação de tais redes configura um grande desafio, especialmente em decorrência dos seguintes aspectos: o alto nível de ruído em experimentos biológicos; a alta dimensionalidade e complexidade do problema em relação ao número de amostras em experimentos típicos; e também de condições experimentais heterogêneas entre bases de dados. Redes biológicas podem ser observadas sob diversos aspectos e configurações (JUNKER; SCHREIBER, 2008; BARABáSI; GULBAHCE; LOSCALZO, 2011), tendo como principal característica a natureza dos nós (e.g. genes ou seus produtos, metabólitos, moléculas externas / drogas, doenças, efeitos colaterais) e o critério de interação utilizado (e.g. co-expressão, interações físicas, similaridade estrutural). A depender dos critérios utilizados, a rede produzida é caracterizada em uma das diversas classes existentes (redes proteína-proteína, droga-proteína, doença-gene, etc).

Ao mesmo tempo, muito esforço tem sido aplicado na criação de bancos de dados de moléculas, i.e. compostos químicos, sintéticos ou naturais, extraídos de animais, plantas ou microorganismos, a fim de catalogar e explorar todo o "espaço químico" de compostos possíveis. Tais bases de dados (físicas ou virtuais) têm sido utilizadas para realização de experimentos e simulações em busca de informações de bioatividade, toxicicidade e aplicabilidade de tais compostos em diversos contextos, especialmente no desenvolvimento de novos fármacos. Entretanto, os relacionamentos entre os espaços químicos e genômicos ainda são pouco conhecidos (JUNKER; SCHREIBER, 2008; DING et al., 2013).

Assim como tem ocorrido em outras áreas do conhecimento, ferramentas e métodos de análise de redes complexas têm sido cada vez mais utilizadas no estudo das redes biológicas (Lü; ZHOU, 2011), dentre os quais destacam-se os métodos de predição de interações (*links*). Neste tipo de problema, o objetivo é identificar a existência de uma ligação entre dois nós, baseado nas interações pré-existentes e nas características dos nós. Isto é extremamente relevante dentro das ciências biológicas, uma vez que o nosso conhecimento de tais redes ainda é bastante limitado, dado que até 80% das interações moleculares em células da levedura e 99.7% de humanos ainda é desconhecida (Lü; ZHOU, 2011). Diferentes métodos para a predição de links em redes complexas podem ser encontrados na literatura, e boas revisões das principais abordagens foram publicadas nos últimos anos (LIBEN-NOWELL; KLEINBERG, 2007; Lü; ZHOU, 2011).

2.2 Predição de interações em redes droga-proteína

A identificação de alvos farmacológicos é um dos primeiros passos no desenvolvimento de novas drogas, na busca por novas aplicações de drogas já comercializadas, ou na análise de seus efeitos colaterais. A identificação de interações droga-proteína *in vitro* é um procedimento caro e demorado (CSERMELY et al., 2013), de maneira que o desenvolvimento de métodos computacionais (*in silico*) é de fundamental importância para redução de custos e tempo de desenvolvimento de novos medicamentos.

Um número crescente de estudos tem investigado a relação de moléculas menores (*small molecules*) e alvos biológicos (proteínas). Uma classificação das proteínas alvo baseadas na estrutura de seus ligantes (KEISER et al., 2007) e análises da rede de interações droga-proteína conhecidas (YILDIRIM et al., 2007), revelaram características topológicas interessantes destas relações, estimulando mais pesquisas na área. Na última década, também tem-se observado um aumento na quantidade de dados de interação droga-proteína para muitas classes de proteínas importantes sob o ponto de vista farmacológico, o que inclui enzimas, canais de íons, receptores acoplados a proteínas G (*G-protein coupled receptors* - GPCR‘s) e receptores nucleares (YAMANISHI et al., 2010; LAARHOVEN; MARCHIORI, 2013a). O acesso irrestrito a tais bases de dados tem demandado o desenvolvimento de novos métodos *in silico* para predição de interações entre drogas e proteínas alvo (LAARHOVEN; NABUURS; MARCHIORI, 2011). A Figura 2.2 apresenta um exemplo de rede de interações entre drogas (círculos verdes) e proteínas GPCR (círculos vermelhos), e suas respectivas relações (neste caso, uma aresta ligando dois nós indica uma interação conhecida entre as duas moléculas). Esta estrutura constitui uma rica fonte de informação para o estudo do modo de ação destes medicamentos em um organismo, revelando relacionamentos sutis entre drogas e proteínas, que podem ajudar na predição de novas interações até então desconhecidas.

É importante ressaltar a distinção aqui dada em relação ao tratamento do problema em termos de interações conhecidas vs. desconhecidas ou em termos de interações vs. não-interações. Neste último caso, é preciso que além das interações conhecidas (i.e., confirmadas) de uma dada

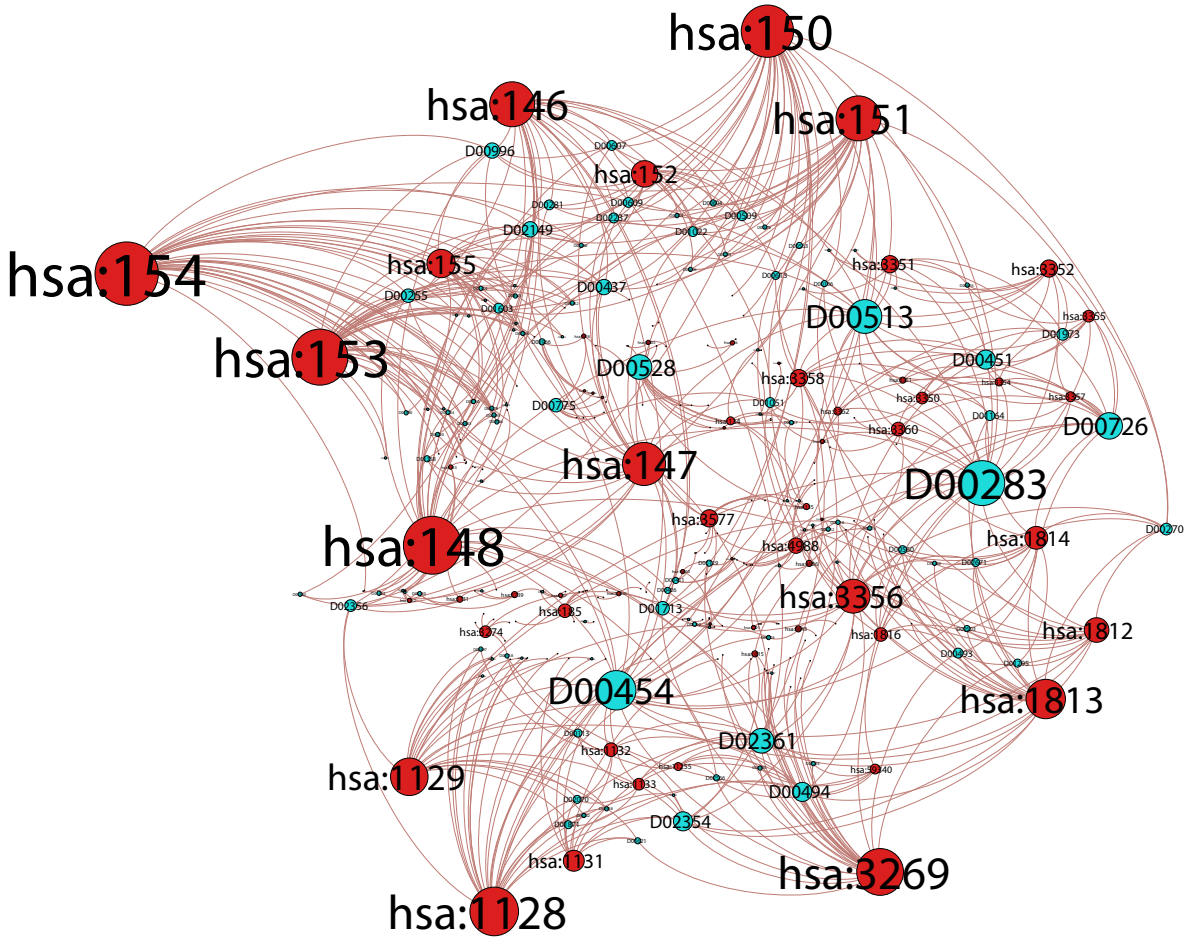

Figura 2.2: Rede de interações droga-proteína (GPCR's) extraída de bases de dados públicas (interações disponibilizadas por YAMANISHI et al. (2008)).

droga, também sejam conhecidos os alvos com os quais a droga não apresenta nenhum tipo de interação. Apesar de ser um cenário mais realista, a raridade desta informação tem estabelecido um padrão nos estudos em redes droga-proteína nos quais são consideradas apenas as interações conhecidas, enquanto o restante dos pares possíveis são tratados como interações desconhecidas.

Em termos práticos, a predição de links em redes droga-proteína pode se dar em quatro contextos distintos (YAMANISHI, 2013; PAHIKKALA et al., 2014):

(a) A droga não possui nenhum alvo conhecido (proteína), e todos os alvos possuem pelo menos uma interação conhecida. Esta seria a situação em que se deseja predizer as possíveis interações para um novo composto;

(b) O alvo não possui nenhuma interação conhecida, e as drogas no conjunto de treinamento possuem ao menos uma interação conhecida. Neste caso, o que se deseja é predizer as possíveis interações para uma proteína com potencial para ser um novo alvo farmacológico;

(c) Tanto a droga quanto o alvo possuem ao menos uma interação conhecida, e o que se deseja é predizer novas interações entre eles. Esta condição seria o equivalente a se investigar a polifarmacologia, ou o reposicionamento de drogas já aprovadas em novas aplicações;

(d) O equivalente à combinação dos itens (a) e (b), no qual podem haver tanto drogas quanto alvos sem nenhuma interação conhecida no conjunto de treinamento.

As abordagens descritas nos contextos (b) e (c) são de significante valor comercial, uma vez que ampliam os mercados de compostos já disponíveis aos consumidores, ou mesmo novos usos para compostos já desenvolvidos, mas que não chegaram a ser lançados, a baixos riscos financeiros e menor tempo de desenvolvimento (EKINS et al., 2011).

Recentemente, uma família de métodos tem se destacado como uma poderosa ferramenta para realizar predições neste tipo de problema, baseados na hipótese de que drogas semelhantes devem ligar-se a proteínas semelhantes (DING et al., 2013; YAMANISHI, 2013). Tais métodos, também conhecidos como métodos de predição baseados em similaridade, utilizam medidas de similaridade droga-droga e proteína-proteína, ou eventualmente a combinação delas, para inferir a probabilidade de interação entre drogas e proteínas na rede subjacente.

2.2.1 Métodos baseados em similaridade

O desenvolvimento de métodos de predição de interações baseados em similaridade tem algumas vantagens em relação a outras abordagens clássicas, como as baseadas em ligantes (e.g., QSAR) ou *docking*, dentre as quais destacam-se (DING et al., 2013):

1. Métodos baseados em similaridade não necessariamente exigem a extração de vetores de características (geralmente grandes) normalmente necessários nos métodos baseados em ligantes, e consequentemente evitam a etapa de filtragem e tratamento de atributos redundantes;

2. Uma vez que não requerem informações de difícil aquisição, como a estrutura tridimensional de proteínas (necessária para realização de experimentos de *docking*), os métodos baseados em similaridade podem eventualmente ser utilizados em escala genômica;

3. O cálculo de medidas de similaridade entre compostos químicos (RALAIVOLA et al., 2005), assim como entre proteínas (LESLIE; ESKIN; NOBLE, 2002; LESLIE; WESTON; NOBLE, 2002; BEN-HUR; NOBLE, 2005), é uma área bem estudada, com diversos métodos já desenvolvidos e largamente utilizados;

4. Abordagens baseadas em similaridade podem ser desenvolvidas baseadas em métodos de kernel bem estabelecidos, como o algoritmo SVM;

5. O fato de permitir que a análise seja realizada em um espaço quimiogenômico maior, possibilita análises mais abrangentes e sistêmicas do problema.

A predição de *links* baseada em similaridade é realizada com a aplicação de técnicas de aprendizagem de máquina. Neste tipo de algoritmo, a solução é composta de duas etapas básicas: construção do modelo, baseado em dados históricos, e predição de novos casos com o modelo aprendido. Estes métodos podem ser agrupados em duas grandes categorias, em relação às características dos dados utilizados na etapa de aprendizagem do modelo: aprendizagem supervisionada, quando o rótulo (ou classe) de cada padrão é conhecido a priori; e não-supervisionada, quando tal rótulo é desconhecido.

Dessa forma, a predição de links pode ser vista como uma instância da aprendizagem supervisionada, onde os dados de treinamento consistem nas informações sobre os nós que compõem a rede, acompanhados das interações conhecidas (rótulos), e o objetivo é predizer as interações com maior chance de ocorrerem na rede. Mais especificamente, a predição de links pode ser vista como um problema de classificação binária, no qual os padrões correspondem aos pares de nós da rede, e as classes positiva e negativa, à existência ou não de um *link* entre os pares de nós, respectivamente. Esta abordagem é vantajosa, uma vez que permite a aplicação de técnicas e algoritmos de aprendizagem de máquina já consolidadas na literatura.

A utilização de métodos de aprendizagem supervisionada tradicionais, i.e., baseados em extração de características e representação vetorial dos exemplos de treinamento, requer que sejam definidos previamente quais são os atributos relevantes ao problema. Em geral, tal abordagem não é diretamente aplicável a grafos, dada sua estrutura complexa e heterogênea (KASHIMA et al., 2009; LI; CHEN, 2013). Métodos de kernel são um tipo particular de algoritmos de aprendizagem que, diferentemente de métodos tradicionais, não requerem conjuntos de atributos explicitamente gerados para os dados em questão.

Métodos de kernel são uma abordagem de aprendizagem de máquina supervisionada, que contempla uma família de algoritmos para construção de métodos lineares sobre espaços multidimensionais. Tais métodos têm obtido sucesso em diversos problemas de classificação (SCHOLKOPF; SMOLA, 2001), inclusive na área de biologia computacional (SCHöLKOPF; TSUDA; VERT, 2004). Estes métodos baseiam-se na definição de uma medida de similaridade na forma de uma função de kernel, $K : X \times X \to \mathbb{R}$. Uma função de kernel válida deve atender basicamente a dois requisitos matemáticos: a matriz produzida (*gram matrix*) deve ser simétrica, i.e., $K(x,x') = K(x',x)$, e ser positiva semi-definida (PSD). A função de kernel é usada para mapear as instâncias do espaço de entrada X para o espaço de características $\mathscr{H}$, também chamado de *reproducing kernel Hilbert space* (RKHS). A função de mapeamento $\varphi(x) : X \to \mathscr{H}$ não precisa ser explicitamente definida, e deve atender a condição $K(x,x') = \langle \varphi(x), \varphi(x') \rangle$. O treinamento e classificação requerem uma máquina de kernel (*kernel machine*), ou seja, um algoritmo que utiliza funções de kernel para inferir padrões na representação dos dados no espaço de características H. Um dos principais representantes desta classe de algoritmos é o algoritmo SVM (SCHOLKOPF; SMOLA, 2001). A utilização de métodos de kernel tem se estabelecido

como estado da arte na tarefa de predição de interações droga-proteína (YAMANISHI et al., 2008, 2010; LAARHOVEN; NABUURS; MARCHIORI, 2011; DING et al., 2013; LAARHOVEN; MARCHIORI, 2013b; YAMANISHI, 2013).

A princípio, a utilização de métodos de kernel para a tarefa de predição de links requer apenas que seja definida uma medida de similaridade entre as instâncias do problema, i.e., é preciso definir uma função de kernel aplicada a pares de nós (similaridade entre pares). Este tipo de kernel, também chamado de kernel de pares, é geralmente obtido a partir de uma composição de um ou mais kernels mais simples, calculados sobre os nós que compõem a rede (kernel de nós, ou simplesmente, kernels base).

Kernels de Nós

Uma grande quantidade de medidas de similaridade de compostos químicos (drogas) e proteínas pode ser encontrada na literatura, podendo ser extraídas das mais diversas fontes. Matrizes de kernel de drogas têm sido calculadas basicamente de cinco maneiras distintas:

1. **Estruturas Químicas**: as estruturas químicas são extraídas de bases de dados públicas (e.g., KEGG (KANEHISA et al., 2008), DrugBank (WISHART et al., 2008), PubChem (WANG et al., 2009), etc.), as quais são modeladas como grafos, onde cada nó corresponde a um átomo, e as arestas às ligações covalentes entre os átomos que compõem a molécula. Em seguida, podem ser aplicadas diversas medidas de similaridade entre grafos, como o número de subgrafos comuns, que indicam o grau de semelhança entre cada par de drogas presente na base de dados;

2. **Efeitos Colaterais**: informações relativas a efeitos colaterais associados a drogas conhecidas podem ser encontradas em bases de dados públicas, como o SIDER (KUHN et al., 2008) e o *Adverse Event Reporting System* (AERS) do FDA (*U.S. Food and Drug Administration*). Uma vez removidos os termos comuns, a co-ocorrência de termos nas descrições de cada medicamento pode ser utilizada como uma medida de similaridade entre os mesmos;

3. **Expressão gênica**: a resposta em termos de expressão gênica de diferentes tecidos a drogas podem ser extraídas de bases de dados específicas, como o Connectivity Map (CMAP) (LAMB et al., 2006) e o Gene Expression Omnibus (GEO) (EDGAR; DOMRACHEV; LASH, 2002). Medidas de correlação podem ser então aplicadas aos diferentes perfis de expressão dos compostos analisados (PERLMAN et al., 2011; WANG et al., 2013);

4. **Informação Farmacológica**: informação farmacológica sobre a ação de drogas conhecidas podem ser extraídas de bases de dados, como o JAPIC[1] (Japan Pharma-

[1] `http://www.japic.or.jp/`

ceutical Information Center). Uma medida de similaridade pode então ser obtida de maneira análoga a descrita em (2) (WANG et al., 2011);

5. **Indicação Terapêutica**: a Organização Mundial da Saúde (OMS) possui um sistema de classificação terapêutica de compostos químicos, chamado de ATC (*Anatomical Therapeutic Chemical*), o qual é estruturado de forma hierárquica. Esta estrutura pode ser utilizada para uma medida de similaridade baseada na proximidade entre os termos associados a cada composto (YAMANISHI et al., 2010; WANG et al., 2011).

No que se refere a medidas de similaridade entre proteínas, estas podem em geral ser agrupadas em três categorias, de acordo com a fonte de informações utilizada:

1. **Sequências**: uma vez que as sequências de aminoácidos que compõem a proteína sejam conhecidas, estas podem ser obtidas de bases de dados públicas, como o KEGG Genes (KANEHISA et al., 2008), e utilizadas para calcular uma série de medidas de similaridade entre sequências já desenvolvidas (SMITH; WATERMAN, 1981; LESLIE; WESTON; NOBLE, 2002; LESLIE; ESKIN; NOBLE, 2002; BEN-HUR; NOBLE, 2005);

2. **Redes proteína-proteína**: medidas de proximidade na rede de interações proteína-proteína (PPI) conhecidas (STARK et al., 2006; SALWINSKI et al., 2004) podem ser utilizadas para derivar uma medida de similaridade entre duas proteínas distintas (PERLMAN et al., 2011);

3. **Anotações Funcionais**: a similaridade semântica entre anotações funcionais relacionadas a proteínas podem ser obtidas comparando-se termos GO (*Gene Ontology*) associados a cada proteína;

4. **Expressão gênica**: uma medida baseada no perfil de resposta de proteínas a diferentes condições pode ser obtida de maneira análoga a descrita anteriormente.

A Figura 2.3 ilustra algumas das principais estratégias de obtenção de kernels de nós nas abordagens de predição de interações droga-proteína baseada em similaridade. É importante ressaltar que a utilização de um método de kernel requer que a matriz de kernel obtida seja positiva semidefinida (PSD). Uma vez que nem toda medida de similaridade necessariamente produz uma matriz de kernel PSD, as matrizes de similaridade produzidas são geralmente tratadas algebricamente para que passem a ser PSD. Nas seções seguintes, matrizes de similaridade que não sejam PSD serão referenciadas como $S_{\langle\cdot\rangle}$, enquanto matrizes de kernel serão referidas por $K_{\langle\cdot\rangle}$.

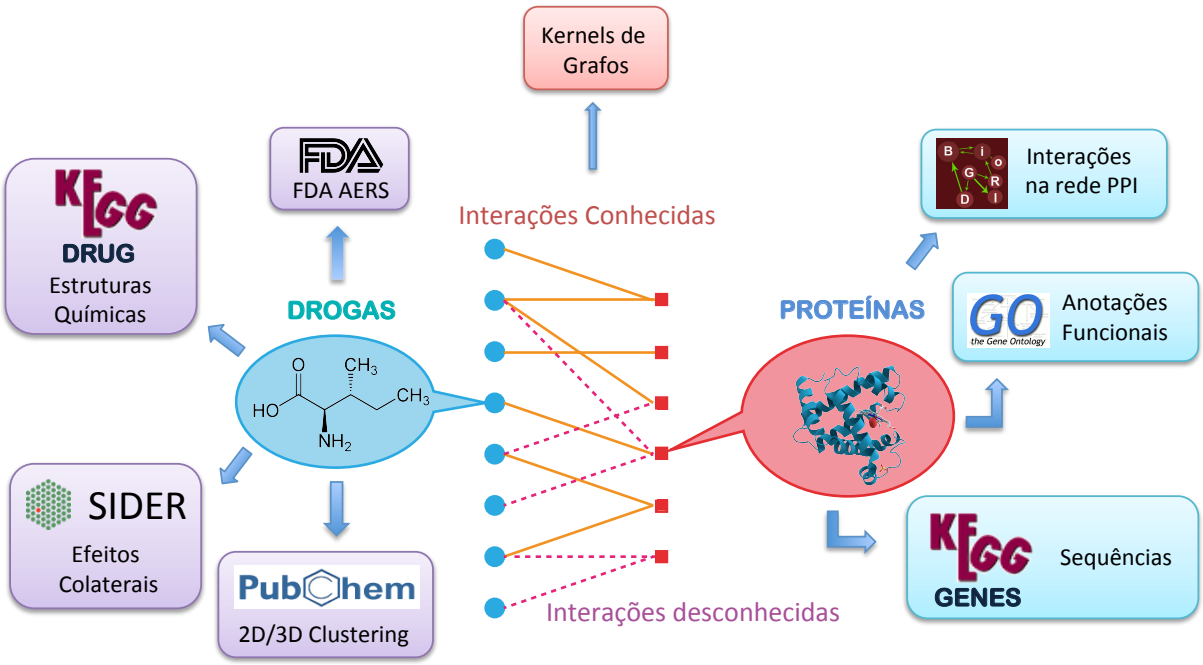

Figura 2.3: Diferentes abordagens podems ser utilizadas para extração de medidas de similaridade sobre drogas e proteínas (kernels de nós), bem como sobre a própria rede (kernels de grafos).

Kernels de Grafos

Uma vez que uma dada estrutura possa ser modelada como um grafo, podemos formular duas perguntas básicas: "Quão similares são dois nós em um dado grafo?"ou "Quão similares são dois grafos entre si?". Comparar nós em um grafo requer a construção de um kernel entre os nós, enquanto comparar grafos envolve o desenvolvimento de um kernel entre grafos. Em ambos os casos, o desafio consiste em desenvolver um kernel que capture o padrão de interação dos nós no grafo, capturando a semântica da estrutura do grafo (VISHWANATHAN; SCHRAUDOLPH, 2010).

Neste trabalho estamos mais interessados na similaridade entre nós de um grafo. A idéia de construir kernels em grafos com este propósito foi introduzida por KONDOR; LAFFERTY (2002). Desde então, muitas funções dessa natureza têm sido propostas, incluindo *diffusion kernels* (SMOLA; KONDOR, 2003; SCHöLKOPF; TSUDA; VERT, 2004), o kernel Laplaciano (YAJIMA, 2006), *commute time kernel* (FOUSS; PIROTTE, 2007), kernel marginalizado (KASHIMA; TSUDA; INOKUCHI, 2003) e perfil de interação Gaussiano (*Gaussian Interaction Profile* - GIP) (LAARHOVEN; NABUURS; MARCHIORI, 2011). Estas funções de kernel podem também ser utilizadas com o objetivo de indicar as possibilidades de existir uma ligação entre pares de nós. KUNEGIS (2011) propôs um algoritmo de ajustamento de curvas modelando o crescimento da rede com a aplicação de kernels de grafos sobre combinações da matriz de adjacências e da Laplaciana do grafo subjacente. Em YAJIMA (2006), é utilizado um kernel sobre

a matriz Laplaciana do grafo em um sistema de recomendação para capturar similaridade entre nós que estejam relacionados à distância entre eles, baseado na hipótese de que nós semelhantes e próximos tem mais tentência a interagir.

Kernels de Pares

Como foi dito anteriormente, redes droga-proteína são modeladas como um grafo bipartido. Neste caso, temos $V = V_A \cup V_B$ e $E = \{(v_a, v_b) | v_a \in V_A, v_b \in V_B\}$. Observe que os objetos pertencentes a cada partição não são comparáveis entre si, uma vez que as funções de kernel operam sobre entidades de tipos distintos. Nesse caso, as matrizes de kernel diferenciam-se não só pela noção de similaridade empregada, como também pela dimensionalidade dos conjuntos de nós presentes no grafo, conforme é demonstrado na Figura 2.4.

Sendo assim, a matriz de kernel base no caso unipartido possui dimensões $|V| \times |V|$, no qual $|V|$ corresponde ao número de vértices no grafo, e no caso bipartido, as matrizes de kernel base possuem dimensões distintas, $K_{V_A} : |V_A| \times |V_A|$ e $K_{V_B} : |V_B| \times |V_B|$, onde $|V_A|$ e $|V_B|$ correspondem ao número de vértices no conjunto V_A e V_B respectivamente. O que se deseja é construir um kernel de arestas $K_E((v_a, v_b), (v'_a, v'_b))$ onde $v_a, v'_a \in V_A$ e $v_b, v'_b \in V_B$, i.e., uma medida de similaridade entre pares de nós.

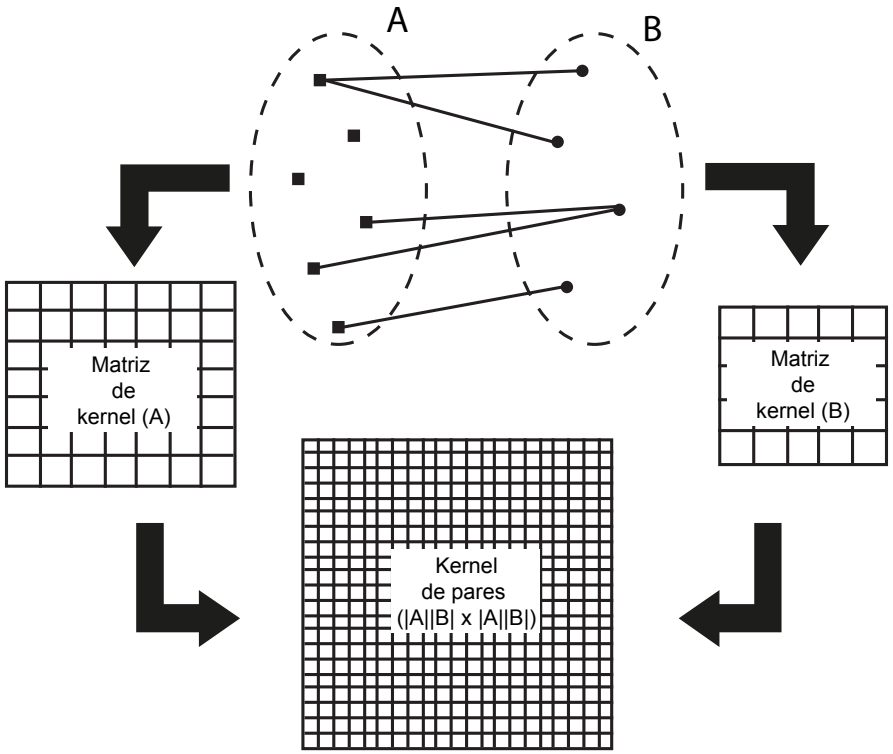

Figura 2.4: Grafo bipartido e respectivos kernels base. Podemos observar que a natureza e a dimensionalidade das matrizes de kernels obtidas são distintas para cada conjunto de vértices.

A produção de um kernel de pares para uma rede bipartida, pode ser feita de diversas formas (KASHIMA et al., 2009; LAARHOVEN; NABUURS; MARCHIORI, 2011; YAMANISHI, 2013), as quais podem ser agrupadas em 3 categorias:

1. Considere uma representação vetorial do par de vértices (v_a, v_b). Supondo que v_a seja representado por um vetor de características $\phi_A(v_a) \in \mathbb{R}^{d_A}$, e o mesmo possa ser feito com v_b, com representação $\phi_B(v_b) \in \mathbb{R}^{d_B}$, onde d_A e d_B correspondem à dimensionalidade da representação ϕ_A e ϕ_B, respectivamente. Tal representação vetorial pode ser feita, por exemplo, com o uso de descritores moleculares físico-químicos em uma droga, ou a composição de aminoácidos em uma proteína. A representação dos pares de objetos é então obtida pela concatenação dos vetores de características ϕ_A e ϕ_B, obtendo uma representação vetorial de tamanho $(d_A + d_B)$. (YU et al., 2012; YAMANISHI, 2013). A obtenção do kernel pares pode então ser obtida aplicando-se uma função adequada a representação vetoriais, como o kernel Gaussiano;

2. Uma outra abordagem é a produção de um kernel de pares como resultado do produto entre dois kernels base distintos, K_{V_A} e K_{V_B}. Ou seja, o kernel produzido corresponte a uma noção de similaridade entre arestas, sendo definido por $K_E((v_a, v_b), (v'_a, v'_b)) = K_{V_A}(v_a, v'_a) \times K_{V_B}(v_b, v'_b)$. Esta formulação equivale ao produto de tensores, o qual, no caso da representação matricial (2 dimensões), é chamado de produto de Kronecker (Apêndice .1), sendo representado por $K_{\otimes} = K_{V_A} \otimes K_{V_B}$, onde K_{V_A} é uma matriz $|V_A| \times |V_A|$ e K_{V_B} é uma matriz $|V_B| \times |V_B|$;

3. De forma análoga, podemos produzir um kernel de pares utilizando a operação de tensores, conhecida como soma de Kronecker (LAUB, 2005), definida por $K_{\oplus} = K_{V_A} \oplus K_{V_B} = I_{|V_B|} \otimes K_{V_A} + K_{V_B} \otimes I_{|V_A|}$ (KASHIMA et al., 2009), onde I_r é uma matriz identidade de dimensões $r \times r$.

Vale ressaltar que é computacionalmente inviável considerar todos os pares possíveis (positivos e desconhecidos) nas abordagens (1), (2) ou (3), uma vez que as matrizes de kernel produzidas contém todas as combinações de pares possíveis, com tamanho $|V|^2 \times |V|^2$ para redes unipartidas, e $|V_A||V_B| \times |V_A||V_B|$ para redes bipartidas, inviabilizando o armazenamento na memória até mesmo para grafos de tamanhos moderados. A utilização prática de alguns métodos de kernel, como o SVM, requer que as instâncias negativas (interações desconhecidas) correspondam a uma amostragem aleatória sobre a base de dados de treinamento, de maneira que a matriz de similaridade resultante seja capaz de ser armazenada na memória (BEN-HUR; NOBLE, 2005; NGUYEN; MAMITSUKA, 2012; LI; CHEN, 2013; DING et al., 2013).

A aplicação do produto de Kronecker (abordagem 2) tem se mostrado promissora quando se deseja construir um kernel de pares a partir de kernels base de proteínas (BEN-HUR; NOBLE, 2005) e drogas (LAARHOVEN; NABUURS; MARCHIORI, 2011; NGUYEN; MAMITSUKA,

2012; LI; CHEN, 2013; DING et al., 2013), que pode ser associado a uma máquina de kernel para realizar predições. A Figura 2.5 apresenta um esquema desta abordagem. Dada uma rede bipartida composta por $|D|$ drogas e $|P|$ proteínas, é calculada uma medida de similaridade sobre os nós da rede (kernel de nós), a partir do qual é obtido um kernel de pares. Por fim, os rótulos de cada par são extraídos da matriz de adjacências da rede, e juntamente com o kernel de pares, são utilizados para treinar uma máquina de kernel.

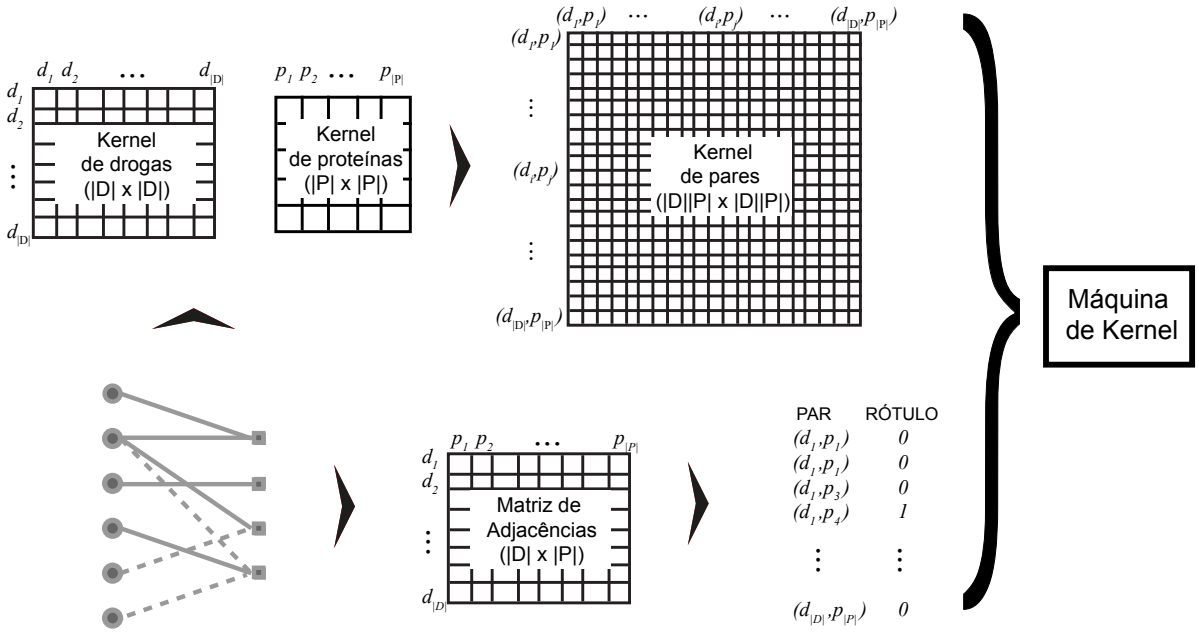

Figura 2.5: Diagrama mostrando os elementos utilizados para a produção do kernel de pares a partir de kernels base em uma rede droga-proteína com $|D|$ drogas e $|P|$ proteínas, e posterior treinamento de uma máquina de kernel.

Dessa forma, o rótulo atribuído a cada padrão de treinamento consiste na existência ou não de uma aresta (interação) entre os elementos que compõem o par em questão, de modo que o conjunto de treinamento $T \subset D \times P$ é formado a partir do conjunto de pares e, para cada par, um rótulo indicando a existência ou não da interação, ou seja, $((d,p), y_{d,p})$, onde $d \in D$, $(d,p) \in T$ e $y_{d,p} = 1$ se existir uma aresta entre os vértices d e p, e $y_{d,p} = 0$ caso contrário.

2.3 Trabalhos relacionados

Nesta seção, serão apresentados de forma breve os métodos de predição de interações em redes droga-proteína baseados em similaridade mais relevantes para este trabalho. A notação adotada será a mesma utilizada nas seções anteriores, aqui revisadas para maior conveniência: seja n_d e n_p o número de drogas e proteínas na rede, respectivamente. Seja também S_d uma matriz de similaridade entre drogas, e S_p uma matriz de similaridade entre proteínas, não necessariamente PSD. Seja K_d e K_p as matrizes de kernel (e portanto, PSD) de drogas e proteínas,

respectivamente.

Seja também $Y \subset D \times P$ a representação matricial (matriz de adjacências) da rede droga-proteína considerada, na qual $Y(d,p) = 1$ se a droga d interage com proteína p, e $Y(d,p) = 0$ caso contrário. O objetivo dada uma rede droga-proteína, é obter uma matriz F contendo os *scores* que represente a perspectiva de existência de novos *links*, calculada de maneira que seja consistente com Y.

2.3.1 Perfil mais próximo (NN)

Em YAMANISHI et al. (2008) são utilizados dois classificadores *baseline* para predição de novas interações entre pares de drogas e proteínas: o perfil mais próximo e o perfil ponderado. Seja $\mathbf{x}_d \in 0,1^{n_d}$ e $\mathbf{x}_p \in 0,1^{n_d}$ vetores representando o perfil de interação da droga d e proteína p, respectivamente. Cada posição no perfil de interação $\mathbf{x}_d$, assume valor 1, se existir uma interação conhecida entre a droga d e a proteína p_i, $1 \leq i \leq n_p$, e 0 caso contrário. Analogamente, os elementos do perfil de interação $\mathbf{x}_p$, assumem valor 1, se existir uma interação conhecida entre a proteína p e a droga d_j, $1 \leq j \leq n_d$, e 0 caso contrário. No método de perfil mais próximo, é dito que uma nova droga d' possui o seguinte perfil de interação:

$$\mathbf{x}_{d'} = S_d(d', d_{nn})\mathbf{x}_{d_{nn}}, \tag{2.1}$$

onde d_{nn} é a droga mais similar a d'. O mesmo é feito para proteínas:

$$\mathbf{x}_{p'} = S_p(p', p_{nn})\mathbf{x}_{p_{nn}}, \tag{2.2}$$

onde p_{nn} é a proteína mais similar a p'. Finalmente, os pares (d', p_i) e (d_j, p') de maior *score* nos vetores $\mathbf{x}_{d'}$ e $\mathbf{x}_{p'}$ são sugeridos como possíveis interações. Dessa forma, dois scores distintos serão atribuídos a cada par, podendo ser calculada a média entre eles para a predição final.

Os autores também apresentam uma variação do método de perfil mais próximo, chamado de *perfil ponderado*. Esta estratégia consiste em uma versão mais geral da abordagem anterior, no qual cada novo composto tem o seguinte perfil de interação:

$$\mathbf{x}_{d'} = \frac{1}{z_{d'}} \sum_{i=1}^{n_d} S_d(d', d_i)\mathbf{x}_{d_i}, \tag{2.3}$$

onde $z_{d'}$ é um termo de normalização definido por:

$$z_{d'} = \sum_{i=1}^{n_d} S_d(d', d_i). \tag{2.4}$$

De maneira análoga, o mesmo é feito para proteínas, e a predição também é feita baseada nos maiores valores de *score* dos pares (d', p_i) e (d_j, p') nos vetores $\mathbf{x}_{d'}$ e $\mathbf{x}_{p'}$. O método de predição pelo perfil mais próximo é muito simples e computacionalmente eficiente, uma vez que

não envolve nenhum tipo de otimização durante a fase de treinamento do modelo. Entretanto, os resultados obtidos por este método em estudos anteriores mostrou-se sempre bem inferior a outras abordagens, especialmente as baseadas em aprendizado supervisionado (DING et al., 2013; YAMANISHI, 2013). Uma possível explicação para o baixo desempenho é que as similaridades droga-droga e proteína-proteína utilizadas isoladamente nem sempre refletem a tendência de ligação entre os pares (YAMANISHI, 2013).

2.3.2 *Kernel Regression Model* (KRM)

YAMANISHI et al. (2008), propôs o algoritmo Kernel Regression Model (KRM), um método de aprendizagem supervisionado baseado em redes bipartidas. O método é composto de basicamente dois passos: primeiramente, a rede de interações é utilizada para construir uma matriz de similaridade,

$$L = \begin{pmatrix} L_{dd} & L_{dp} \\ L_{dp}^T & L_{pp} \end{pmatrix},$$

onde os elementos de L_{dd}, L_{pp} e L_{dp} são calculados utilizando funções Gaussianas:

$$(L_{dd})_{ij} = exp\left(\frac{-dist(d_i,d_j)^2}{h^2}\right), i,j = 1,...,n_d$$
$$(L_{pp})_{ij} = exp\left(\frac{-dist(p_i,p_j)^2}{h^2}\right), i,j = 1,...,n_p$$
$$(L_{dp})_{ij} = exp\left(\frac{-dist(d_i,p_j)^2}{h^2}\right), i = 1,...,n_d, j = 1,...,n_p,$$

onde *dist* corresponde a menor distância entre todos os objetos (compostos e proteínas) no grafo subjacente. A distância entre objetos desconectados é definida como infinita e h é um parâmetro de largura.

Vale ressaltar que o tamanho da matriz resultante é $(n_d + n_p)(n_d + n_p)$. A matriz K não é necessariamente PSD, de modo que uma matriz identidade apropriada é somada à L a fim de que ela se torne PSD. Em seguida, é feita a decomposição em autovalores da matriz $L = \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^T$. A diagonal principal da matriz Λ contém os autovalores enquanto as colunas de Γ correspondem aos autovetores de L.

Os r maiores autovalores de K são selecionados para construir uma matriz de características de dimensões $(n_d + n_p) \times r$, definida como $U = \Gamma_d\Lambda_d^{1/2}$. As drogas e proteínas passam então a ser representados por vetores de características r-dimensionais, i.e., as linhas da matriz U, ou seja $U = (\mathbf{u}_{d_1},...,\mathbf{u}_{d_{n_d}},\mathbf{u}_{p_1},...,\mathbf{u}_{p_{n_p}})^T$.

Em seguida, são treinados dois modelos, f_d e f_p, para drogas e proteínas, respectivamente.

As interações para uma dada droga d' ou proteína p', são dadas por:

$$\mathbf{u}_{d'} = f_d(d', d_i) = \sum_{i=1}^{n_d} K_d(d', d_i)\mathbf{w}_i^d$$
$$\mathbf{u}_{p'} = f_p(p', p_j) = \sum_{j=1}^{n_p} K_p(p', p_j)\mathbf{w}_j^p$$

onde $\mathbf{w}_i^d \in \mathbb{R}^{n_d}$ e $\mathbf{w}_j^p \in \mathbb{R}^{n_p}$ são ambos vetores de pesos. A otimização dos modelos é obtida pela minimização da função de perda definida como

$$\min_{\mathbf{w}_d} \| U - K_d W_d \|_F^2$$
$$\min_{\mathbf{w}_p} \| U - K_p W_p \|_F^2$$

onde $W_d = (\mathbf{w}_1^d, ..., \mathbf{w}_{n_d}^d)$ e $W_p = (\mathbf{w}_1^p, ..., \mathbf{w}_{n_p}^p)$, e $\| \cdot \|_F$ é a norma de Frobenius.

Por fim, a proximidade entre drogas e proteínas é dada pelo produto interno dos vetores de características correspondentes na matriz U, i.e., $f(d_i, p_j) = \mathbf{u}_{d_i} \cdot \mathbf{u}_{p_j}$. Os pares cujos valores de proximidade ultrapassarem um determinado limiar são então preditos como novas interações.

O KRM é considerado um marco na predição de interações droga-proteína em larga escala dado que foi o primeiro trabalho a incorporar drogas e proteínas em um único espaço de características. Uma das principais limitações deste método se dá pelo fato de que apenas um tipo de medida de similaridade pode ser utilizada para drogas e proteínas.

2.3.3 *Bipartite Local Model* (BLM)

O algoritmo *Bipartite Local Model* (BLM) (BLEAKLEY; YAMANISHI, 2009), é um método supervisionado que visa predizer interações desconhecidas combinando-se os resultados obtidos em predições baseadas em drogas e alvos isoladamente. Para isso, diversos modelos locais (classificadores) são treinados.

Mais especificamente, o método consiste na utilização do algoritmo SVM, com o qual são treinados diversos modelos locais, um para cada par droga-proteína possível. Ou seja, para uma droga d_i, um modelo SVM_{d_i} é treinado. O modelo é treinado considerando a matriz de kernel de proteínas K_p e o vetor de rótulos dado pelo perfil de interação da droga d_i, i.e., um vetor $\mathbf{y}_{d_i} \in \mathbb{R}^{n_p}$, no qual cada posição recebe $+1$ para uma interação observada e -1 caso contrário (Figura 2.6). O mesmo é feito para proteínas. A predição final consiste na média dos *scores* de ambos os classificadores.

Uma vez que o algoritmo BLM constrói um classificador para cada par droga-proteína possível, o custo computacional para construção de tantos modelos é consideravelmente alto, mesmo para redes de tamanho moderado (DING et al., 2013).

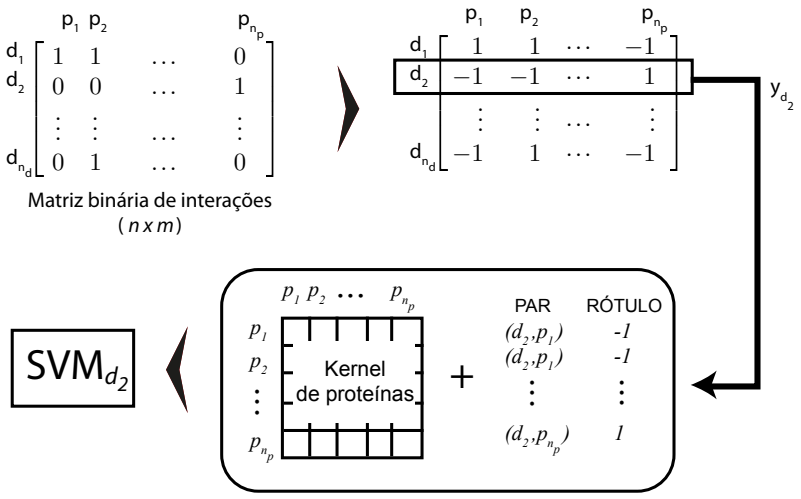

Figura 2.6: Procedimento de treinamento do algoritmo BLM para a droga d_2. Este procedimento é repetido para todas as drogas, e analogamente para todas as proteínas. Este processo resulta em um total de n_d classificadores SVM para drogas e n_p para proteínas.

2.3.4 *Pairwise Kernel Method* (PKM)

Como foi dito na seção anterior, uma abordagem mais direta para a predição de interações utilizando o algoritmo SVM é obtida constuindo um kernel de pares, a partir de kernels base. O algoritmo *Pairwise kernel method* (PKM) (JACOB; VERT, 2008) representa esta abordagem, de modo que a matriz de kernel de pares é calculada a partir das similaridades dos nós, i.e., $K_E((d,p),(d',p')) = K_d(d,d') \times K_p(p,p')$. Uma vez que o kernel de pares é contruído, o algoritmo SVM é utilizado para treinar um modelo com a matriz de kernel de pares obtida.

Apesar de ser mais eficiente que o algoritmo BLM, pois requer o treinamento de apenas um modelo para realizar as predições, a abordagem PKM possui limitações relacionadas ao uso de memória para construção do kernel de pares. Desse modo, na prática, não é possível utilizar todo o conjunto de pares possíveis no conjunto de treinamento, de modo que uma amostragem aleatória de pares negativos (interações desconhecidas) se faz necessária.

2.3.5 *Kronecker Regularized Least Squares* (KronRLS)

O algoritmo Kronecker *Regularized Least Squares* (KronRLS) (LAARHOVEN; NABU-URS; MARCHIORI, 2011; PAHIKKALA et al., 2014) é um método de kernel para problemas bipartidos, baseado no algoritmo *Regularized Least Squares* (RLS) (RIFKIN; YEO; POGGIO, 2003). O RLS tem como objetivo a minimização de uma função de custo, no caso, o erro quadrático:

$$J(f) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2} \parallel f \parallel_K^2 , \qquad (2.5)$$

onde $\| f \|_K$ é a norma da função de predição f no espaço de Hilbert associado ao kernel K, e $\lambda > 0$ é um parâmetro de regularização que determina o compromisso entre o erro de predição e a complexidade do modelo. De acordo com o teorema da representação (KIMELDORF; WAHBA, 1971), um minimizador da função objetivo acima admite uma representação dual

$$f(x) = \sum_{i=1}^{n} a_i K_E(x, x_i) \,, \tag{2.6}$$

onde $K_E : n_d n_t \times n_d n_t \to \mathbb{R}$ é a função de kernel de pares e $\boldsymbol{a}$ é o vetor de variáveis duais correspondentes a cada restrição de separação. O algoritmo RLS obtém um minimizador da Equação 2.5 pela solução do sistema de equações lineares:

$$(K_E + \lambda I)\boldsymbol{a} = \boldsymbol{y}, \tag{2.7}$$

onde $\boldsymbol{a}$ e $\boldsymbol{y} = vec(Y)$ são ambos vetores $(n_d n_t)$-dimensionais compostos dos parâmetros a_i e rótulos y_i.

Como foi dito anteriormente, um kernel de pares pode ser obtido pelo produto de dois kernels base, i.e., $K_E((d,p),(d',p')) = K_d(d,d')K_p(p,p')$, onde K_d e K_p são os kernels base para drogas e proteínas, respectivamente. Isso é equivalente ao produto de Kronecker de dois kernels base (KASHIMA et al., 2009; YAMANISHI, 2013): $K_E = K_d \otimes K_p$.

A fim de estender o algoritmo RLS para tratar de problemas de natureza bipartida, e.g., com o kernel de pares, o algoritmo KronRLS se aproveita de duas propriedades algébricas específicas do produto de Kronecker, otimizando o consumo de memória e o cálculo do modelo: o chamado *vec trick* (KASHIMA et al., 2009) e a relação entre a auto decomposição do produto de Kronecker e a auto decomposição dos seus fatores (LAUB, 2005; PAHIKKALA et al., 2014).

Sejam $K_d = Q_d \Lambda_d Q_d^T$ e $K_p = Q_p \Lambda_p Q_p^T$ as decomposições em autovalores e autovetores das matrizes de kernel K_d e K_p. A solução $\boldsymbol{a}$ pode ser dada solucionando-se a sequinte equação (PAHIKKALA et al., 2014):

$$\boldsymbol{a} = vec(Q_p C Q_d^T) \,, \tag{2.8}$$

onde $vec(\cdot)$ é o operador de vetorização que empilha as colunas de uma matriz em um vetor, e C é uma matriz definida como:

$$C = (\Lambda_d \otimes \Lambda_p + \lambda I)^{-1} vec(Q_p^T Y^T Q_d) \,. \tag{2.9}$$

Substituindo $\boldsymbol{a}$ na Equação 2.7, e reescrevendo a Equação 2.6 em forma matricial, i.e., $f(x) = K_E \boldsymbol{a}$, pode-se obter uma solução matricial, dada por (LAARHOVEN; NABUURS; MARCHIORI, 2011):

$$F = Q_d A^T Q_p^T \,. \tag{2.10}$$

O algoritmo KronRLS é bastante eficiente em tanto em termos de complexidade computacional quanto em consumo de memória. Isto se dá basicamente pelo fato de que apenas um modelo precisa ser treinado, além de não ser necessário o cálculo explícito da matriz de kernel de pares. Entretanto, o método suporta apenas um único kernel para cada tipo de entidade.

2.3.6 Outras abordagens

A abordagem proposta por CHENG et al. (2012) explora a similaridade topológica na rede de interações subjacente, e difere dos métodos anteriores pelo fato de que não utiliza uma matriz de similaridade calculada diretamente sobre drogas e proteínas, mas na verdade considera processos de transição sobre o grafo bipartido para calcular a probabilidade de interação entre objetos. Em ZHAO; LI (2010) foi apresentado um outro método de integração de características fenotípicas de drogas, índices químicos e interações PPI em um espaço genômico utilizado para predição de novas interações. Em CHEN; LIU; YAN (2012) é introduzida uma abordagem baseada em *random walks* para a mesma tarefa, no qual as predições são feitas sobre uma rede heterogênea composta de informações a respeito de similaridade entre drogas, alvos e entre interações. O trabalho de MEI et al. (2013) estende a abordagem BLM incorporando informações de vizinhos para realizar as predições.

Além das abordagens descritas acima, nas quais são consideradas informações químicas e genômicas, outras informações tem sido integradas com o mesmo objetivo, como por exemplo perfis de efeitos colaterais (KUHN et al., 2008), de expressão gênica (IORIO; TAGLIAFERRI; BERNARDO, 2009) e dados de resposta transcricional (IORIO et al., 2010).

Entretanto, apesar de existir uma grande diversidade de técnicas para predição de interações droga-proteína encontradas na literatura, a grande maioria dos trabalhos consiste em utilizar uma quantidade restrita de medidas de similaridade (kernels), na maior parte dos casos com apenas uma para cada classe de objetos (drogas e proteínas), e a combinação destas informações, quando realizada, se dá de maneira uniforme e predefinida. A aplicação uniforme deste parâmetro desconsidera a heterogeneidade dos perfis de interação de cada nó na rede, e desconsidera a grande quantidade de visões distintas sobre as entidades envolvidas no problema de predição de interações droga-proteína.

2.4 Considerações Finais

Nesta seção apresentamos alguns aspectos da tarefa de predição de interações em redes droga-proteína baseadas em similaridade, bem como uma visão geral das principais estratégias de predição. Métodos de kernel têm mostrado desempenho favorável em relação à outros métodos de aprendizagem de máquina neste contexto, em geral através da definição de kernels de pares de objetos construídos com base em kernels mais simples.

A predição de interações em redes droga-proteína, apresenta uma dificuldade no que se

refere ao desbalanceamento em relação ao número de interações conhecidas (classe positiva) e o número de interações desconhecidas (classe negativa). O desbalanceamento de classes é um problema já conhecido na literatura da área de aprendizagem de máquina (CHAWLA; JAPKOWICZ; DRIVE, 2004; TANG; ZHANG; CHAWLA, 2009). A construção de um kernel de pares é uma tarefa computacionalmente custosa, de modo que alguns métodos (em especial, aqueles baseados em SVM) utilizam uma estratégia de subamostragem de exemplos negativos durante a fase de treinamento do modelo. A principal motivação é viabilizar o cálculo e armazenamento do kernel de pares na memória. Entretanto, um problema com este tipo de abordagem é que ela pode levar a remoção de pares importantes, e limitam o espaço quimiogenômico a ser explorado (CHAWLA; JAPKOWICZ; DRIVE, 2004; DING et al., 2013).

Além disso, já foi demonstrado que o algoritmo SVM, embora seja robusto para um desbalanceamento moderado de classes, pode ter seu desempenho afetado por um alto grau de desbalanceamento entre a classe positiva e negativa (TANG; ZHANG; CHAWLA, 2009).

A utilização de diferentes fontes de dados se apresenta como uma estratégia promissora na elucidação de redes biológicas, uma vez que há uma grande disponibilidade de informações acerca destas entidades, sejam dados sobre sequências de aminoácidos, expressão gênica, estruturas, efeitos colaterais, ou ontologias de genes. Entretanto, os estudos sobre estratégias integrativas na tarefa de predição de links ainda se mostram bastante incipientes, uma vez que as estratégias de combinação em geral adotam a combinação não ponderada (uniforme) das diferentes medidas de similaridade utilizadas. Acreditamos que o desenvolvimento de métodos de combinação de informação, no formato de kernels, específicos para o contexto de predição de links em redes complexas pode trazer benefícios, levando a uma melhor qualidade nas predições.

3

Aprendizagem de múltiplos kernels

Métodos de aprendizagem de máquina têm como principal objetivo encontrar padrões em dados, i.e., seus métodos constroem modelos computacionais com o objetivo de extrair conhecimento a partir da análise de dados. Problemas de aprendizagem de máquina podem ser agrupados em duas grandes categorias: aprendizagem supervisionada e não supervisionada.

Na aprendizagem supervisionada, os padrões de treinamento são apresentados ao algoritmo de classificação acompanhados de um conjunto de rótulos (ou classes) previamente conhecidos. O objetivo consiste em inferir uma função, ou simplesmente, um classificador, que associe as características dos padrões de entrada ao rótulo correto, de modo que seja capaz de realizar a classificação até mesmo para novos padrões (sem informação de rótulo). A aprendizagem não supervisionada se caracteriza pela ausência de dados rotulados, e seu objetivo consiste em identificar padrões na distribuição dos dados de entrada.

Este trabalho pertence ao contexto de aprendizagem supervisionada, mais especificamente, ao problema de classificação binária. Dado um conjunto de exemplos de treinamento composto por vetores com C características, i.e., $X \subseteq \mathbb{R}^C$, e os seus respectivos rótulos (classes) $Y \in \{-1, 1\}$, temos que as tuplas $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, onde $(\mathbf{x}_i, y_i) \in X \times Y$ compõem o conjunto de entrada do algoritmo. O objetivo consiste em obter uma função de classificação sobre o conjunto de entrada que separe novos objetos com o menor erro possível.

Métodos de kernel são uma família de algoritmos para construção de métodos lineares sobre espaços multidimensionais, e que têm sido aplicados com sucesso em diversos problemas de aprendizagem supervisionada (SCHOLKOPF; SMOLA, 2001). O objetivo é obter um hiperplano de separação entre as classes do problema, que por sua vez é usado como fronteira entre classes por uma função de classificação (SCHOLKOPF; SMOLA, 2001):

$$f(\mathbf{x}) = sgn(\langle \mathbf{w}, \varphi(\mathbf{x}) \rangle) + b, \tag{3.1}$$

no qual $\mathbf{w} \in \mathscr{H}$ é um vetor de pesos e $\varphi : X \to \mathscr{H}$ é uma função que mapeia cada instância em um espaço de características $\mathscr{H}$. O espaço de características $\mathscr{H}$ pode ser multidimensional ou até mesmo de infinitas dimensões (SCHOLKOPF; SMOLA, 2001). Isto se faz possível devido a uma propriedade dos métodos de kernel, que possibilita construir a função f sobre espaços

multidimensionais, sem calcular explicitamente $\varphi(X)$. A única informação que o algoritmo de kernel precisa é o produto interno dos pares de instâncias no espaço de características. Este produto interno é definido através de uma função de kernel:

$$K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle. \tag{3.2}$$

Uma função de kernel $K : X \times X \to \mathbb{R}$ válida produz uma matriz simétrica positiva definida. Ou seja, k produz uma matriz quadrada de tamanho $n \times n$, devendo satisfazer $K_{ij} = K_{ji}$ para todo $1 \leq i, j \leq n$, e $c^T Kc \geq 0$ para todo $c \in \mathbb{R}^n$. Um kernel define similaridades entre os diferentes padrões do conjunto de treinamento, e tem um papel fundamental em tais técnicas. Existem muitas funções de kernel válidas, dentre as quais as mais populares são os kernels lineares, polinomiais e Gaussiano:

$$K_{LIN}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.3}$$

$$K_{POL}(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^q, q \in \mathbb{N} \tag{3.4}$$

$$K_{GAU}(\mathbf{x}_i, \mathbf{x}_j) = exp(- \parallel \mathbf{x}_i - \mathbf{x}_j \parallel_2^2 / s^2), s \in \mathbb{R} \tag{3.5}$$

O desempenho dos métodos baseados em kernel é extremamente dependente da seleção e projeto das funções de kernel utilizadas (SCHöLKOPF; TSUDA; VERT, 2004; GöNEN; ALPAYDıN, 2011). Embora a utilização de kernels mais gerais como os apresentados acima apresente desempenho satisfatório em muitos casos, o desenvolvimento de kernels específicos para diferentes representações dos dados pode melhorar bastante os resultados obtidos (KAWANABE; NAKAJIMA; BINDER, 2009). Por exemplo, o *spectrum kernel* proposto em LESLIE; ESKIN; NOBLE (2002), o *motif kernel* (BEN-HUR; BRUTLAG, 2003), o *Pfam kernel* (GOMEZ; NOBLE; RZHETSKY, 2003) e o *Mismatch kernel* (LESLIE; WESTON; NOBLE, 2002), são kernels específicos para problemas cujo espaço de entrada consiste em sequências de aminoácidos. Dado que, em certos casos, mais de uma medida de similaridade está disponível, diferentes abordagens de combinação de tais medidas têm sido propostas nos últimos anos. Tais abordagens têm sido chamadas de métodos de aprendizado de múltiplos kernels (*multiple kernel learning* - MKL), e frequentemente tem levado a uma melhoria do desempenho do modelo produzido (GöNEN; ALPAYDıN, 2011).

A utilização de múltiplos kernels baseia-se na mesma hipótese utilizada na combinação de diferentes classificadores em sistemas híbridos: ao invés de utilizar um único kernel (ou visão dos dados), é preferível utilizar um conjunto de kernels distintos, e deixar que o algoritmo encontre os melhores, ou sua combinação. GöNEN; ALPAYDıN (2011) define duas importantes motivações para o aprendizado de múltiplos kernels:

(I) Diferentes noções de similaridade nos dados: utilizar um único kernel pode adicionar um viés ao classificador, e uma seleção (ou combinação) automática de vários kernels pode levar a melhores resultados.

(II) Diferentes representações dos dados: possivelmente de diferentes fontes ou modalidades. Uma vez que são representações distintas, eles tem diferentes medidas de similaridade correspondendo a cada kernel, de modo que a combinação de kernels é uma forma de combinar múltiplas fontes de informação.

A principal característica de métodos MKL é buscar uma combinação ótima de kernels base, que minimize erros, produzindo um novo kernel, dado por:

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = f_\eta(\{K_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\}_{m=1}^P), \tag{3.6}$$

onde a função de combinação $f_\eta : \mathbb{R}^P \to \mathbb{R}$ pode ser uma função linear ou não-linear dos kernels de entrada. Funções de kernel $\{k_m(\cdot,\cdot)\}_{m=1}^P$ recebem P representações de características dos padrões de entrada, onde $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^P \in \mathbb{R}^{C_m}$, e C_m é a dimensionalidade da representação correspondente (GöNEN; ALPAYDıN, 2011).

Nas seções seguintes, apresentamos algumas das abordagens de combinação de kernels mais relacionadas com este trabalho. Por fim, uma revisão dos trabalhos cujo foco é a integração de informações sob a forma de kernels para predição de interações droga-proteína é apresentada na Seção 3.4.

3.1 Combinação por regras fixas

A forma mais frequentemente utilizada de combinação de kernels consiste na combinação linear dos mesmos, sem nenhum tipo de treinamento para definição da função f da Equação (3.6). Ou seja, dado um conjunto de kernels base $K_1, K_2, ..., K_P$, obtidos por métodos distintos, uma combinação deles pode ser obtida, por exemplo, calculando-se o somatório ou o produto dos mesmos (CRISTIANINI; SHAWE-TAYLOR, 2000; SCHöLKOPF; TSUDA; VERT, 2004):

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} K_m(\mathbf{x}_i^m, \mathbf{x}_j^m) \tag{3.7}$$

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = \prod_{m=1}^{P} K_m(\mathbf{x}_i^m, \mathbf{x}_j^m), \tag{3.8}$$

Outras abordagens incluem a seleção dos valores de maior similaridade dentre todos os kernels base considerados (WANG et al., 2011) ou a sua média:

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = max((K_m(\mathbf{x}_i^m, \mathbf{x}_j^m))_{m=1}^P) \tag{3.9}$$

$$K_\eta = \frac{1}{P} \sum_{m=1}^{P} K_m(\mathbf{x}_i^m, \mathbf{x}_j^m). \tag{3.10}$$

Uma vez que o procedimento de combinação é obtido, uma máquina de kernel pode ser treinada com o kernel K_η resultante. Já foi demonstrado que este tipo de combinação é capaz de

produzir melhores resultados em comparação ao treinamento de múltiplas máquinas de kernel, treinadas com cada um dos kernels base (PAVLIDIS et al., 2001; BEN-HUR; NOBLE, 2005; LAARHOVEN; NABUURS; MARCHIORI, 2011).

3.2 Combinação via funções parametrizadas

No lugar de utilizar uma estratégia fixa para determinar o kernel resultante da combinação dos kernels base, podemos utilizar uma variação parametrizada de f, que, em seu caso mais simples, se caracteriza pela soma ponderada dos kernels de acordo com um vetor de pesos $\eta \in \mathbb{R}^P$:

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} \eta_m K_m(\mathbf{x}_i^m, \mathbf{x}_j^m). \tag{3.11}$$

Este tipo de abordagem pode ainda diferenciar-se em relação às restrições impostas aos valores de $\{\eta_m\}_m^P$, por exemplo: soma cônica ($\eta \in \mathbb{R}_+^P$) ou soma convexa ($\eta \in \mathbb{R}_+^P$ e $\sum_{m=1}^{P} \eta_m = 1$). Os pesos nas formas cônica e convexa denotam a importância de cada kernel na combinação.

Uma outra característica importante na combinação via funções parametrizadas é se o processo de obtenção dos pesos é realizado em conjunto com o treinamento do classificador (passo único), ou se ocorre em um processo separado (passo duplo). Uma estratégia comum neste último caso, consiste em alternar dois modos de otimização: fixar os parâmetros da função de combinação e otimizar o classificador, e em seguida utilizar o classificador obtido para otimizar os coeficientes de combinação. Este processo se repete até a convergência. Existem ainda abordagens sequenciais, onde os parâmetros da função de combinação são determinados previamente, e só depois o classificador é treinado.

A combinação de kernels utilizando a soma cônica e convexa tem se mostrado promissora, geralmente superando os resultados obtidos com cada kernel separadamente (GöNEN; ALPAYDıN, 2011). LANCKRIET et al. (2004) utilizou uma abordagem de combinação cônica ao problema de classificação de proteínas. Os autores integraram o procedimento de otimização dos pesos ao algoritmo SVM, no qual os coeficientes de combinação são obtidos durante o processo de treinamento do classificador. Foram utilizadas visões heterogêneas de dados genômicos obtidos através de diferentes procedimentos experimentais (sequências de aminoácidos, perfis de hidropatia, expressão gênica e interações proteína-proteína conhecidas). Os resultados demonstraram ganhos de desempenho quando comparados ao uso dos kernels de forma isolada ou à soma não ponderada dos kernels. Os pesos finais obtidos refletem claramente a qualidade dos kernels para a tarefa em questão, atribuindo valores próximos de zero a kernels contruídos com valores aleatórios adicionados ao conjunto de kernels antes do treinamento.

A combinação de diferentes medidas de similaridade através de uma soma ponderada ou não ponderada atribui o mesmo peso para um dado kernel sobre todo o espaço de entrada. No entanto, em certos contextos, a atribuição de pesos diferentes em diferentes regiões do espaço de

entrada em um mesmo kernel pode produzir melhores resultados (GöNEN; ALPAYDıN, 2013). Esta classe de métodos de MKL, chamada de combinação derivada dos dados, busca identificar estruturas locais nos dados, nas quais cada kernel pode exercer maior influência. Neste contexto, os parâmetros de combinação (pesos) são representados por um tensor de ordem 3 ($\Omega : n \times n \times P$), que realiza a combinação dos kernels elemento a elemento (MOGUERZA; MUñOZ; DIEGO, 2004; DIEGO; MUNOZ; MOGUERZA, 2010; YANG et al., 2009):

$$K_\Omega(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} \Omega_m(\mathbf{x}_i^m, \mathbf{x}_j^m) K_m(\mathbf{x}_i^m, \mathbf{x}_j^m). \tag{3.12}$$

Uma desvantagem da combinação derivada dos dados é o alto custo para estimação de pesos individuais dos kernels para cada padrão de entrada. Em YANG et al. (2012), é proposta a utilização de uma etapa de preprocessamento, na qual os objetos (no caso imagens) são agrupados em grupos, e posteriormente o método, chamado de *Group-sensitive multiple kernel learning* (GS-MKL) procura estimar os coeficientes de combinação de kernel para cada grupo. O treinamento é feito em um único passo, ou seja, ao mesmo tempo em que os pesos são estimados, o classificador baseado em kernel é otimizado. A função de kernel obtida ao final do processo pode ser escrita como:

$$K_\eta(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} \eta_{c_i}^m K_m(\mathbf{x}_i^m, \mathbf{x}_j^m) \eta_{c_j}^m, \tag{3.13}$$

onde $\eta_{c_i}^m$ corresponde ao peso do kernel $K_m(\cdot,\cdot)$ no cluster c_i que contém o exemplo x_i.

Um ponto importante para definir a aplicabilidade do aprendizado de múltiplos kernels a determinados problemas, é o custo para se obter a função de parametrização. Este aspecto é largamente influenciado pelo método adotado para a identificação dos pesos de cada kernel. Em geral, duas categorias básicas podem ser identificadas na literatura: a métodos baseados em heurísticas e métodos baseados em otimização.

3.2.1 Combinação baseada em heurísticas

É possível estimar os parâmetros de combinação de kernels através da aplicação de uma medida de qualidade calculada sobre os kernels base. O valor de cada parâmetro pode ser extraído da própria matriz de kernel, ou mesmo do desempenho de classificadores construídos com cada kernel separadamente.

Em TANABE et al. (2008), o parâmetro η_m da equação (3.7) é selecionado analisando o desempenho de cada kernel separadamente:

$$\eta_m = \frac{\pi_m - \theta}{\sum_{h=1}^{P} (\pi_h - \theta)}, \tag{3.14}$$

onde π_m é a acurácia do modelo construído utilizando apenas K_m e θ é o limiar que deve

ser menor ou igual a acurácia mínima obtida pelos classificadores com kernel simples. Uma desvantagem desta abordagem é a necessidade de se avaliar cada kernel isoladamente, o que eleva consideravelmente o custo computacional de estimação dos pesos. Extensões a esta abordagem para tratar problemas de regressão foram apresentadas em QIU; LANE (2009).

Uma noção de *alinhamento de kernels* é apresentada em CRISTIANINI et al. (2002), a qual pode ser vista como uma medida de similaridade entre dois kernels, dada pelo coseno do ângulo entre eles. A definição empírica é definida como:

$$A(K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}, \tag{3.15}$$

onde $\langle K_1, K_2 \rangle_F = \sum_{i=1}^{N} \sum_{j=1}^{N} K_1(\mathbf{x}_i^1, \mathbf{x}_j^1) K_2(\mathbf{x}_i^2, \mathbf{x}_j^2)$. Em tarefas de classificação binária, dado o vetor de rótulos $\mathbf{y} \in \{-1, +1\}$, um kernel ideal é obtido por $K_{\text{ideal}} = \mathbf{y}\mathbf{y}^T$, uma vez que $K_{\text{ideal}}(\mathbf{x}_i, \mathbf{x}_j) = 1$ se $y_i = y_j$ e $K_{\text{ideal}}(\mathbf{x}_i, \mathbf{x}_j) = -1$ caso contrário (LANCKRIET et al., 2004). Dessa forma, o alinhamento entre um kernel e o kernel ideal é:

$$A(K, \mathbf{y}\mathbf{y}^T) = \frac{\langle K, \mathbf{y}\mathbf{y}^T \rangle_F}{N \sqrt{\langle K, K \rangle_F}}. \tag{3.16}$$

A noção de alinhamento também foi utilizada como heurística para estimar os parâmetros da função de combinação em QIU; LANE (2009):

$$\eta_m = \frac{A(K_m, \mathbf{y}\mathbf{y}^T)}{\sum_{h=1}^{P} A(K_h, \mathbf{y}\mathbf{y}^T)}. \tag{3.17}$$

Em WANG; XIE; HU (2013) é apresentada uma estratégia de combinação heurística baseada no critério do kernel de Fischer (*Kernel Fischer Criterion* - KFC). De maneira que os pesos podem ser obtidos como:

$$\eta_m = \frac{J_m}{\sum_{h=1}^{P} J_h}, \tag{3.18}$$

onde J_m corresponde ao KFC do kernel K_m, dado por:

$$J_m = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} A_B(i,j)(K_m(\mathbf{x}_i, \mathbf{x}_i) + 2K_m(\mathbf{x}_i, \mathbf{x}_j) + K_m(\mathbf{x}_j, \mathbf{x}_j))}{\sum_{i=1}^{n} \sum_{j=1}^{n} A_W(i,j)(K_m(\mathbf{x}_i, \mathbf{x}_i) + 2K_m(\mathbf{x}_i, \mathbf{x}_j) + K_m(\mathbf{x}_j, \mathbf{x}_j))}, \tag{3.19}$$

A_B e A_W representam as matrizes de afinidade entre e inter classes, respectivamente, definidas

como:

$$A_B(i,j) = \begin{cases} \frac{1}{n} - \frac{1}{n_r}, & y_i = y_j = r \\ \frac{1}{n}, & y_i \neq y_j \end{cases}$$
$$A_W(i,j) = \begin{cases} \frac{1}{n_r}, & y_i = y_j = r \\ 0, & y_i \neq y_j \end{cases},$$

n_r corresponde ao número de padrões na r-ésima classe.

3.2.2 Combinação como um problema de otimização

Os parâmetros da função de combinação também podem ser obtidos solucionando um problema de otimização. Esta solução pode ser obtida de maneira integrada ao método de kernel (aprendiz base), ou como um processo separado. Em geral, os métodos deste tipo diferenciam-se pela função objetivo adotada e suas respectivas restrições. Por exemplo, a heurística de alinhamento de kernels (CRISTIANINI et al., 2002) foi utilizada como objetivo a ser maximizado por KANDOLA; SHAWE-TAYLOR; CRISTIANINI (2002), onde o alinhamento de um dado kernel e o kernel ideal é dado por:

$$A(K_\eta, \boldsymbol{y}\boldsymbol{y}^T) = \frac{\sum_{m=1}^{P} \eta_m \langle K_m, \boldsymbol{y}\boldsymbol{y}^T \rangle_F}{N\sqrt{\sum_{m=1}^{P}\sum_{h=1}^{P} \eta_m \eta_h \langle K_m, K_h \rangle_F}}, \tag{3.20}$$

e a estimação dos coeficientes η é obtida resolvendo-se um problema de otimização:

$$\max_{\eta \in \mathbb{R}_+} \; A(K_\eta, \boldsymbol{y}\boldsymbol{y}^T). \tag{3.21}$$

De maneira semelhante, LANCKRIET et al. (2004) propõe maximizar o alinhamento entre o kernel resultante da combinação (K_η) com o kernel ideal, com a restrição sobre o traço da matriz (Apêndice .1) resultante $tr(K_\eta) = 1$ imposta sobre K_η. Outras abordagens incluem a otimização da distância entre o kernel resultado da combinação e o kernel ideal (HE; CHANG; XIE, 2008), ou ainda a utilização de outras medidas de qualidade, como o *FSM* (*feature space-based kernel matrix evaluation measure* (NGUYEN; HO, 2007; TANABE et al., 2008) e a divergência de Kullback-Leibler (KL) (YING; HUANG; CAMPBELL, 2009).

Uma grande quantidade de trabalhos integra o processo de otimização dos pesos ao aprendizado de um classificador SVM (GöNEN; ALPAYDıN, 2011; RAKOTOMAMONJY; BACH, 2008; LANCKRIET et al., 2004). O algoritmo SVM busca construir um plano de separação com margem máxima no espaço de características induzido pela função de mapeamento ϕ. Um classificador (consideramos aqui o SVM de margem suave) pode ser obtido através da

minimização do erro de classificação no conjunto de treinamento:

$$\begin{aligned} \min \quad & \frac{1}{2} \parallel \mathbf{w} \parallel_2^2 + C \sum_{i=1}^{n} \xi_i \\ w.r.t.^1 \quad & \mathbf{w} \in \mathscr{H}, \xi \in \mathbb{R}_+^n, b \in \mathbb{R} \\ s.t.^2 \quad & y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, \quad \forall i, \end{aligned}$$

onde C é uma constante que impõe um peso diferente para o treinamento em relação à capacidade de generalização do modelo (determinada empiricamente) e ξ é um vetor de variáveis de relaxamento que suavisam as restrições impostas na determinação do hiperplano de separação ótimo. A solução deste problema é obtida pela introdução de multiplicadores de Lagrange, que leva ao problema de otimização *dual* (o qual é efetivamente resolvido na prática):

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j}^{n} a_i a_j y_i y_j K(x_i, x_j) \\ w.r.t. \quad & \boldsymbol{a} \in \mathbb{R}^n, \\ s.t. \quad & a_i \geq 0 \quad \forall i = 1, ..., n, \quad \sum_{i=1}^{n} a_i y_i = 0. \end{aligned}$$

Seja $\boldsymbol{a}^*$ a solução ótima do problema *dual* descrito acima, e sabendo que no contexto de MKL, são considerados múltiplos kernels, ou seja, $K_\eta = \sum_{m=1}^{P} \eta_m K_m$, pode-se chegar a seguinte função objetivo (RAKOTOMAMONJY; BACH, 2008):

$$J(\eta) = \sum_{i=1}^{n} a_i^* - \frac{1}{2} \sum_{i,j}^{n} a_i^* a_j^* y_i y_j \sum_{m=1}^{P} \eta_m K_m(x_i, x_j).$$

Diversas estratégias de otimização dos coeficientes de combinação dos kernels base feitas concomitantemente com a construção de um classificador SVM foram propostas (GöNEN; ALPAYDıN (2011) apresenta uma extensa revisão métodos deste tipo). Uma estratégia bastante comum é envolver o otimizador usual do SVM com um procedimento de otimização exclusivo para os coeficiente de combinação η, cujo principal representante é o algoritmo SimpleMKL (RAKOTOMAMONJY; BACH, 2008). A principal diferença entre as diferentes abordagens consiste na metodologia de otimização utilizada, no classificador base, bem como nas restrições impostas à matriz de kernel resultante e ao vetor de pesos usados na combinação. Por exemplo, LANCKRIET et al. (2004) resolve o problema de otimização acima usando programação semidefinida com a restrição $trace(K_\eta) = t$, onde t é uma constante. DINUZZO (2011) propõe uma extensão ao algoritmo *Regularized Least Squares* (RLS) para lidar com múltiplos kernels, no qual também é adotado um procedimento de otimização alternada semelhante à utilizada no

[2] *w.r.t* corresponde a notação em problemas de otimização para "em relação a"("*with relation to*").

[2] *s.t* corresponde a notação em problemas de otimização para indicar as restrições impostas a solução ("*subject to*").

SimpleMKL.

3.3 Combinação não esparsa de kernels

Grande parte dos trabalhos sobre MKL propostos na literatura impõem uma regularização através da norma l_1 sobre os coeficientes de combinação dos kernels. Isto se dá especialmente por que este tipo de norma é caracterizada pela esparsidade no vetor η obtido (KLOFT et al., 2008; YU et al., 2010), atribuindo valores dominantes sobre apenas um ou dois kernels.

Este é um resultado desejável em situações em que o que se quer é selecionar um pequeno conjunto de fontes de informação relevantes, dentre um grande número de fontes de baixa qualidade. Entretanto, esta não é a regra geral em abordagens integrativas em problemas biológicos, nas quais muitas vezes se dispõe de poucas representações dos dados, as quais são cuidadosamente extraídas e podem conter informações complementares e relevantes ao problema em questão. YAN et al. (2009) demonstrou que a norma l_2 apresenta resultados superiores a norma l_1 quando comparados em problemas de classificação de imagens e vídeos, e concluem que a norma l_2 deve ser utilizada quando os kernels base contém informações complementares. LONGWORTH; GALES (2008) introduz um termo de regularização ao algoritmo SimpleMKL (RAKOTOMAMONJY; BACH, 2008) a fim de controlar a esparsidade do vetor de coeficientes usados na combinação dos kernels:

$$\sigma \sum_{m=1}^{P} \eta_m^2,$$

onde σ é um coeficiente de regularização. YU et al. (2010) avaliou a utilização da norma l_2 na combinação de diversos conjuntos heterogêneos de kernels genômicos em 6 bases de dados biomédicas, e demonstrou que métodos baseados na norma l_2 apresentaram melhores resultados na maior parte das bases avaliadas. Resultados semelhantes também foram apresentados por KLOFT et al. (2008). Em KLOFT; LASKOV; ZIEN (2009) os autores estendem o problema para o caso de normas $l_p(p \geq 1)$ arbitrárias, adicionando o termo $\sigma \parallel \eta \parallel_p^p$ a função objetivo, o que leva ao seguinte problema de otimização:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i - \frac{1}{2} \left(\sum_{m=1}^{P} \left(\sum_{i,j}^{n} a_i a_j y_i y_j K_m(x_i, x_j) \right)^{\frac{p-1}{p}} \right)^{\frac{p}{p-1}} \\ w.r.t. \quad & \boldsymbol{a} \in \mathscr{R}^n, \\ s.t. \quad & a_i \geq 0 \quad \forall i = 1, ..., n, \quad \sum_{i=1}^{n} a_i y_i = 0. \end{aligned}$$

3.4 Abordagens integrativas na predição de interações droga-proteína

Nos últimos anos, a popularização de bancos de dados de compostos químicos, proteínas, e também sobre as suas interconexões, tem motivado o desenvolvimento de métodos com o objetivo de inferir novas interações baseadas nas interações já conhecidas. Entretanto, embora pareça natural que cada fonte de dados seja importante em determinados aspectos, e contribua de forma singular ao problema, muitas dessas abordagens baseiam-se em apenas um único kernel para cada tipo de entidade, como foi visto no Capítulo 2. A integração de dados biológicos através de métodos de kernel tem se mostrado promissora (WANG et al., 2011; PERLMAN et al., 2011; WANG; ZENG, 2013; CSERMELY et al., 2013), muito embora muitas destas iniciativas não sejam explicitamente inspiradas no arcabouço MKL.

É importante frisar que a aplicação direta de abordagens MKL tradicionais a problemas de predição de links se restringe a combinação de kernels calculados sobre um conjunto homogêneo de padrões de entrada, i.e., em geral, operam sobre matrizes com a mesma dimensão. De modo que tais métodos podem ser aplicados essencialmente a redes unipartidas, utilizando, por exemplo, um kernel de pares, em conjunto com uma máquina de kernel (e.g., SVM). Estes trabalhos integram matrizes de similaridade construídas sobre as mesmas entidades (i.e., proteínas), e não se aplicam ao contexto de redes bipartidas no qual se encontram as redes droga-proteína (bipartidas). O número de trabalhos com foco na integração de dados em problemas com estrutura bipartida ainda é muito limitado, de modo que os métodos propostos em geral possuem pelo menos uma das limitações abaixo (WANG et al., 2011; ARANY et al., 2012; YANG; XU; ZENG, 2014; SAWADA; KOTERA; YAMANISHI, 2014):

- não são escaláveis em relação ao número de kernels utilizados;
- utilizam apenas combinações simples de kernels base (e.g., soma não ponderada e média);
- baseiam-se no algoritmo SVM, e em geral realizam sub-amostragem de exemplos negativos (conforme descrito no Capítulo 2).

Dito isto, nos concentraremos nesta seção nos trabalhos cujo foco tenha sido a integração de dados em redes droga-proteína, e destacamos em quais aspectos esta proposta se diferencia delas.

3.4.1 Wang, 2011

Em WANG et al. (2011), foi realizada a integração de diversas medidas de similaridade de compostos químicos, com uma medida de similaridade de proteínas, para elucidação de interações droga-proteína. O método baseia-se em métodos de kernel (SVM), e a estratégia

de integração consiste em um kernel de pares produzido para cada possível combinação de kernels, conforme $K((d,p),(d',p')) = K_d(d,d')K_p(p,p')$. No total foram utilizados três kernels de drogas (baseados em informações terapêuticas, farmacológicas e estruturas químicas) e apenas um kernel para proteínas (baseado na similaridade de sequência genômica).

Cada combinação possível de kernels de drogas e proteína foi avaliada separadamente, e em seguida foi utilizada uma estratégia simples para combinação de pares de kernels de drogas. A fusão entre um par de kernels é feita selecionando o maior valor de similaridade entre as duas matrizes, i.e., seja K_{chem} um kernel baseado em estruturas químicas, e K_{ther} um kernel baseado em informações terapêuticas, o resultado da combinação entre eles é dado por $K_d(d,d') = max(K_{chem}(d,d'),K_{ther}(d,d'))$. Em seguida, os autores selecionam aleatoriamente o mesmo número de exemplos positivos (interações conhecidas) dentre os exemplos negativos (interações desconhecidas). O conjunto resultante é utilizado para treinar um classificador *one-class* SVM, utilizando o kernel de pares construído.

Observe que o algoritmo proposto pode ser reduzido ao método PKM, descrito na Seção 2.3.4, com a diferença de que os kernels base utilizados são construídos através de uma heurística de combinação prévia (*max*). Por este motivo, neste trabalho, referenciaremos esta abordagem por `PKM-MAX`. Uma desvantagem clara desta abordagem é que as interações negativas descartadas durante o processo de amostragem aleatória podem conter interações potenciais, limitando o espaço amostral do modelo. Um outro fator a ser destacado é que a abordagem de combinação não é capaz de identificar e selecionar os kernels relevantes ao problema.

3.4.2 WANG-MKL

Em WANG et al. (2011), também é sugerida (embora não tenha sido considerada nos experimentos realizados no artigo) uma heurística para identificação da relevância de cada kernel, utilizando uma combinação convexa dos mesmos, baseado na proximidade dos nós na rede de interações subjacente. Neste trabalho, chamaremos esta abordagem de `WANG-MKL`.

Os autores sugerem o aprendizado dos vetores de pesos ótimos para o problema, através da maximização do coeficiente de correlação entre a matriz de kernel obtida e a topologia da rede de interações droga-proteína. Este objetivo pode ser alcançado solucionando um problema de otimização definido como:

$$\max_{\boldsymbol{\eta}_d} \quad |corr(K_d(\boldsymbol{\eta}_d), dist)|,$$

onde $K_d(\boldsymbol{\eta}_d)$ corresponde a combinação das matrizes de kernel de drogas com o vetor de pesos $\boldsymbol{\eta}_d$, *dist* corresponde a distância (número de saltos) entre as drogas na rede subjacente, e *corr* representa o coeficiente de correlação. Analogamente, o mesmo pode ser realizado para proteínas. O problema acima pode ser solucionado através de programação linear, e leva a uma seleção explícita dos kernels de entrada, cujos pesos podem ser utilizados para denotar a relevância de

cada kernel para o problema.

3.4.3 SITAR

No trabalho proposto por PERLMAN et al. (2011), é apresentado o algoritmo SITAR (*Similarity-based Inference of drug-TARgets*), no qual o processo de predição de novas interações consiste em três passos: construção de um vetor de características para cada par droga-proteína a partir das matrizes de similaridade; seleção das características relevantes através de um procedimento de otimização guloso baseado na acurácia (*forward selection and backward elimination*), no qual dois procedimentos ocorrem em paralelo: um se inicia com um conjunto vazio de atributos e a cada passo acrescenta a característica que mais contribui para o resultado, e outro, que inicia com o conjunto completo de características, e a cada passo remove a característica que menos contribui com o desempenho. Ao final, o conjunto de medidas com maior acurácia é selecionado.

Em seguida, para classificar um dado par (d, p), para cada interação verdadeira (d', p'), é calculada sua similaridade droga-droga $S_d(d, d')$ e proteína-proteína $S_p(p, p')$. As medidas obtidas são então combinadas em um único score, através de uma versão ponderada da média geométrica:

$$max\left(S_d(d,d')^r S_p(p,p')^{(1-r)}\right), 0 \leq r \leq 1. \tag{3.22}$$

Por fim, os scores de todas as interações positivas são integradas em um vetor de características, que é utilizado para treinar um classificador de regressão logística.

Os experimentos realizados utilizaram um conjunto de cinco medidas de similaridade para drogas (baseadas em estruturas químicas, perfis de ligantes, perfis de expressão gênica, uso terapêutico e efeitos colaterais associados) e três para proteínas (baseados em sequências genômicas, anotações funcionais e proximidade na rede proteína-proteína).

O algoritmo SITAR não é um método MKL propriamente dito, uma vez que não realiza a combinação linear das medidas de similaridade utilizadas, além de não ser baseado em um método de kernel. Na verdade, a abordagem utilizada está mais próxima de uma seleção de características rígida, e não permite uma medida clara da qualidade das medidas de similaridades adotadas. Além disso, o algoritmo não é escalável para o treinamento em todo o espaço quimiogenômico, visto que também requer a sub-amostragem aleatória de exemplos negativos.

3.5 Considerações Finais

Neste capítulo, apresentamos os principais conceitos da aprendizagem de múltiplos kernels, desde estratégias básicas de integração de dados, como a média ou a soma de kernels base, como também os principais avanços na seleção e ponderação automática de kernels. Entretanto, vimos que a aplicação de estratégias clássicas de MKL na predição de interações em redes droga-proteína leva a necessidade de realizar adaptações, devido a natureza bipartida em

tais redes. Também foram apresentados trabalhos cujo foco consiste na integração de informações baseada em kernels em redes farmaco-genômicas.

Apesar das muitas iniciativas na direção da integração de dados no contexto da predição de interações em redes biológicas, os estudos sobre estratégias de combinação de dados sob um ponto de vista de aprendizagem de máquina ainda são bastante limitados. Dentre os métodos revisados, apenas as estratégias baseadas em heurísticas (e.g., similaridade, etc.) e o algoritmo `WANG-MKL` são capazes de retornar uma medida direta da qualidade dos kernels considerados. A fim de preencher o espaço na utilização do framework MKL em problemas de predição de interação em redes bipartidas (e.g. droga-proteína), e com foco na escalabilidade do método para redes com grande número de nós, este trabalho propõe um método baseado em otimização alternada, capaz de realizar a predição de interações baseado em múltiplos kernels.

4

Aprendizado de múltiplos kernels para predição de interações em redes droga-proteína

Neste capítulo, apresentamos um método de combinação de múltiplos kernels (MKL) com foco no problema de predição de interações em redes droga-proteína. Métodos de MKL tradicionais buscam encontrar uma combinação ótima de kernels para tarefas de classificação ou regressão, e permitem que diferentes visões dos dados, na forma de kernels, possam ser incorporadas ao modelo, com o objetivo de classificar as instâncias do problema corretamente. Entretanto, estes kernels representam uma noção de similaridade construída sobre um mesmo tipo de entidade, e.g., usuários em uma rede social, proteínas em um organismo, compostos químicos, etc.

Redes droga-proteína podem ser modeladas como grafos bipartidos, e a predição de interações em tais redes pode ser abordada como um problema de aprendizagem de máquina, conforme apresentado no Capítulo 2. Dessa forma, diferentemente do que ocorre na aprendizagem de kernels tradicional, os kernels em redes bipartidas representam noções de similaridade para classes de objetos distintos, neste caso, drogas e proteínas, de modo que não podem ser tratados como um conjunto único de kernels.

Recentemente, estratégias de MKL têm sido usadas em problemas de predição de interações, especialmente em redes unipartidas, e.g., proteína-proteína (PPI) (TANABE et al., 2008; BEN-HUR; NOBLE, 2005). Entretanto, a maior parte destas iniciativas utiliza combinações simples (soma, média ou produto) de kernels base (BEN-HUR; NOBLE, 2005), ou a utilização uniforme de pesos sobre todo o espaço de entrada (LAARHOVEN; NABUURS; MARCHIORI, 2011; SAWADA; KOTERA; YAMANISHI, 2014), de modo que estes estudos têm limitações a serem ainda exploradas. A Tabela 4.1 categoriza as abordagens integrativas na predição de interações droga-proteína recentemente propostas, em termos da quantidade de kernels de drogas e proteínas, tipo de combinação, algoritmo de aprendizado utilizado e se requer a sub-amostragem de pares desconhecidos.

Embora o interesse neste tipo de abordagem tenha crescido bastante nos últimos anos, os estudos ainda são bastante limitados, seja em termos da quantidade de kernels base utilizados,

Tabela 4.1: Trabalhos recentes com abordagens de integração de dados heterogêneos através de métodos de kernel para predição de interações droga-proteína.

Referência	Kernels de Drogas	Kernels de proteínas	Tipo de Combinação	Aprendiz Base	Sub-amostragem
JACOB; VERT (2008)	1	5	Heurística	SVM	Sim
PERLMAN et al. (2011)	5	3	Seleção de características	Regressão Logística	Sim
WANG et al. (2011)	3	1	Heurística	SVM	Sim
LAARHOVEN et al. (2011)	2	2	Média	KronRLS	Não
SAWADA et al. (2014)	18	4	Média	PKR	Não

como em relação aos métodos de combinação e seleção de kernels. Dito isto, neste trabalho buscamos explorar deficiências em trabalhos anteriores, dentre as quais destacamos:

1. A combinação de kernels na predição de interações em redes bipartidas requer o desenvolvimento de métodos específicos, dadas as caracerísticas deste tipo de problema. Poucos trabalhos foram desenvolvidos neste contexto anteriormente, muitos dos quais realizam combinações simples de kernels, como a média (LAARHOVEN; NABUURS; MARCHIORI, 2011; SAWADA; KOTERA; YAMANISHI, 2014). Utilizamos estratégias de aprendizagem de kernels para desenvolver um método para combinação (não esparsa) de kernels em redes bipartidas. A aprendizagem da função de kernel induz um processo automático de seleção das fontes de informação mais relevantes ao problema;

2. Uma vez que grande parte dos algoritmos MKL utiliza o algoritmo SVM como aprendiz base (GöNEN; ALPAYDıN, 2011), estes sofrem das mesmas limitações para matrizes de kernel de grandes dimensões, como é o caso do kernel de pares (KASHIMA et al., 2009), bem como o alto grau de desbalanceamento de classes neste contexto. Tais limitações associadas ao custo inerente da otimização no procedimento de aprendizagem dos kernels, demandam uma solução específica para o problema. Dessa forma, optamos por uma adaptação de uma técnica mais adequada ao kernel de pares, o algoritmo *Kronecker Regularized Least Squares* (KronRLS) (PAHIKKALA; WAEGEMAN, 2010; PAHIKKALA et al., 2013), mas que até então era restrita ao uso de um único par de kernels;

3. Não temos o conhecimento de nenhum trabalho que tenha avaliado o mesmo conjunto de kernels na predição de interações droga-proteína, sob diferentes estratégias de combinação. Recentemente, SAWADA; KOTERA; YAMANISHI (2014) propôs um estudo comparativo avaliando a qualidade de kernels para a tarefa de predição de interações em redes droga-proteína. Entretanto, a nossa proposta diferencia-se tanto em termos da quantidade de kernels considerados (10 kernels de drogas e 10 kernels de proteínas) como em termos dos métodos de aprendizagem e de seleção e combinação de kernels base.

Dessa forma, construímos um novo método MKL, específico para o problema de predição de interações em redes bipartidas, o algoritmo KronRLS-MKL (Figura 4.1). O algoritmo proposto consiste em um método MKL, no qual os coeficientes de combinação são aprendidos em conjunto com o modelo de predição. Dois métodos distintos de otimização dos pesos foram desenvolvidos: um considerando uma combinação de kernels com pesos arbitrários, e um segundo que realiza a combinação convexa de kernels. Ambos são baseados em um processo de otimização alternada, em que os parâmetros do modelo e os pesos na combinação dos kernels base são otimizados de forma intercalada.

Uma vez que é esperado que os kernels base contenham informações complementares, introduzimos um termo de regularização (l_2) ao processo de otimização, a fim de controlar a esparsidade da solução. Esta é uma abordagem recente na aprendizagem de kernels (KLOFT et al., 2008), cujos resultados indicam ser apropriadas a problemas com esta característica, i.e., complementaridade de informações embutida nos kernels base.

Apesar dos estudos aqui apresentados estarem limitados ao escopo de redes droga-proteína, o algoritmo proposto é aplicável a outros domínios com estrutura bipartida, como, por exemplo, filtragem colaborativa, redes de autoria ou redes de documentos e suas características.

Este capítulo está estruturado da seguinte maneira: na Seção 4.1 é apresentado o algoritmo KronRLS, utilizado como método de aprendizagem base para os métodos propostos. Na Seção 4.2 são apresentadas as extensões feitas ao algoritmo KronRLS para contemplar a combinação de múltiplos kernels. A Seção 4.3 apresenta a metodologia para combinação não esparsa de bases de dados heterogêneas e respectivas estratégias de otimização. Considerações finais são discutidas na Seção 4.4.

4.1 KronRLS

O algoritmo KronRLS consiste em uma adaptação do algoritmo RLS (RIFKIN; YEO; POGGIO, 2003) para lidar com kernels construídos sobre arestas em um grafo. Neste caso, seja um conjunto de drogas $D = \{d_1, ..., d_{n_d}\}$, e $p = \{p_1, ..., p_{n_p}\}$ um conjunto de proteínas, bem como o conjunto de padrões de treinamento x_i (pares droga-proteína) e seus respectivos rótulos binários $y_i \in \mathbb{R}$ (em que 1 corresponde a uma interação conhecida e 0 caso contrário), com $1 < i \leq n$, $n = |D||P|$ (número de pares droga-proteína). Neste caso, a função objetivo do algoritmo RLS (equação (2.5)) pode ser escrita em formato matricial, no qual é considerado um kernel de pares (produto de Kronecker entre dois kernels base):

$$(\boldsymbol{y} - K_E\boldsymbol{a})^T(\boldsymbol{y} - K_E\boldsymbol{a}) + \lambda \boldsymbol{a}^T K_E \boldsymbol{a}\,, \tag{4.1}$$

Derivando a equação acima em relação a $\boldsymbol{a}$ e a igualando a zero, chega-se a um sistema

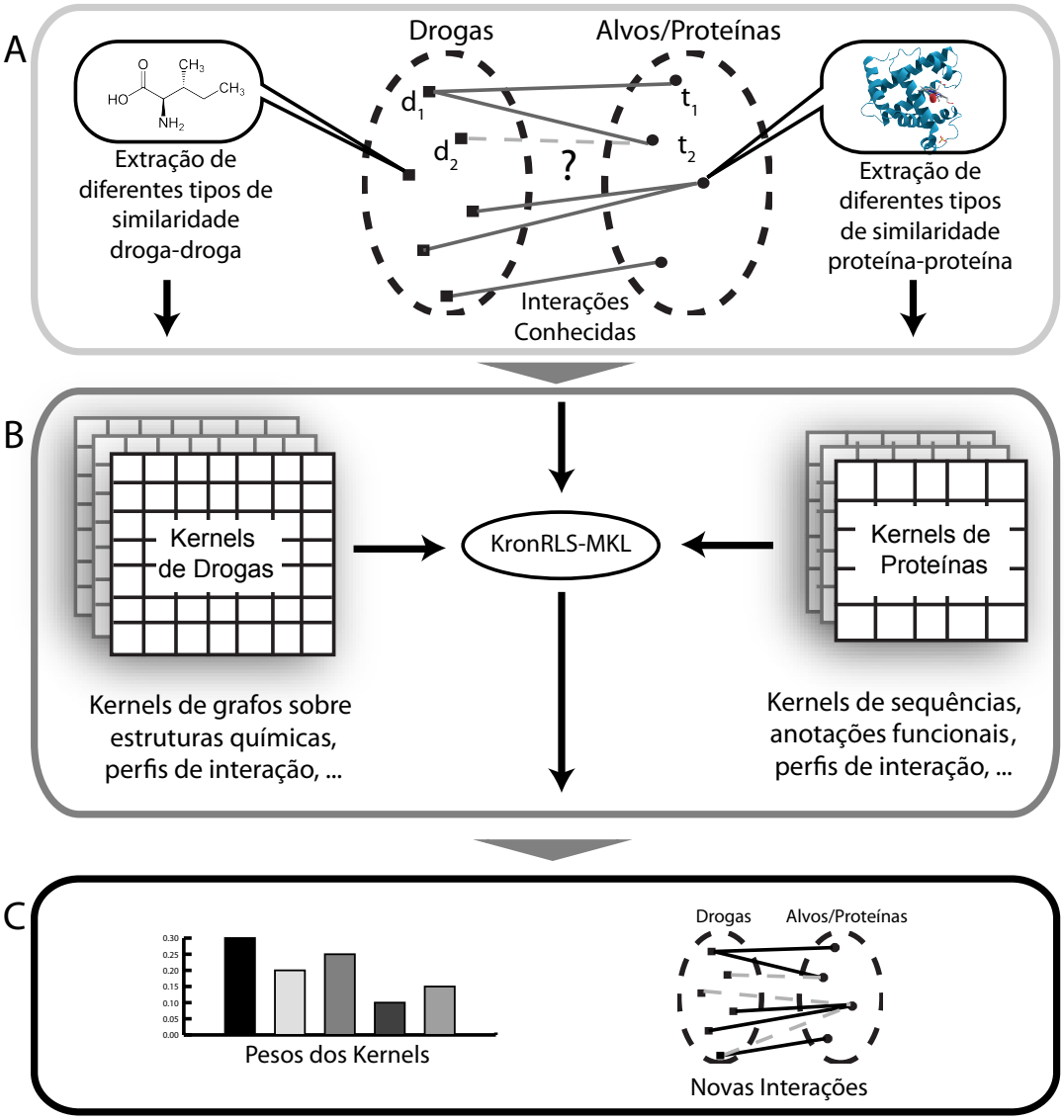

Figura 4.1: (A) A rede droga-proteína é representada como um grafo bipartido, com drogas (esquerda) e proteínas (direita). Arestas (linhas contínuas) entre drogas e proteínas indicam uma interação droga-proteína conhecida. A predição de novas interações é definida como a busca por arestas desconhecidas (linhas tracejadas), baseado na hipótese de que drogas (ou proteínas) semelhantes devem compartilhar os mesmos alvos. (B) O algoritmo proposto, KronRLS-MKL, utiliza um conjunto de kernels de drogas e proteínas para predizer novas interações. Diferentes kernels são obtidos comparando-se a similaridade entre drogas (ou proteínas), utilizando fontes de informação distintas. (C) O algoritmo KronRLS-MKL não só é capaz de indicar novas interações, como também identifica a relevância (peso) de cada kernel utilizado nas predições.

de equações lineares dada por (PAHIKKALA et al., 2013):

$$(K_E K_E + \lambda K_E)\boldsymbol{a} = K_E \boldsymbol{y} \tag{4.2}$$

$$(K_E + \lambda I)\boldsymbol{a} = \boldsymbol{y}, \tag{4.3}$$

a qual possui uma única solução uma vez que a matriz K_E é positiva semidefinida. PAHIKKALA et al. (2013) apresenta uma solução analítica para o caso em que o kernel K_E corresponde a um produto de Kronecker, utilizando duas propriedades algébricas do produto de Kronecker de duas matrizes: o *vec trick* e a relação entre a autodecomposição de um produto de Kronecker e a autodecomposição de seus fatores (PAHIKKALA et al., 2014; LAUB, 2005).

Dessa forma, seja $K_d = Q_d \Lambda_d Q_d^T$ e $K_p = Q_p \Lambda_p Q_p^T$ as autodecomposições das matrizes de kernel K_d e K_p. Utilizando as propriedades supracitadas e multiplicando ambos os lados da equação (4.3) por $(K_E + \lambda I)^{-1}$, chega-se a solução $\boldsymbol{a}$ (PAHIKKALA et al., 2013, 2014):

$$\boldsymbol{a} = vec(Q_p C Q_d^T)\ , \tag{4.4}$$

em que C é uma matriz definida como:

$$C = (\Lambda_D \otimes \Lambda_T + \lambda I)^{-1} vec(Q_T^T Y^T Q_D)\ . \tag{4.5}$$

4.2 O algoritmo KronRLS-MKL

Nesta seção, apresentamos uma extensão ao algoritmo KronRLS, chamada de KronRLS-MKL, capaz de incorporar múltiplos kernels a tarefa de predição de interações. Seja um conjunto de diferentes kernels de drogas e proteínas, i.e., $\boldsymbol{k}_d = (K_d^1, K_d^2, ..., K_d^{P_d})^T$ e $\boldsymbol{k}_p = (K_p^1, K_p^2, ..., K_p^{P_p})^T$, onde P_d e P_p indicam o número de kernels base definidos sobre o conjunto de drogas e proteínas, respectivamente.

Os kernels podem ser combinados por uma função linear, i.e., a soma poderada dos kernels base, correspondendo aos kernels ótimos K_d^* e K_p^*:

$$K_d^* = \sum_{i=1}^{P_d} \beta_{d,i}^* K_d^i \ \ , \ \ K_p^* = \sum_{j=1}^{P_p} \beta_{p,j}^* K_p^j.$$

DINUZZO (2011) demonstrou que métodos MKL podem ser interpretados como uma instância particular de uma máquina de kernel com múltiplas camadas, na qual a segunda camada é uma função linear, fornecendo o embasamento teórico para o desenvolvimento de uma extensão MKL para o algoritmo KronRLS.

A combinação ótima dos kernels base é então definida pelos vetores de pesos $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$. Uma vez que a função de classificação da equação 2.6 pode ser escrita em sua forma matricial, $f_a = K\boldsymbol{a}$ (RIFKIN; YEO; POGGIO, 2003), bem como aplicando a já conhecida propriedade do

produto de Kronecker, $(A \otimes B)vec(X) = vec(BXA^T)$(LAUB, 2005), temos:

$$\begin{aligned} f_a &= K\boldsymbol{a} \\ &= (K_d^* \otimes K_p^*)vec(Q_p C Q_d^T) \\ &= (K_p^*(Q_p C Q_d^T)(K_d^*)^T), \end{aligned}$$

onde $\boldsymbol{a} = vec(Q_p C Q_d^T)$ corresponde a solução dada pelo algoritmo KronRLS. Desse modo, podemos reescrever a função de classificação como $(K_d^* A^T (K_p^*)^T)$, em que $A = unvec(\boldsymbol{a})$ (o operador *unvec* é definido tal que $vec(unvec(\boldsymbol{x})) = \boldsymbol{x}$). A solução para o problema de otimização derivado da combinação de múltiplos kernels consiste em encontrar os valores ótimos dos coeficientes $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$, bem como do vetor de parâmetros do modelo, $\boldsymbol{a}$.

Utilizando uma abordagem interativa, considerada em estratégias MKL anteriores (GÖNEN; ALPAYDıN, 2011), propomos o uso de um processo de otimização de passo duplo, no qual a otimização do vetor $\boldsymbol{a}$ é intercalado com a otimização dos vetores de pesos dos kernels. Dados dois vetores de pesos iniciais, $\boldsymbol{\beta}_{d,0}$ e $\boldsymbol{\beta}_{p,0}$, um valor ótimo para o vetor $\boldsymbol{a}$, é encontrado utilizando a equação (2.8). Uma vez encontrado o vetor $\boldsymbol{a}$ ótimo para os vetores de peso iniciais, podemos proceder a busca pelos vetores de pesos $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$. Mais especificamente, a equação (2.5) pode ser redefinida quando $\boldsymbol{a}$ é fixo, e sabendo que $\| f \|_F^2 = \boldsymbol{a}^T K \boldsymbol{a}$ (RIFKIN; YEO; POGGIO, 2003; DINUZZO, 2011). Fazendo:

$$\mathbf{u} = \left(\boldsymbol{y} - \frac{\lambda \boldsymbol{a}}{2}\right) \therefore \boldsymbol{y} = \boldsymbol{u} + \frac{\lambda}{2}\boldsymbol{a},$$

e sabendo que,

$$\begin{aligned} (K + \lambda I)\boldsymbol{a} &= \boldsymbol{y} \\ K\boldsymbol{a} + \lambda \boldsymbol{a} &= \boldsymbol{y} \\ K\boldsymbol{a} &= \boldsymbol{y} - \lambda \boldsymbol{a}, \end{aligned}$$

De modo que podemos reescrever a equação (2.5):

$$\begin{aligned} J(f_a) &= \frac{1}{2n\lambda} \| \boldsymbol{y} - K\boldsymbol{a} \|_2^2 + \frac{1}{2}\lambda \mathbf{a}^T K \mathbf{a} \\ &= \frac{1}{2n\lambda} \| \boldsymbol{u} + \frac{\lambda}{2}\boldsymbol{a} - K\boldsymbol{a} \|_2^2 + \frac{1}{2}\lambda \ \boldsymbol{a}^T K \boldsymbol{a} \\ &= \frac{1}{2n\lambda} \| \boldsymbol{u} - K\boldsymbol{a} \|_2^2 + \frac{\lambda \boldsymbol{a}^T \boldsymbol{a}}{8n} + \frac{1}{2}\lambda \boldsymbol{a}^T K \boldsymbol{a} \\ &= \frac{1}{2n\lambda} \| \boldsymbol{u} - K\boldsymbol{a} \|_2^2 + \frac{\lambda \boldsymbol{a}^T}{2}\left(\frac{\boldsymbol{a}}{4n} + (\boldsymbol{y} - \lambda \boldsymbol{a})\right). \end{aligned}$$

Dado que o segundo termo não depende de K (e portanto não depende dos pesos dos kernels), e, sendo $\boldsymbol{y}$ e $\boldsymbol{a}$ fixos, podemos descartá-lo do procedimento de otimização dos pesos. Também

podemos converter $\boldsymbol{u}$ para uma forma matricial pela aplicação do operador *unvec*, i.e., $U = unvec(\boldsymbol{u})$, e também utilizar uma norma mais adequada para matrizes (a norma de Frobenius, $\| A \|_2 \leq \| A \|_F$ (LAUB, 2005)):

$$J = \frac{1}{2n\lambda} \| U - (K_d^* A^T (K_p^*)^T) \|_F^2 . \tag{4.6}$$

4.3 Combinação não esparsa de kernels

Note que não estamos interessados em uma seleção esparsa de kernels, como em DINUZZO (2011). Desse modo, introduzimos um termo de regularização baseado na norma l_2 para controlar a esparcidade (KLOFT et al., 2008) dos vetores de pesos. Uma vez que K_p^* e K_d^* correspondem aos kernels resultantes da soma ponderada dos kernels base com os pesos $\boldsymbol{\beta}_p$ e $\boldsymbol{\beta}_d$, respectivamente, a função objetivo pode ser escrita como:

$$J = \frac{1}{2\lambda n} \left\| U - \left(\sum_j \beta_{d,j} K_d^j \right) A^T \left(\sum_i \beta_{p,i} K_p^i \right)^T \right\|_F^2 + \sigma(\| \boldsymbol{\beta}_d \|_2 + \| \boldsymbol{\beta}_p \|_2), \tag{4.7}$$

onde σ corresponde ao coeficiente de regularização que controla a esparcidade dos vetores de pesos $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$.

4.3.1 Abordagem com combinação com pesos arbitrários (KRONRLS-MKL^{arb})

Uma regra de atualização iterativa dos pesos pode ser obtida através das derivações parciais da equação (4.7) em termos de $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$, com $\boldsymbol{\beta}_p$ e $\boldsymbol{a}$ fixo. Podemos definir $B_p \in \mathbb{R}^{n_p \times n_p}$ e $M_i \in \mathbb{R}^{n_d \times n_p}$:

$$B_p = \left(\sum_{j=1}^{P_p} \beta_{p,j} K_{p,j} \right)^T$$
$$M_i = K_{d,i} A^T B_p, \;\; 1 \leq i \leq P_d.$$

Calculando-se a derivada $dJ/d\beta_{d,s}, 1 \leq s \leq P_d$, temos,

$$\frac{dJ}{d\beta_{d,s}} = \left[U - \sum_i^{P_d} \beta_{d,i} M_i\right] : [-M_s] + \sigma\beta_{d,s} \tag{4.8}$$

$$= \left[\sum_i^{P_d} \beta_{d,i} M_i - U\right] : [M_s] + \sigma\beta_{d,s} \tag{4.9}$$

$$= \sum_i^{P_d} \beta_{d,i} M_i : M_s - U : M_s + \sigma\beta_{d,s}, \tag{4.10}$$

em que $M : N = \sum M_{ij} N_{ij} = vec(M)^T vec(N) = tr(M^T N)$ corresponde ao produto de Frobenius. A equação acima pode ser reescrita em termos do um vetor $\boldsymbol{v} = \{v_i\} \in \mathbb{R}^{P_d}$, cujos elementos são dados por:

$$v_i = U : M_i,$$

e da matriz simétrica $X^d \in \mathbb{R}^{P_d \times P_d}$, definida como:

$$X_d(s,i) = M_i : M_s.$$

Fazendo $dJ/d\beta_{d,s} = 0$, temos:

$$X_d \boldsymbol{\beta}_d - \boldsymbol{v} + \sigma\boldsymbol{\beta}_d = 0 \tag{4.11}$$

$$(X_d + \sigma I_{P_d})\boldsymbol{\beta}_d = \boldsymbol{v} \tag{4.12}$$

$$\boldsymbol{\beta}_d = (X_d + \sigma I_{P_d})^{-1} \boldsymbol{v}. \tag{4.13}$$

De maneira análoga, podemos obter a regra de atualização de $\boldsymbol{\beta}_p$:

$$B_d = \left(\sum_{i=1}^{P_d} \beta_{d,i} K_{d,i}\right)^T$$

$$M_j = B_d A^T (K_{p,j})^T$$

$$X_p(r,j) = M_j : M_r$$

$$\frac{dJ}{d\beta_{p,r}} = \left[U - \sum_{j=1}^{P_p} M_i \beta_{p,j}\right] : [-M_r] + \sigma\beta_{p,r}$$

$$= \left[\sum_{j=1}^{P_p} \beta_{p,j} M_j - U\right] : [M_r] + \sigma\beta_{p,r}$$

$$= \sum_{j=1}^{P_p} \beta_{p,j} M_j : M_r - U : M_r + \sigma\beta_{p,r},$$

em que $1 \leq r \leq P_p$ e $X_p \in \mathbb{R}^{P_p \times P_p}$. Dessa forma, temos:

$$\boldsymbol{\beta}_p = (X_p + \sigma I_{P_p})^{-1} \boldsymbol{v}. \tag{4.14}$$

Os pesos finais são então submetidos a uma normalização, i.e., $\boldsymbol{\beta}_{\cdot,i} = \boldsymbol{\beta}_{\cdot,i} / \sum_i \boldsymbol{\beta}_{\cdot,i}$. O Algoritmo 1 descreve os passos do método proposto para predição de interações via combinação de kernels com pesos arbitrários.

Algoritmo 1: Algoritmo para predição de interações com combinação arbitrária de kernels base (KRONRLS-MKL^{arb}).

início

Inicialização uniforme dos pesos $\boldsymbol{\beta}_{d,0}$ e $\boldsymbol{\beta}_{p,0}$;

Calcular combinações de kernels base com pesos iniciais

$K_d = \boldsymbol{k}_d^T \boldsymbol{\beta}_{d,0}$

$K_p = \boldsymbol{k}_p^T \boldsymbol{\beta}_{p,0}$

Calcular $\boldsymbol{a}$ segundo a Eq. (2.8);

while *Critério de parada* **do**

Com $\boldsymbol{a}$ e $\boldsymbol{\beta}_p$ fixos, encontrar a solução $\boldsymbol{\beta}'_d$ da Eq. 4.13 ;

Com $\boldsymbol{a}$ e $\boldsymbol{\beta}_d$ fixos, encontrar a solução $\boldsymbol{\beta}'_p$ da Eq. 4.14 ;

Calcular combinações de kernels K'_d e K'_p com os pesos $\boldsymbol{\beta}'_d$ e $\boldsymbol{\beta}'_p$;

Calcular $\boldsymbol{a}$ utilizando K'_d e K'_p, segundo a Eq. 2.8;

Normalizar $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$;

end

Fazer a decomposição em autovalores $K'_d = Q_d \Lambda_d Q_d^T$ e $K'_p = Q_p \Lambda_p Q_p^T$;

$A = unvec(\boldsymbol{a})$;

Scores das novas interações é dado por $F = Q_d A^T Q_p^T$;

fim

4.3.2 Abordagem com combinação convexa dos pesos (KRONRLS-MKL^{conv})

A fim de obter uma combinação convexa dos pesos dos kernels base, i.e., $\sum_{i=1}^{P_\cdot} \beta_{\cdot,i} = 1,\ 0 \leq \beta_{\cdot,i} \leq 1$, optamos por uma formulação genérica, que permita a utilização de pacotes de otimização tradicionais. Dessa forma, para quaisquer valores fixos de $\boldsymbol{a}$ e $\boldsymbol{\beta}_p$, o valor ótimo para os vetores de pesos é obtido solucionando o problema de otimização definido como:

$$\begin{aligned} \min_{\boldsymbol{\beta}_d} \quad & J_{\boldsymbol{\beta}_d} = \frac{1}{2\lambda n} \parallel U - \boldsymbol{\beta}_d \boldsymbol{m}_d \parallel_F^2 + \sigma \parallel \boldsymbol{\beta}_d \parallel_2^2 \\ \text{s.t.} \quad & \sum_{i=1}^{P_d} \beta_{d,i} = 1,\ 0 \leq \beta_{d,i} \leq 1 \\ \boldsymbol{m}_d = & (K_{d,1} A^T (K_p^*)^T, K_{d,2} A^T (K_p^*)^T, ..., K_{d,P_d} A^T (K_p^*)^T) \end{aligned} \tag{4.15}$$

enquanto o $\boldsymbol{\beta}_p$ ótimo pode ser encontrado fixando os valores de $\boldsymbol{a}$ e $\boldsymbol{\beta}_d$, de acordo com:

$$\begin{aligned} \min_{\boldsymbol{\beta}_p} \; & J_{\boldsymbol{\beta}_p} = \frac{1}{2\lambda n} \parallel U - \boldsymbol{m}_p \boldsymbol{\beta}_p \parallel_F^2 + \sigma \parallel \boldsymbol{\beta}_p \parallel_2^2 \\ \text{s.t.} \; & \sum_{i=1}^{P_p} \beta_{p,i} = 1, \; 0 \leq \beta_{p,i} \leq 1 \\ \boldsymbol{m}_p = & (K_d^* A^T (K_{p,1})^T, K_d^* A^T (K_{p,2}), ..., K_d^* A^T (K_{p,P_p})^T) \end{aligned} \tag{4.16}$$

onde K_d^* e K_p^* correspondem a combinação dos vetores de kernel $\boldsymbol{k}_d$ e $\boldsymbol{k}_p$ conforme os pesos $\boldsymbol{\beta}_d$ e $\boldsymbol{\beta}_p$, respectivamente.

Uma vez que o problema de otimização encontra-se definido, este pode ser resolvido por um dos diferentes pacotes de otimização convexa disponíveis. Nesta tese, optamos por utilizar o pacote de otimização da ferramenta Matlab (MATLAB, 2013). Foi adotado o algoritmo de otimização de ponto interior, uma vez que é capaz de solucionar problemas de grande dimensionalidade com eficiência (BYRD; HRIBAR; NOCEDAL, 1999). O método incorpora duas poderosas ferramentas para resolução de problemas com restrições: a otimização baseada nas condições de KKT e multiplicadores de Lagrange e estratégias de região de confiança, para tratar problemas de otimização convexa e não convexa (BYRD; HRIBAR; NOCEDAL, 1999). O algoritmo proposto (Algoritmo 2) distingue-se da abordagem proposta na seção anterior no que se refere a abordagem utilizada para solução do problema de otimização.

Algoritmo 2: Algoritmo de otimização alternada para combinação convexa dos pesos dos kernels (KRONRLS-MKL^{conv}).

início

Inicialização uniforme dos pesos $\boldsymbol{\beta}_{d,0}$ e $\boldsymbol{\beta}_{p,0}$;

Calcular combinações de kernels base com pesos iniciais

$K_d = \boldsymbol{k}_d^T \boldsymbol{\beta}_{d,0}$

$K_p = \boldsymbol{k}_p^T \boldsymbol{\beta}_{p,0}$

Calcular $\boldsymbol{a}$ segundo a Eq. (2.8);

while *Critério de parada* **do**

Com $\boldsymbol{a}$ e $\boldsymbol{\beta}_p$ fixos, encontrar a solução $\boldsymbol{\beta}'_d$ do prob. de otimização 4.15 ;

Com $\boldsymbol{a}$ e $\boldsymbol{\beta}_d$ fixos, encontrar a solução $\boldsymbol{\beta}'_p$ do prob. de otimização 4.16 ;

Calcular combinações de kernels K'_d e K'_p com os pesos $\boldsymbol{\beta}'_d$ e $\boldsymbol{\beta}'_p$;

Calcular $\boldsymbol{a}$ utilizando K'_d e K'_p, segundo a Eq. 2.8;

end

Fazer a decomposição em autovalores $K'_d = Q_d \Lambda_d Q_d^T$ e $K'_p = Q_p \Lambda_p Q_p^T$;

$A = unvec(\boldsymbol{a})$;

Scores das novas interações é dado por $F = Q_d A^T Q_p^T$;

fim

4.4 Considerações Finais

Neste capítulo foi apresentado um método de combinação de kernels para problemas de predição de interações em redes droga-proteína. Este tipo de rede é modelada na forma de um grafo bipartido, no qual interações existem apenas entre conjuntos disjuntos de vértices, i.e., entre drogas e proteínas. Apesar do fato de que muitos métodos de predição de interações já terem sido propostos na literatura, poucos trabalhos realizam a integração de múltiplas fontes de informação. Como foi apresentado no Capítulo 3, este tipo de fusão de dados pode ser obtida através da utilização do MKL em combinação com um método de kernel para a tarefa de predição.

O método proposto explora duas deficiências em trabalhos anteriores, propondo uma extensão ao algoritmo KronRLS, um método eficiente para predição de interações e que até então estava limitado ao uso de kernels simples, para ser utilizado em um contexto de MKL. A utilização do algoritmo KronRLS como aprendiz base reduz o impacto do uso do kernel de pares, normalmente observado em métodos MKL baseados em SVM. O método proposto não requer a sub-amostragem de padrões negativos, e pode ser utilizado sobre todo o espaço quimiogenômico, podendo ser aplicado diretamente a redes com milhares de interações. Os kernels base são selecionados automaticamente, de forma paralela ao treinamento do modelo.

5

Experimentos e resultados

Neste capítulo são apresentados os procedimentos adotados para validação experimental dos modelos propostos. A Seção 5.1 apresenta uma visão geral dos experimentos realizados, incluindo informações sobre os conjuntos de interações droga-proteína conhecidas e respectivos kernels (Seção 5.1.1), os métodos competidores utilizados como referência (Seção 5.1.2), bem como o desenho experimental adotado para avaliação dos resultados (Seção 5.1.3). Os resultados obtidos são descritos na Seção 5.2, incluindo análises da acurácia e custo computacional em termos de tempo de processamento e consumo de memória. Por fim, as considerações finais são apresentadas na Seção 5.3.

5.1 Experimentos

A fim de avaliar o desempenho dos métodos propostos, foram realizados experimentos sistemáticos simulando o processo de predição de novas interações, a partir de um conjunto de interações droga-proteína já conhecidas. O método proposto foi comparado tanto com o próprio algoritmo KronRLS (considerando apenas um kernel para cada tipo de entidade), bem como com outras estratégias de combinação de informações propostas na literatura.

5.1.1 Dados

Os experimentos foram realizados com um conjunto de bases de dados de interação droga-proteína já considerada em estudos anteriores (LAARHOVEN; NABUURS; MARCHIORI, 2011; PAHIKKALA et al., 2014; DING et al., 2013; WANG et al., 2011), proposta inicialmente por YAMANISHI et al. (2008). Cada base consiste em uma matriz binária, contendo todas as interações conhecidas de um determinado grupo de alvos moleculares, i.e., enzimas (E), canais de íons (IC), receptores acoplados a proteínas G (GPCR) e receptores nucleares (NR). As interações foram extraídas do KEGG BRITE (KANEHISA et al., 2008), SuperTarget (GüNTHER et al., 2008) e do DrugBank (WISHART et al., 2008). Todas as bases de dados são extremamente desbalanceadas, se considerarmos todo o espaço de interações possível entre o conjunto de drogas e proteínas, como demonstrado na Tabela 5.1.

Tabela 5.1: Número de drogas, proteínas e instâncias positivas (interações conhecidas) vs. o número de instâncias negativas (ou desconhecidas) em cada base de dados.

	Base de dados			
	Receptores Nucleares	GPCR	Canais de Íons	Enzimas
Interações				
Conhecidas	90	635	1476	2926
	(6.41%)	(3%)	(3.45%)	(1%)
Desconhecidas	1314	20550	41364	292554
	(93.59%)	(97%)	(96.55%)	(99%)
Entidade				
Drogas	54	223	210	445
Proteínas	26	95	204	664

5.1.1.1 Kernels Base

Cada tipo de entidade pode ser analisada sob diferentes pontos de vista, e.g., proteínas podem ser comparadas em relação a sua sequência de amino ácidos, função biológica, ou até mesmo ao seu perfil de expressão gênica. Consideramos um total de 20 fontes de informação distintas (10 para proteínas e 10 para drogas, ver Tabela 5.2), as quais foram avaliadas em relação à sua relevância na tarefa de predição de novas interações. As similaridades finais droga-droga (K_d) e proteína-proteína (K_p) são representadas como matrizes de kernel, pela aplicação dos diferentes tipos de medidas de similaridade químicas e genômicas:

$$K_d = \begin{bmatrix} K_d(d_1,d_1) & \cdots & K_d(d_1,d_{n_d}) \\ \vdots & \ddots & \vdots \\ K_d(d_1,d_{n_d}) & \cdots & K_d(d_{n_d},d_{n_d}) \end{bmatrix} \qquad K_p = \begin{bmatrix} K_p(p_1,p_1) & \cdots & K_p(p_1,p_{n_p}) \\ \vdots & \ddots & \vdots \\ K_p(p_1,p_{n_p}) & \cdots & K_p(p_{n_p},p_{n_p}) \end{bmatrix}$$

A Tabela 5.2 lista todos os kernels utilizados neste trabalho, a sua fonte de informação e os parâmetros utilizados.

Quando não especificado de outra forma, todos os kernels foram normalizados utilizando a equação abaixo:

$$K_{norm}(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}. \tag{5.1}$$

Em alguns casos, a medida de similaridade adotada não produz um matriz de kernel válida (i.e., PSD). Nesses casos a matriz de similaridade é convertida em uma matriz de kernel somando ela a uma matriz diagonal composta de pequenos múltiplos do menor autovalor da matriz original, i.e., $K + \delta\lambda_{min}I$ (SCHOLKOPF; SMOLA, 2001), em que utilizamos $\delta = 0.0001$ nos experimentos realizados.

Tabela 5.2: Kernels utilizados para drogas e proteínas considerados, e suas respectivas fontes de informação

Entidade	Kernels	Fonte de informação
Drogas	AERS-bit - AERS bit	Efeitos Colaterais
	AERS-freq - AERS freq	Efeitos Colaterais
	GIP - Gaussian Interaction Profile	Network
	LAMBDA - Lambda-k Kernel	Estrut. Químicas
	MARG - Marginalized Kernel	Estrut. Químicas
	MINMAX - MinMax Kernel	Estrut. Químicas
	SIMCOMP - Graph kernel	Estrut. Químicas
	SIDER - Similaridade de efeitos colaterais	Efeitos Colaterais
	SPEC - Spectrum Kernel	Estrut. Químicas
	TAN - Tanimoto Kernel	Estrut. Químicas
Proteínas	GIP - Gaussian Interaction Profile	Network
	GO - Similaridade semântica no Gene Ontology	Anotações funcionais
	MIS-k3m1 - Mismatch kernel ($k = 3, m = 1$)	Sequência genômica
	MIS-k4m1 - Mismatch kernel ($k = 4, m = 1$)	Sequência genômica
	MIS-k3m2 - Mismatch kernel ($k = 3, m = 2$)	Sequência genômica
	MIS-k4m2 - Mismatch kernel ($k = 4, m = 2$)	Sequência genômica
	PPI - Proximidade na rede proteína-proteína	Interações proteína-proteína
	SPEC-k3 - Spectrum kernel ($k = 3$)	Sequência genômica
	SPEC-k4 - Spectrum kernel ($k = 4$)	Sequência genômica
	SW - score no alinhamenhto Smith-Waterman	Sequência genômica

Kernels de Proteínas

Utilizamos as seguintes fontes de informação acerca de proteínas: sequência de aminoácidos, anotações funcionais e proximidade na rede proteína-proteína (PPI). Como kernels de sequências, consideramos o score no alinhamento de sequências segundo o algoritmo de alinhamento Smith-Waterman (SW) (SMITH; WATERMAN, 1981), quatro diferentes combinações de parâmetros do kernel Mismatch (MIS) (LESLIE; WESTON; NOBLE, 2002) e duas parametrizações do kernel Spectrum (SPEC) (LESLIE; ESKIN; NOBLE, 2002), todos calculados sobre as sequências de aminoácidos extraídas da base de dados KEGG (KANEHISA et al., 2008). A similaridade funcional foi obtida com base nos termos associados na base de dados *Gene Ontology* (GO) (CONSORTIUM; OTHERS, 2004), enquanto a similaridade na rede PPI foi calculada sobre a rede de interações proteína-proteína obtida da base de dados Bio GRID (STARK et al., 2006). Apresentamos a seguir uma breve explanação de cada kernel.

Smith-Waterman (SW)

O algoritmo Smith-Waterman (SMITH; WATERMAN, 1981) é um dos principais algoritmos de alinhamento local entre duas sequências. O algoritmo utiliza programação dinâmica para comparar segmentos de todos os tamanhos possíveis entre um par de sequências, otimizando a medida de similaridade obtida. Uma grande vantagem deste método é o fato de que é garantido que um alinhamento local ótimo seja obtido, dadas duas sequências de aminoácidos. Neste trabalho, utilizamos a matriz de kernel obtida por YAMANISHI et al. (2008) e dispobibilizada em http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

Spectrum (`SPEC`*)*

Seja o conjunto de todas as substrings de tamanho k (k-mers) encontradas na sequências de aminoácidos das proteínas p e p'. O kernel Spectrum efetua uma busca, contabilizando o número de ocorrências de cada k-mer na sequência, de modo que o espaço de busca cresce exponencialmente com k, ou seja, seu tamanho é dado por $|\mathscr{A}|^k$, onde $|\mathscr{A}|$ é o tamanho do alfabeto ($|\mathscr{A}| = 21$ para sequências de aminoácidos).

Para cada k-mer m, o algoritmo calcula o seu respectivo número de ocorrências, $N(m,p)$ e $N(m,p')$, de maneira que é obtido um vetor de características para cada sequência, e.g., $\phi_{SPEC}(p) = \{N(m,p)\}_{m\in\mathscr{A}^k}$ e $\phi_{SPEC}(p') = \{N(m,p')\}_{m\in\mathscr{A}^k}$. Em seguida, a similaridade entre p e p' é dada pelo produto interno dos vetores de características obtidos, i.e., $K_{SPEC}(p,p') = \langle \phi_{SPEC}(p), \phi_{SPEC}(p') \rangle$.

Neste trabalho, consideramos dois valores possíveis para o número de k-mers, $k = 3$ (`SPEC-k3`) e $k = 4$ (`SPEC-k4`), calculados com a implementação disponível no pacote KeBABS (PALME; BODENHOFER, 2014) da ferramenta estatística R.

Mismatch (`MIS`*)*

O kernel Mismatch (LESLIE; WESTON; NOBLE, 2002) é similar ao kernel Spectrum, de modo que o parâmetro k possui o mesmo significado. A distinção se dá pelo fato de que o kernel Mismatch permite um número de caracteres $m < k$ diferentes em cada k-mer (*mismatches*). Observe que, para $m = 0$, obtemos o kernel Spectrum apresentado anteriormente. O cálculo do kernel também é feito de maneira análoga à realizada no kernel Spectrum (produto interno dos vetores de características). Avaliamos quatro combinações de valores para parâmetros k e m: $k = 3$ e $k = 4$, e $m = 1$ e $m = 2$, chamados de `MIS-k3m1`, `MIS-k3m2`, `MIS-k4m1` e `MIS-k4m2`.

Similaridade semântica de anotações funcionais (`GO`*)*

O *Gene Ontology* (GO) é um vocabulário controlado de genes e respectivos produtos, para descrever características e sintetizar o conhecimento existente sobre tais entidades. Os termos são especificados em uma linguagem formal, o que o torna legível tanto para seres humanos quanto para máquinas. O GO é composto de três ontologias, que capturam diferentes aspectos da biologia celular: processos biológicos (*biological processes* - BP), componentes celulares (*cellular components* - CC) e funções moleculares (*molecular functions* - MF). Cada ontologia é composta por um conjunto de termos e suas respectivas relações hierárquicas.

O formalismo estruturado utilizado no GO permite que sejam definidas medidas de similaridade semântica entre genes, baseado no princípio de que genes similares compartilham anotações GO. Diversas medidas deste tipo podem ser encontradas na literatura (GUZZI et al., 2012), dentre as quais uma das mais simples e utilizadas é a medida de similaridade de Resnik (RESNIK, 1999; GUZZI et al., 2012). Dados dois genes e os seus respectivos termos GO associados, a similaridade de Resnik inicia identificando termos menos frequentes, i.e., mais específicos, de maneira que é dada maior importância a termos menos frequentes. Mais especificamente, dado um corpus de anotações de um determinado gene, a quantidade de informação (*information content* - *IC*) de um termo c é definido como:

$$IC(c) = -log(f_{anot}(c)), \tag{5.2}$$

onde $f_{anot}(c)$ é a fração de genes que são anotados com o termo c, ou seus descendentes. Dessa forma, a similaridade de Resnik (Sim_{Res}) é definida como:

$$Sim_{Res}(p,p') = IC[argmax(IC(p''))], \tag{5.3}$$

$$p'' \in ancestrais(p,p'). \tag{5.4}$$

Para o cálculo da similaridade semântica, extraímos os termos GO da base de dados BioMART (SMEDLEY et al., 2015) para cada proteína dos conjuntos NR, GPCR, IC e Enzimas. A similaridade semântica entre proteínas, foi realizada com a implementação do algoritmo de Resnik no pacote csbl.go (OVASKA; LAAKSO; HAUTANIEMI, 2008) da ferramenta estatística R.

Proximidade na rede proteína-proteína (`PPI`*)*
A rede de interações proteína-proteína em seres humanos foi obtida da base de dados BioGRID (STARK et al., 2006). A similaridade entre cada par de proteínas foi calculado baseado na menor distância na rede PPI correspondente, de acordo com:

$$S(p,p') = Ae^{bD(p,p')},$$

onde os parâmetros A e b foram ajustados como em PERLMAN et al. (2011) ($A = 0.9, b = 1$), e $D(p,p')$ é a menor distância (número de saltos) entre as proteínas p e p'.

Kernels de drogas

Dentre um total de 10 kernels de drogas utilizados, seis deles baseiam-se na similaridade de estrutura química de cada droga, e três baseiam-se na similaridade farmacológica (i.e., efeitos-colaterais). Todos os kernels baseados em estruturas químicas consideram cada molécula como um grafo (2D), no qual os vértices correspondem aos átomos, e as arestas às ligações covalentes entre os átomos, sendo considerados kernels de grafos. Em geral, baseiam-se no conceito de fragmentos moleculares (*caminhos* no grafo ou subgrafos), i.e., uma sequência de átomos conectados por ligações, ou subárvores. Se considerarmos duas moléculas d e d', então o kernel K é dado por:

$$K(d,d') = \sum_{f \in \mathscr{F}} N(f,d) \cdot N(f,d'), \tag{5.5}$$

onde $\mathscr{F}$ é o conjunto de fragmentos moleculares, i.e., todos os possíveis caminhos ou subgrafos, e a função $N(f,d)$ representa a frequência observada do padrão f no grafo da molécula d. Em geral, os métodos utilizados diferenciam-se em relação a abordagem utilizada para calcular $N(f,d)$.

SIMCOMP
O algoritmo SIMCOMP (HATTORI et al., 2003) é um método baseado em grafos para comparação de estruturas químicas, que busca por cliques no subgrafo comum induzido máximo (MCIS). A busca pelo MCIS é um problema NP-difícil, que recebe um par de grafos como entrada e busca por um grafo de ordem máxima, i.e., maior número de nós) que seja isomorfo a um subgrafo induzido[1] de cada um

[1] Um subgrafo induzido consiste em um subconjunto de vértices de um dado grafo, associado a todas as arestas cujos nós também encontram-se neste subconjunto.

dos grafos de entrada. O algoritmo SIMCOMP adota diversas heurísticas para reduzir a dificuldade computacional da busca por cliques, bem como para aumentar a relevância biológica da similaridade utilizando testes de quiralidade[2] e diferenciando os tipos de átomos de cada nó. Uma vez encontrado o MCIS, o índice é convertido em uma medida de similaridade aplicando-se o coeficiente de Jaccard:

$$K(d,d') = \frac{|MCIS(d,d')|}{|d|+|d'|-MCIS(d,d')}.$$

Marginalizado (`MARG`)

No kernel Marginalizado (KASHIMA; TSUDA; INOKUCHI, 2003), a quantidade de vezes que o caminho f é encontrado no grafo da molécula d é ponderado pela probabilidade do caminho f ocorrer:

$$N_{MARG}(f,d) = \sum_{h \in \mathscr{H}(d)} w(h,d) \cdot \mathbf{1}(h=f),$$

onde $\mathscr{H}(d)$ é o conjunto de caminhos em d, e $w(h,d)$ é a probabilidade de que o caminho h ocorra em d e $\mathbf{1}(h=f)$ é 1 se o caminho h é igual a f, e zero caso contrário. Uma vez que o número de caminhos possíveis em $\mathscr{H}(d)$ pode ser infinito, a implementação adotada utiliza uma aproximação, na qual é limitado o tamanho máximo de cada caminho (max_n). A probabilidade de ocorrência de um caminho $w(h,d)$ também é limitada pela probabilidade de parada p_{stop}. O cálculo deste kernel foi realizado na ferramenta Rchemcpp da plataforma estatística R, com valores padrão dos parâmetros: $max_n = 3, p_{stop} = 0.1$.

Spectrum (`SPEC`)

O kernel Spectrum (MAHR; KLAMBAUER; HOCHREITER, 2012) aplica o mesmo princípio utilizado no kernel Spectrum para sequências (LESLIE; ESKIN; NOBLE, 2002), com a distinção que neste caso, as sequências consistem em caminhos no grafo. Dessa forma, a função $N(f,d)$ contabiliza quantas vezes o caminho f ocorre no grafo d.

$$N_{SPEC}(f,d) = \#\{f \in d\}.$$

Tanimoto (`TAN`)

O kernel Tanimoto (RALAIVOLA et al., 2005) por sua vez é uma versão binária do kernel Spectrum, ou seja:

$$N_{TAN}(f,d) = \begin{cases} 1, \text{ se } f \text{ ocorre em } d, \\ 0, \text{ caso contrário.} \end{cases}$$

Foi utilizada a versão normalizada, segundo o coeficiente de Jaccard:

$$K(d,d') = \frac{K(d,d')}{K(d,d)+K(d',d')-K(d,d')}.$$

[2]Quiralidade é uma propriedade de assimetria, i.e., um objeto é dito quiral se não pode ser sobreposto ao seu reflexo em um espelho (imagem especular).

Minimax (MINMAX)

Uma outra versão do kernel Spectrum é o chamado kernel MinMax (RALAIVOLA et al., 2005), onde é considerado o número mínimo e máximo de ocorrências do padrão f em d e d', ou seja:

$$K_{MINMAX}(d,d') = \frac{\sum_{f\in\mathscr{F}} min(N_{SPEC}(f,d),N_{SPEC}(f,d'))}{\sum_{f\in\mathscr{F}} max(N_{SPEC}(f,d),N_{SPEC}(f,d'))}.$$

Lambda-k (LAMBDA)

Assim como o kernel marginalizado, o kernel lambda-k (MAHR; KLAMBAUER; HOCHREITER, 2012) pondera a quantidade de ocorrências de um dado caminho f no grafo d por uma função do comprimento do caminho:

$$K_{LAMBDA}(d,d') = \lambda^{|f|} \cdot \#\{f \in d\}.$$

Observe que para $\lambda > 1$ a influência de caminhos mais longos é maior, enquanto para $\lambda < 1$ a medida de similaridade é mais influenciada por caminhos mais curtos.

Adverse Event Reporting System (AERS)

O *adverse event reporting system* (AERS) é uma base de dados compilada pelo FDA (*U.S. Food and Drug Administration*), que contém informações sobre reações adversas relacionadas ao uso de medicamentos submetidas ao FDA. TAKARABE et al. (2012) propôs utilizar os dados do AERS para extrair uma medida de similaridade baseada nas palavras-chave de reações associadas a cada medicamento. Os autores introduziram dois tipos de perfis farmacológicos para drogas: o primeiro baseado na frequência de cada palavra nas informações das reações para cada droga (AERS-freq), e o segundo baseado na presença ou ausência de um efeito colateral particular, levando a um perfil binário para cada droga (AERS-bit). Uma vez que nem todas as drogas presentes nas bases NR, IC, GPCR e Enzima estão também presentes nos dados AERS, foram extraídas as similaridades apenas para as drogas encontradas em tais dados, e foi atribuída similaridade zero as não existentes.

A medida de similaridade final foi obtida aplicando o coeficiente de correlação cosseno ponderada entre cada par de perfis (TAKARABE et al., 2012). A idéia básica é pôr mais ênfase em palavras menos frequentes do que em termos frequentemente encontrados em tais efeitos colaterais. Ou seja:

$$S_{EC}(d,d') = \frac{\sum_{k=1}^{K} w_k d_k d'_k}{\sqrt{\sum_{k=1}^{K} w_k d_k}\sqrt{\sum_{k=1}^{K} w_k d'_k}} \tag{5.6}$$

onde w_k é uma função que associa um peso a k-ésima palavra-chave, definida como $w_k = exp(-f_k^2/\sigma^2 h^2), k = 1,2,...,K$, onde f_k é a frequência associada a palavra k, σ é a média de $\{f_k\}_{k=1}^{K}$ e h é um parâmetro (neste trabalho, foi utilizado $h = 1$ (TAKARABE et al., 2012)).

SIDER

O SIDER [3] KUHN et al. (2010) é uma base de dados que contém informação sobre drogas comerciais e

[3] http://sideeffects.embl.de/

efeitos colaterais e reações adversas registradas. Cada droga é representada por um perfil binário, no qual a presença ou ausência de cada efeito colateral é representado por 1 e 0, respectivamente. A base cobre um total de 888 drogas e 1.450 efeitos colaterais, resultando em 62.269 pares droga-efeito colateral. Para obter a matriz de similaridade, também foi aplicado o coeficiente de correlação cosseno ponderada sobre os perfis binários.

Kernel de rede

A fim de incorporar informações sobre a topologia da rede droga-proteína ao modelo, aplicamos o kernel *Gaussian Interaction Profile* (GIP), tanto para drogas e proteínas. Utilizamos a mesma definição do kernel GIP apresentada em LAARHOVEN; NABUURS; MARCHIORI (2011), o qual baseia-se no princípio de que drogas (ou proteínas) que apresentam o mesmo padrão de interação na rede tem maior probabilidade de demonstrar comportamento semelhante em relação a novas proteínas (ou drogas).

Para cada droga d é extraído um perfil binário de interação na rede, i.e., o equivalente à linha d da matriz de adjacências Y, que chamaremos de $\boldsymbol{y}_d$. Este vetor é utilizado como um descritor da droga em questão, cujos elementos representam a presença ou ausência de interações conhecidas entre a droga d e todas as proteínas na base de treinamento. Em seguida, é aplicada uma função de kernel. Uma vez construídos os vetores de características, é aplicada uma função Gaussiana sobre cada par de vetores. Esta função produz um kernel, também chamado de kernel de função de base radial (RBF), definido como:

$$K_{GIP}(d,d') = exp(-\gamma_d \parallel \boldsymbol{y}_d - \boldsymbol{y}_{d'} \parallel^2),$$

onde γ_d é um parâmetro de largura, normalizado pelo número de interações médias de cada droga, ou seja:

$$\gamma_d = \frac{\widehat{\gamma_d}}{\frac{1}{n_d}\sum_{i=1}^{n_d} |\boldsymbol{y}_d|^2}.$$

Assim como em LAARHOVEN; NABUURS; MARCHIORI (2011), adotamos o valor $\widehat{\gamma_d} = 1$.

5.1.2 Métodos competidores

Comparamos os métodos propostos, KronRLS-MKL^{arb} e KronRLS-MKL^{conv}, com o desempenho do algoritmo KronRLS com pares de kernels simples, bem como com três abordagens integrativas para predição de interação droga-proteína propostas recentemente (Seção 3.4). Além disso, propomos a utilização de duas heurísticas distintas de combinação prévia, utilizadas em conjunto com dois classificadores base: o próprio algoritmo KronRLS e o algoritmo SVM. A escolha destes classificadores baseia-se no fato de que estes métodos já demonstraram superar abordagens de kernels simples propostas anteriormente, como perfil mais próximo, KRM e BLM (LAARHOVEN; NABUURS; MARCHIORI, 2011; DING et al., 2013).

O algoritmo KronRLS, recentemente indicado como melhor método para predição de interação entre pares de droga-proteína com um kernels base simples DING et al. (2013), foi utilizado para avaliar o desempenho de todas as possíveis combinações de kernels base da Tabela 5.2. Isto resultou em um total de $10 \times 10 = 100$ diferentes combinações. Os melhores pares de kernels foram selecionados de acordo com

dois critérios distintos: o par de kernels que obteve maior area sob a curva de precisão-sensitividade (*area under the precision recall curve* - AUPR) no conjunto de treinamento (SINGLE†), e, em uma abordagem mais otimista, considerando a maior AUPR no conjunto de teste (SINGLE*). Estes experimentos serão usados como métodos *baseline* na nossa avaliação.

Duas heurísticas distintas foram adotadas para combinar múltiplos kernels em um único kernel, também utilizadas como bases de comparação: a média de kernels base e a heurística de alinhamento de kernels (KA) (QIU; LANE, 2009). A média dos kernels base é obtida como $K_D^* = 1/P_D \sum_{i=1}^{P_D} K_D^i$ para drogas, e, analogamente, o mesmo pode ser feito para proteínas. KA é uma heurística para a estimação de pesos para cada kernel, introduzida por QIU; LANE (2009), baseado na noção de alinhamento de kernels (CRISTIANINI et al., 2002) (Equação 3.16)). Uma vez que tal combinação é obtida, os kernels de droga e proteína resultantes são então usados para treinar um classificador KronRLS. Estas estratégias são referenciadas como KRONRLS-MEAN e KRONRLS-KA para a média dos kernels base e a combinação baseada em KA, respectivamente.

Consideramos também a utilização de outro classificador base, o SVM, considerando um kernel de pares (PKM (JACOB; VERT, 2008; WANG et al., 2011)), conforme descrito na Seção 2.3.4. Como o algoritmo PKM básico não se aplica a múltiplos kernels, este é utilizado em conjunto com as heurísticas KA e média de kernels base, sendo referenciadas como PKM-KA e PKM-MEAN, respectivamente. Utilizamos ainda uma outra heurística, adotada anteriormente por WANG et al. (2011) e descrita na Seção 3.4.1, a qual referenciamos por PKM-MAX. Além dos métodos supracitados, consideramos ainda uma abordagem proposta como trabalhos futuros por WANG et al. (2011) aqui chamado de WANG-MKL (Seção 3.4.2) e o algoritmo *Similarity-based Inference of drug-TARgets* (SITAR) (PERLMAN et al., 2011).

5.1.3 Desenho Experimental

As matrizes de adjacência com as interações conhecidas de cada rede droga-proteína foram utilizadas como rótulos das ligações para treinamento dos modelos, onde uma interação conhecida foi considerada como um padrão positivo, e interações desconhecidas foram considerados padrões negativos. Os experimentos consistiram em 5 execuções (i.e., repetições) de um procedimento de validação cruzada 5-folds, no qual cada base de dados foi separada aleatoriamente em 5 conjuntos disjuntos de pares droga-proteína (folds), dos quais quatro (04) foram utilizados para treinamento do algoritmo, e hum (01) foi utilizado para teste. Alguns dos métodos competidores (SITAR, WANG-MKL e métodos baseados no PKM) requerem a sub-amostragem de instâncias negativas, por limitações de memória impostas pela construção da matriz de kernel de pares. Entretanto, a fim de obter uma visão mais concreta da capacidade e limitações de cada método, avaliamos o desempenho preditivo, o uso de memória e o custo computacional destes métodos também em um contexto sem sub-amostragem de pares negativos, sempre que os recursos computacionais disponíveis permitiram, para fins de comparação. A sub-amostragem é obtida selecionando-se aleatoriamente padrões negativos na mesma quantidade de padrões positivos (WANG et al., 2011; PERLMAN et al., 2011)

Desse modo, os experimentos foram realizados em dois contextos de treinamento distintos (Figura 5.1): (1) o treinamento dos métodos competidores SITAR, WANG-MKL, PKM-MEAN, PKM-KA e PKM-MAX, foi realizado segundo o procedimento proposto originalmente pelos autores destes métodos, com sub-amostragem de padrões negativos, i.e., selecionando padrões negativos aleatoriamente na mesma

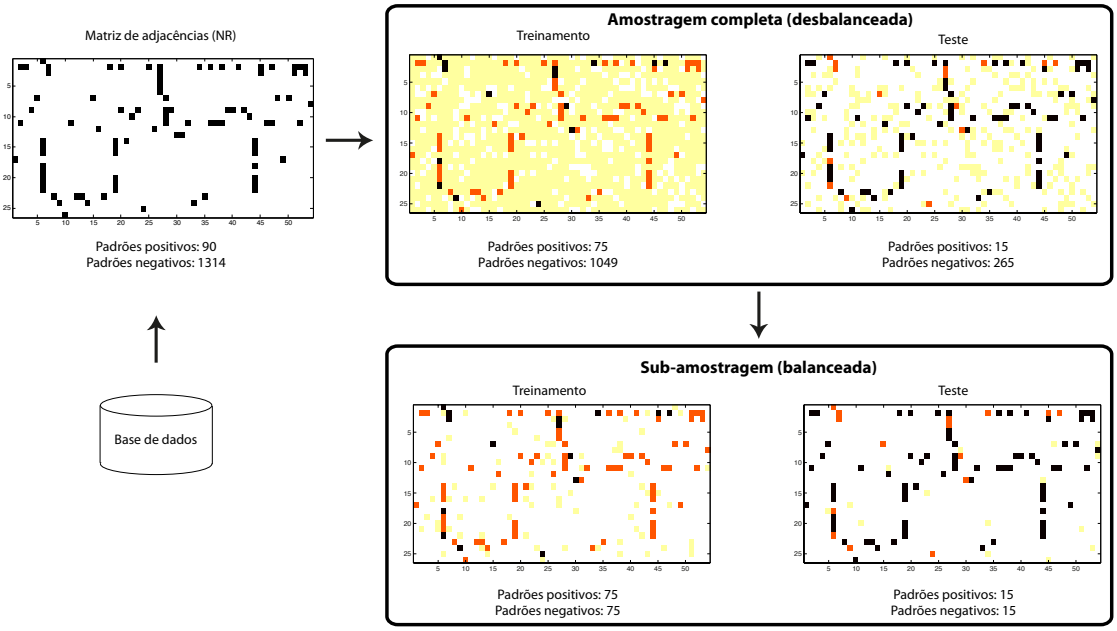

Figura 5.1: O processo de seleção das instâncias de treinamento e teste na validação cruzada. A partir das bases de dados, a matriz de adjacências da rede de interações é extraída, na qual linhas e colunas correspondem às drogas e proteínas, respectivamente, do conjunto de dados '*Nuclear Receptors*'. As posições preenchidas com preto correspondem às interações conhecidas, enquanto as preenchidas com branco correspondem às interações desconhecidas. Primeiramente, é elaborado o conjunto sem sub-amostragem, no qual as instâncias de treinamento e teste são selecionadas aleatoriamente sobre todo o espaço de interações, o que leva ao desbalanceamento das classes (indicadas pela cor laranja e amarela para as classes positiva e negativa, respectivamente), chamado Contexto 2. O conjunto de treinamento e teste do contexto (1) é selecionado sobre o mesmo conjunto de padrões positivos de (2), porém apenas um subconjunto das instâncias negativas de (2) são selecionadas, a fim de manter o balanceamento entre as classes positivas e negativas.

quantidade de padrões positivos. Dessa forma, foi produzido um conjunto de treinamento balanceado, e portanto, com dimensões inferiores ao utilizado no contexto (2); e (2) todos os modelos foram treinados com o conjunto de interações completo, i.e., toda a matriz de adjacências foi considerada como o conjunto de pares a ser particionado em treinamento e teste, o que inevitavelmente leva a folds com distribuição desbalanceada das classes. Os resultados no contexto (2) foram omitidos sempre que os requisitos de memória excederam 64 GB, ou cujo tempo computacional para uma execução do experimento excedeu 24 horas.

De maneira análoga, o procedimento de teste também foi realizado sob as duas perspectivas, i.e., com e sem sub-amostragem de padrões negativos, embora classes balanceadas sejam improváveis em cenários reais. Todos os experimentos foram executados no cluster de computação de alto desempenho da Universidade RWTH (Alemanha).

Os hiperparâmetros de cada método foram otimizados através de uma busca em conjuntos de valores pré-estabelecidos em um procedimento de validação cruzada interna. Este procedimento

ocorreu da seguinte maneira: durante a fase de treinamento de um dado fold, após a separação entre os conjuntos de treinamento $T1$ (04 folds) e teste $T2$ (01 fold), é realizado um novo processo de validação cruzada 3-fold considerando apenas o conjunto de treinamento ($T1$). Este procedimento de validação cruzada interna é repetido para cada valor avaliado do parâmetro em questão. Por fim, é selecionado o parâmetro que apresentou maior AUPR na validação cruzada interna. Para os métodos baseados em SVM (`PKM` e `WANG-MKL`), foi selecionado o parâmetro de custo C dentro do intervalo $\{2^{-1},...,2^{3}\}$; nos métodos baseados no algoritmo KronRLS, foi otimizado o valor do parâmetro de regularização do modelo, λ, dentro do conjunto $\{2^{-15},2^{-10},...,2^{30}\}$. No caso específico dos algoritmos `KRONRLS-MKL`^{arb} e `KRONRLS-MKL`^{conv}, também foi otimizado o valor do coeficiente σ de regularização dos pesos dos kernels, dentro do conjunto $\{0,0.25,0.5,0.75,1\}$, escolhido empiricamente. Uma vez selecionados os melhores hiperparâmetros para cada método, realizamos a construção do modelo, e seu desempenho é avaliado na partição de teste da validação cruzada externa.

A métrica de avaliação dos resultados foi a área sob a curva de precisão e sensitividade (*area under precision recall curve* - AUPR). A AUPR permite uma avaliação quantitativa do desempenho médio, e mede como a predição dos scores de interações positivas se separa dos scores das não-interações verdadeiras. Segundo DAVIS; GOADRICH (2006) e LAARHOVEN; NABUURS; MARCHIORI (2011), esta métrica apresenta maior qualidade para dados com alto grau de desbalanceamento nas classes, uma vez que pune a existência de falsos positivos entre os exemplos mais bem rankeados. Este é exatamente o caso, como demonstrado na Tabela 5.1, dado que todas as bases de dados são extremamente desbalanceadas.

Tabela 5.3: Matriz de confusão.

	Verdadeiros Positivos	Verdadeiros Negativos
Classif. como Positivos	TP	FP
Classif. como Negativos	FN	TN

Dada a matriz de confusão em um problema de classificação binária, contendo as seguintes informações: verdadeiros positivos (*True positives* - TP), i.e., os exemplos positivos classificados corretamente como positivos; falsos positivos (*False positives* - FP), i.e., os exemplos negativos classificados incorretamente como positivos; Verdadeiros negativos (*True negatives* - TN), i.e., os exemplos negativos classificados corretamente como negativos; e, finalmente, falsos negativos (*false negatives* - FN), indicando os exemplos positivos classificados incorretamente como negativos (Tabela 5.3). A taxa de precisão é definida como:

$$\text{Precisão} = \frac{TP}{TP+FP}, \tag{5.7}$$

e mede qual a fração de exemplos classificados como positivos que realmente o são. E a taxa de sensitividade (*Recall*) é dada por:

$$\text{Sensitividade} = \frac{TP}{TP+FN}, \tag{5.8}$$

a qual mede a fração de exemplos positivos que são corretamente classificados dentre o total de exemplos positivos. Dessa forma, podemos gerar um gráfico onde o eixo x corresponde a sensitividade , enquanto o eixo y corresponde a precisão, e assim obter a área obtida sob a curva gerada (AUPR).

5.2 Resultados

Os resultados obtidos com as diferentes estratégias de integração consideradas são apresentados nesta seção. A fim de avaliar os benefícios da combinação múltiplos kernels, apresentamos também os resultados obtidos em modelos construídos com o algoritmo KronRLS utilizando apenas um kernel para cada tipo de entidade (kernels simples), apresentados na Seção 5.2.1, os quais são usados como referência da qualidade individual de cada kernel. Em seguida, são apresentados os resultados dos experimentos com combinação de kernels (Seção 5.2.2). Uma avaliação da capacidade de predição utilizando versões atuais de bases de dados de interações droga-proteína é apresentada na Seção 5.2.4.

5.2.1 Pares de kernels simples

Como caso base, avaliamos o desempenho do algoritmo KronRLS sobre todos os pares de kernels base (10×10 combinações). Os resultados dos experimentos em cada uma das bases, para cada par de kernels, são apresentados nos *heatmaps* das Figuras 5.2 e 5.3 (valores em vermelho indicam maior AUPR)[4].

A Figura 5.4 apresenta *boxplots* com o desempenho médio de cada kernel, quando combinado com todos os outros kernels para cada entidade, nos quais podemos observar que o desempenho do algoritmo KronRLS varia drasticamente com a escolha dos kernels. Para a base de receptores nucleares (NR), o melhor par de kernels foi o `SPEC-k4` (proteínas) e o `GIP` (drogas), enquanto o `SW` (proteínas) e o `GIP` (drogas) o apresentaram melhor desempenho em todas as outras bases. Em geral, os kernels de drogas `AERS`, `SIMCOMP`, `GIP`, `MINIMAX` e `SIDER` apresentaram desempenho acima da média, enquanto os kernels `LAMBDA`, `MARG`, `SPEC` e `TAN` apresentaram os piores resultados. Para proteínas, os kernels `GIP`, `GO`, `MIS-k4m1`, `SPEC` e `SW` demonstraram melhor desempenho, enquanto os kernels `MIS-k3m2` e `PPI` foram inferiores em todas as bases. O custo computacional para execução da combinação exaustiva de pares de kernels base é moderada, até mesmo para um algoritmo de maior eficiência computacional como o KronRLS. O tempo para realizar todos os experimentos foi aproximadamente 8, 147, 385 e 12.516 minutos para as bases NR, GPCR, IC e Enzima, respectivamente.

5.2.2 Análise comparativa (Múltiplos kernels)

Nesta seção, comparamos os métodos competidores em termos de AUPR em todas as bases analisadas, considerando a combinação de todos os kernels base. Como descrito na Seção 5.1.3, esta análise é realizada sob duas condições distintas de treinamento dos métodos `SITAR`, `WANG-MKL`, `PKM-MEAN`, `PKM-KA` e `PKM-MAX`: (1) o treinamento destes métodos é realizado conforme proposto originalmente pelos autores, i.e., com sub-amostragem de padrões negativos, produzindo assim um conjunto de treinamento balanceado; e (2) todos os modelos foram treinados com o conjunto de interações completo, i.e., toda a matriz de adjacências foi considerada como o conjunto de pares a ser particionado em treinamento e teste, o que inevitavelmente leva a folds com distribuição desbalanceada das classes.

[4] As escalas de cores foram ajustadas a cada figura a fim de deixar as diferenças de desempenho entre os kernels mais explícitas.

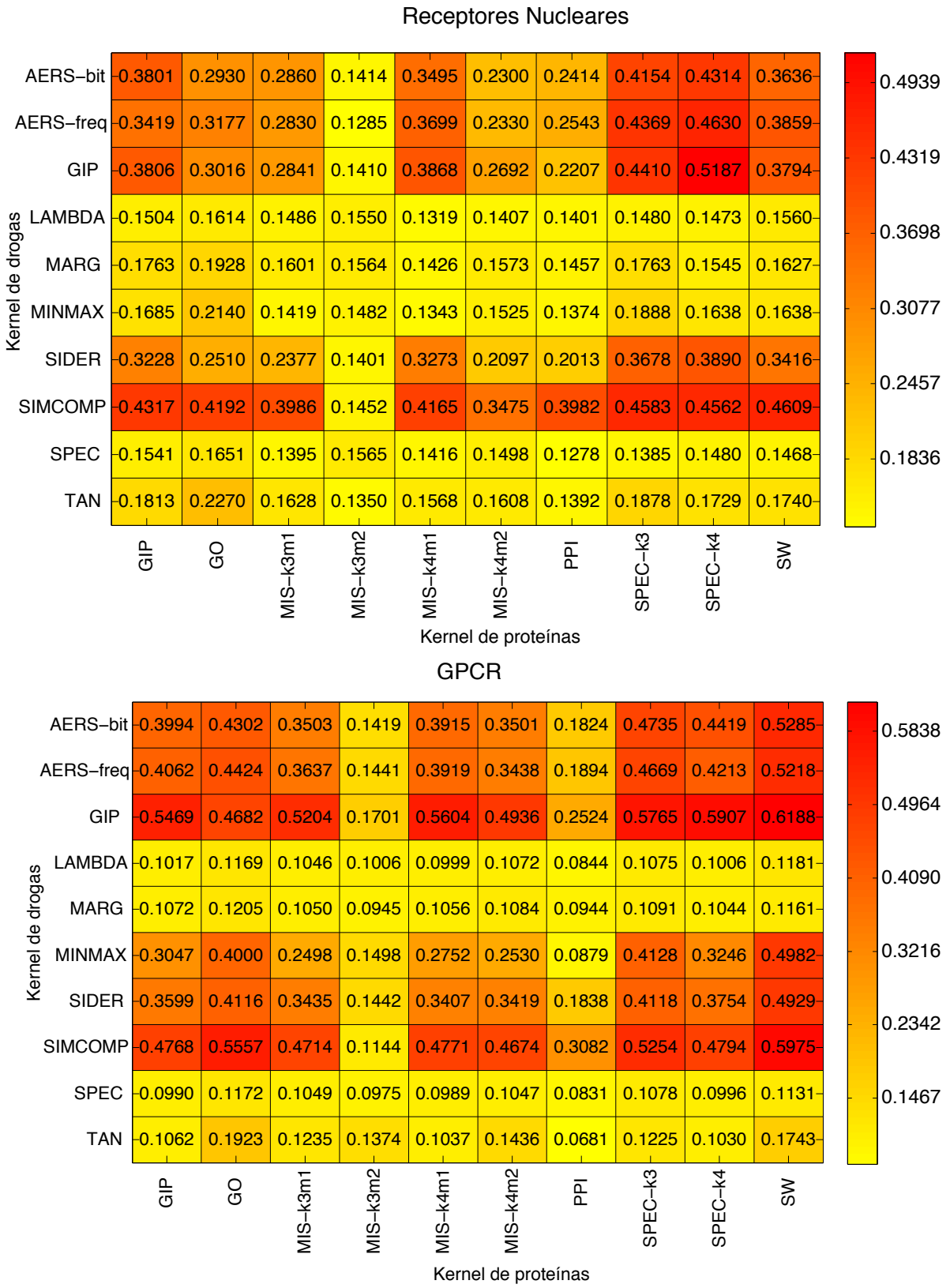

Figura 5.2: Desempenho individual (heatmap) de cada par de kernels (5 x 5-fold CV) nas bases NR e GPCR. Valores em vermelho indicam maior AUPR.

Para cada uma das condições acima, dois conjuntos de teste são considerados: um conjunto

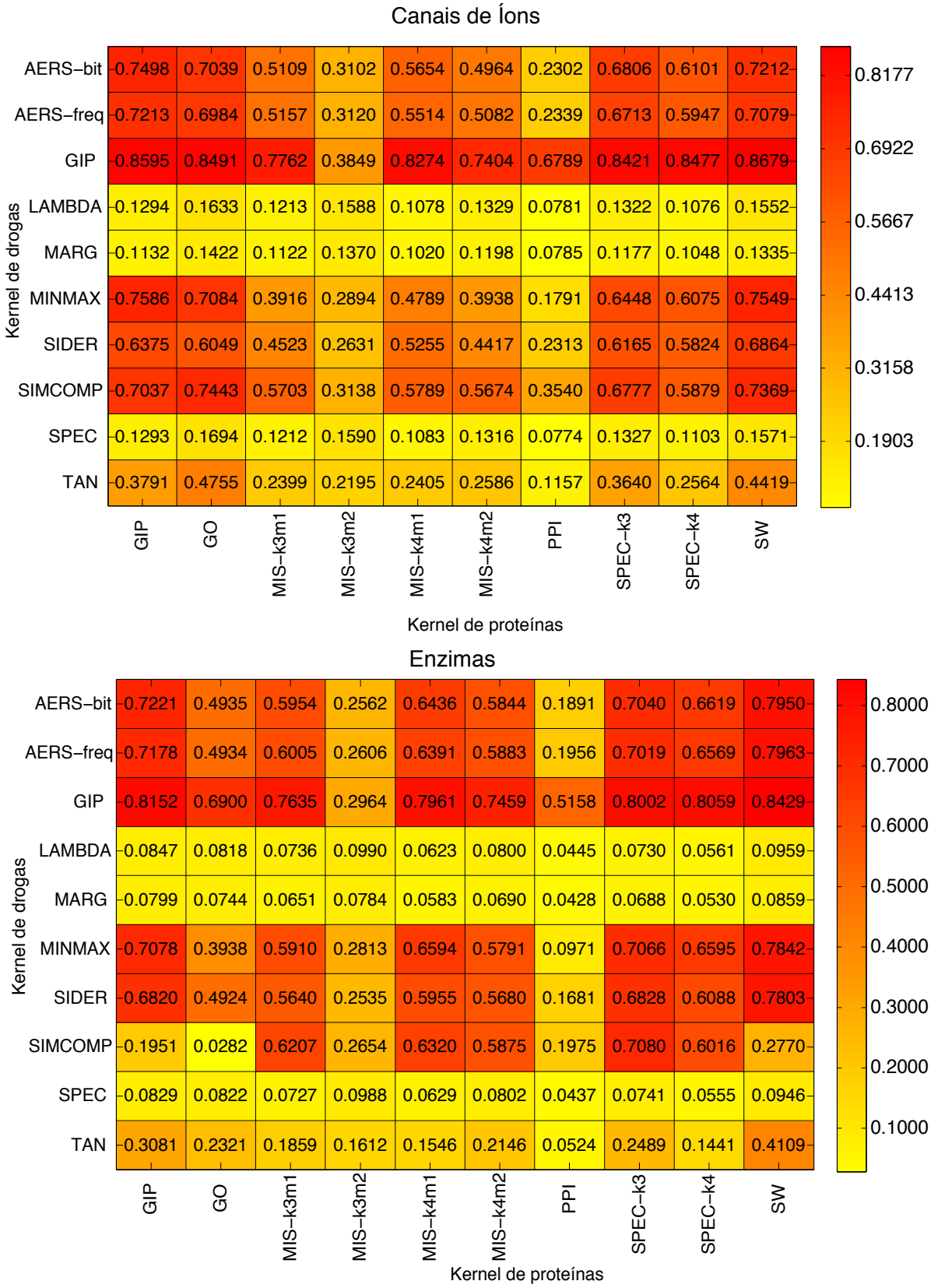

Figura 5.3: Desempenho individual (heatmap) de cada par de kernels (5 x 5-fold CV) nas bases IC e Enzima. Valores em vermelho indicam maior AUPR.

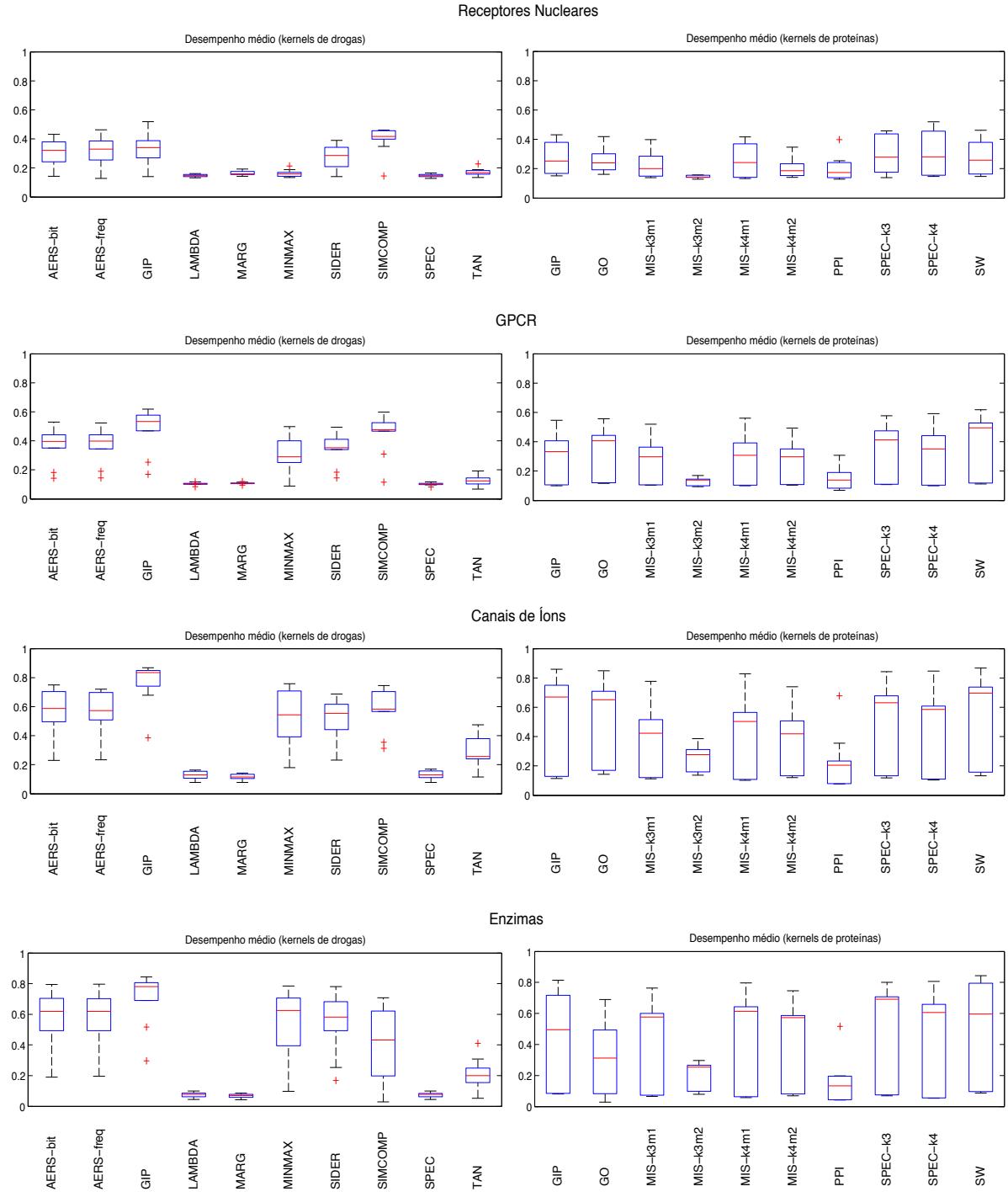

Figura 5.4: Boxplots do desempenho médio (AUPR) de cada de kernels.

completo, utilizando uma amostra desbalanceada, e também um segundo conjunto obtido pela amostragem balanceada dos pares no conjunto de testes completo.

5.2.2.1 Contexto 1

Neste contexto, os métodos `PKM-KA`, `PKM-MEAN`, `PKM-MAX`, `WANG-MKL` e `SITAR` foram treinados com um conjunto de treinamento balanceado, i.e., um número igual ao de interações conhecidas

foi selecionado aleatoriamente dentre as interações desconhecidas. Este cenário, apesar de irreal, é o adotado por todos esses métodos em trabalhos anteriores, dadas as limitações relativas ao cálculo explícito do kernel de pares. Conforme foi descrito na Seção 5.1, todos os métodos também foram avaliados com sub-amostragem de exemplos negativos.

Tabela 5.4: Resultados dos experimentos (5 x 5 CV) utilizando combinação de kernels. O melhor método em cada base/condição encontra-se destacado em negrito. O desvio padrão é apresentado entre parênteses. O treinamento dos métodos PKM−KA, PKM−MEAN, PKM−MAX, WANG−MKL e SITAR foi realizado com sub-amostragem de pares desconhecidos, como proposto originalmente pelos autores, e o teste foi realizado tanto no conjunto de testes completo quanto no conjunto de testes com sub-amostragem de pares desconhecidos (balanceado).

Contexto 1						
Dataset	Metodo	Teste		Teste (sub-amostragem)		Tempo Médio (seg.)
NR	SINGLE† ([SW-N] - [SIDER])	0.3416	(±0.0205)	0.7009	(±0.0062)	0.19
	SINGLE* ([SPEC-K4] - [GIP])	0.5187	(±0.0255)	0.8389	(±0.0176)	0.18
	KRONRLS-KA	0.4383	(±0.0089)	0.8043	(±0.0172)	0.21
	KRONRLS-MEAN	0.4130	(±0.0118)	0.7943	(±0.0189)	0.20
	KRONRLS-MKL^{conv}	**0.5308**	(±0.0184)	**0.8294**	(±0.0076)	72.90
	KRONRLS-MKL^{arb}	0.4942	(±0.0189)	0.8065	(±0.0117)	2.94
	PKM-KA	0.1951	(±0.0151)	0.7290	(±0.0380)	0.15
	PKM-MAX	0.0992	(±0.0104)	0.5952	(±0.0212)	0.13
	PKM-MEAN	0.1792	(±0.0158)	0.7170	(±0.0363)	0.14
	SITAR	0.4126	(±0.0392)	0.7793	(±0.0498)	73.87
	WANG-MKL	0.3600	(±0.0387)	0.7662	(±0.0115)	0.76
GPCR	SINGLE† ([MIS-K4M1] - [MINMAX])	0.2752	(±0.0132)	0.8486	(±0.0092)	3.99
	SINGLE* ([SW-N] - [GIP])	0.6188	(±0.0075)	0.9249	(±0.0028)	3.43
	KRONRLS-KA	0.6117	(±0.0136)	0.9438	(±0.0030)	3.72
	KRONRLS-MEAN	0.6105	(±0.0155)	0.9417	(±0.0030)	3.62
	KRONRLS-MKL^{conv}	0.6284	(±0.0200)	0.9460	(±0.0038)	408.01
	KRONRLS-MKL^{arb}	**0.6472**	(±0.0090)	0.9391	(±0.0048)	62.39
	PKM-KA	0.2600	(±0.0149)	0.8711	(±0.0041)	7.75
	PKM-MAX	0.1138	(±0.0058)	0.7642	(±0.0119)	7.46
	PKM-MEAN	0.2653	(±0.0100)	0.8685	(±0.0033)	7.45
	SITAR	0.5348	(±0.0133)	**0.9534**	(±0.0058)	516.03
	WANG-MKL	0.4184	(±0.0172)	0.9165	(±0.0029)	8.97
IC	SINGLE† ([PPI] - [GIP])	0.6789	(±0.0078)	0.9372	(±0.0049)	9.07
	SINGLE* ([SW-N] - [GIP])	0.8679	(±0.0056)	0.9815	(±0.0006)	9.94
	KRONRLS-KA	0.8548	(±0.0015)	0.9757	(±0.0024)	9.99
	KRONRLS-MEAN	0.8690	(±0.0011)	0.9775	(±0.0020)	9.70
	KRONRLS-MKL^{conv}	**0.8765**	(±0.0014)	**0.9802**	(±0.0017)	950.98
	KRONRLS-MKL^{arb}	0.8726	(±0.0031)	0.9791	(±0.0007)	114.08
	PKM-KA	0.5076	(±0.0104)	0.9428	(±0.0067)	38.68
	PKM-MAX	0.0883	(±0.0017)	0.7060	(±0.0109)	35.54
	PKM-MEAN	0.5422	(±0.0107)	0.9467	(±0.0066)	37.10
	SITAR	0.7472	(±0.0191)	0.9752	(±0.0029)	1379.99
	WANG-MKL	0.7204	(±0.0088)	0.9605	(±0.0032)	39.72
E	SINGLE† ([GO] - [MINMAX])	0.3938	(±0.0019)	0.9244	(±0.0021)	267.03
	SINGLE* ([SW-N] - [GIP])	0.8429	(±0.0054)	0.9777	(±0.0012)	268.43
	KRONRLS-KA	0.8630	(±0.0013)	0.9775	(±0.0005)	287.83
	KRONRLS-MEAN	0.8667	(±0.0011)	0.9783	(±0.0005)	282.41
	KRONRLS-MKL^{conv}	0.8818	(±0.0025)	0.9806	(±0.0005)	15435.96
	KRONRLS-MKL^{arb}	**0.8837**	(±0.0017)	0.9780	(±0.0003)	4233.98
	PKM-KA	0.2278	(±0.0032)	0.9427	(±0.0050)	639.34
	PKM-MAX	0.0742	(±0.0009)	0.8676	(±0.0047)	628.80
	PKM-MEAN	0.2034	(±0.0044)	0.9394	(±0.0054)	582.37
	SITAR	0.7478	(±0.0076)	**0.9878**	(±0.0009)	7236.16
	WANG-MKL	0.7256	(±0.0032)	0.9746	(±0.0011)	812.90

†melhor combinação no conjunto de treinamento
*melhor combinação no conjunto de teste

Em um cenário mais realista (primeira coluna da Tabela 5.4), o algoritmo KRONRLS−MKL^{conv} obteve maior AUPR em duas das bases de dados, até mesmo superando o desempenho obtido com

uma combinação ótima de kernels, sob a perspectiva otimista da seleção baseada nos resultados obtidos no conjunto de testes. O algoritmo KRONRLS−MKL^{arb} obteve resultados semelhantes. Com relação aos resultados no conjunto de testes com sub-amostragem, o método proposto apresentou desempenho superior nas em duas bases (NR e IC) e inferior ao método SITAR nas outras bases (GPCR e Enzima). Os maiores valores de AUPR obtidos nos testes com sub-amostragem em comparação ao teste em dados desbalanceados, demonstram claramente que realizar predições neste cenário é uma tarefa mais difícil.

A significância estatística dos resultados nos experimentos do Contexto 1 foram avaliados de acordo com o teste de *Friedman* (FRIEDMAN, 1937), conforme indicado por DEMSAR (2006). Os *p-values* calculados considerando como método base o algoritmo KronRLS-MKL^{conv} (primeira coluna) e o KronRLS-MKL^{arb} (segunda coluna) são apresentados na Tabela 5.5. Os resultados apresentados em negrito tiveram diferença estatística significativa ($p < 0.05$). Os resultados do teste confirmam a superioridade dos algoritmos propostos quando comparados aos métodos competidores, ao mesmo tempo que indicam que não há diferença significativa entre os dois métodos. Esta informação reforça a idéia de que o método de atualização utilizado no algoritmo KronRLS-MKL^{arb} deve ser priorizado, quando não houver a necessidade de considerar restrições sobre os pesos dos kernels.

Tabela 5.5: *p-values* dos resultados dos algoritmos propostos quando comparados com os métodos competidores, no teste de Friedman (Contexto 1). Valores em negrito indicam diferença significativa ($p < 0.05$).

	KRONRLS-MKL^{conv}	KRONRLS-MKL^{arb}
KRONRLS-KA	**1.30E-02**	**1.53E-02**
KRONRLS-MEAN	**2.09E-02**	**2.43E-02**
KRONRLS-MKL^{conv}	-	9.54E-01
KRONRLS-MKL^{arb}	9.54E-01	-
PKM-KA	**6.40E-12**	**9.58E-12**
PKM-MAX	**0.00E+00**	**0.00E+00**
PKM-MEAN	**2.83E-12**	**4.26E-12**
SITAR	**8.64E-05**	**1.10E-04**
WANG-MKL	**3.76E-07**	**5.09E-07**

Em relação ao custo computacional (Figura 5.5), os algoritmos de combinação prévia KRONRLS−KA, KRONRLS−MEAN, PKM−KA, PKM−MEAN, PKM−MAX, e WANG−MKL foram os mais rápidos, obtendo os menores tempos de execução em todos os experimentos. No maior conjunto de dados (Enzima), as abordagens KRONRLS−KA e KRONRLS−MEAN foram mais rápidas, levando em média 24 minutos para o treinamento e teste do modelo, enquanto os métodos baseados em SVM levaram cerca de 53 minutos (5-fold CV). O menor tempo de execução destes métodos se dá majoritariamente pelo fato de que eles utilizam uma combinação prévia de kernels, e portanto não demandam processos iterativos de otimização dos pesos. O algoritmo SITAR levou em média 6, 43, 114 e 603 minutos para as bases NR, GPCR, IC e Enzima, respectivamente. O tempo de execução por ambas versões do KRONRLS−MKL foram inferiores apenas ao obtido pelo método SITAR, exceto na base Enzima, na qual o procedimento de otimização convexa obteve pior tempo (257 minutos). É importante lembrar que, exceto os métodos baseados no algoritmo KRONRLS, todos os outros realizam sub-amostragem de pares negativos, o que leva a conjuntos de treinamento menores.

Comparamos também a quantidade máxima de memória requisitada durante os experimentos. Podemos observar que, tanto os métodos baseados no KronRLS assim como o SITAR, mantêm um consumo estável de memória, independente do tamanho da base. Entretanto, o cálculo explícito da matriz

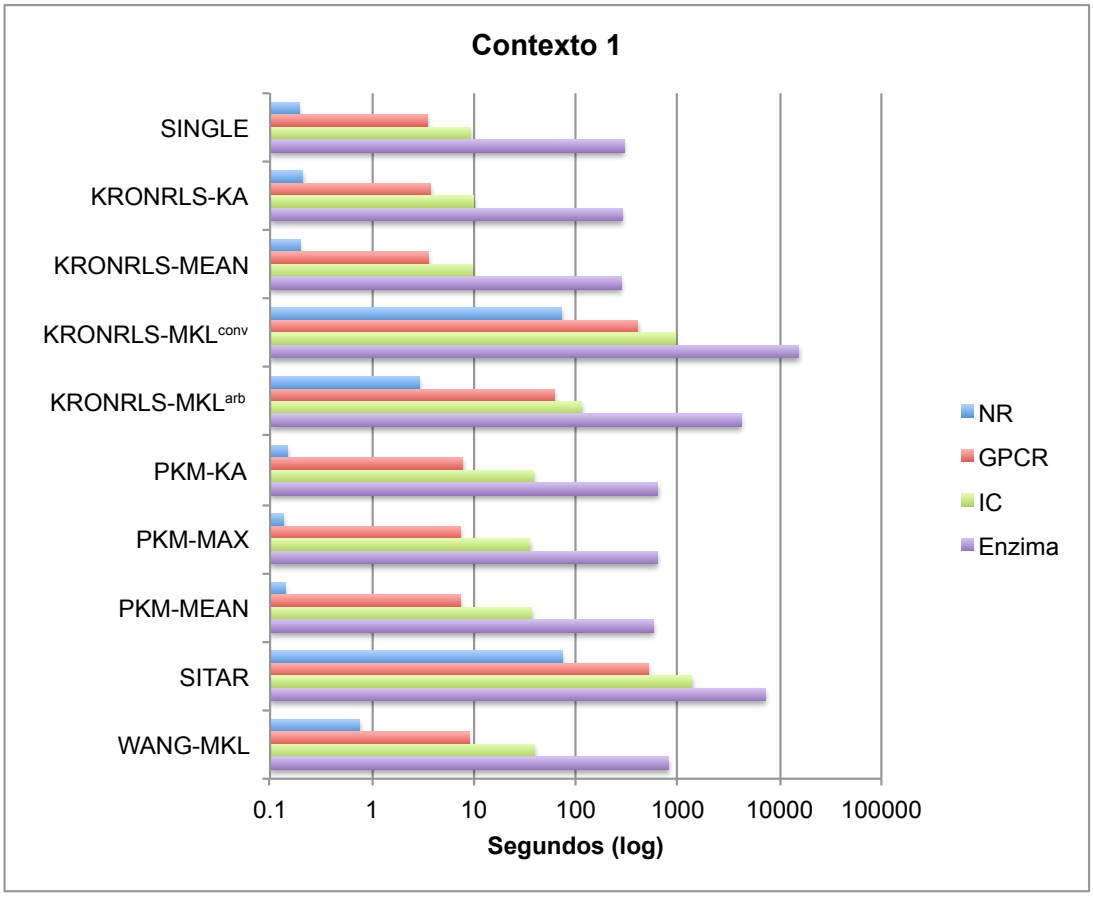

Figura 5.5: Tempo médio de execução dos experimentos nas bases de receptores nucleares (NR), GPCRs, canais de íons (IC) e Enzimas (Enzyme).

de kernel de pares nos métodos baseados em SVM leva a um crescimento exponencial no consumo de memória de tais métodos, mesmo com a sub-amostragem de exemplos negativos, como pode ser observado na Figura 5.6 (escala logarítimica).

5.2.2.2 Contexto 2

Com o intuito de comparar os métodos considerados nas mesmas condições de treinamento, realizamos também experimentos nos quais o treinamento dos métodos baseados em SVM e do algoritmo SITAR foi realizado sem sub-amostragem de padrões negativos (i.e., com classes desbalanceadas). Como esperado, foi observado um aumento considerável no desempenho destes métodos (Tabela 5.6), entretanto, associado a um aumento do custo computacional e de memória (Figura 5.7 e Figura 5.8). Ressaltamos que o algoritmo SITAR e todos os métodos baseados em SVM não puderam ser executados na maior base de dados (Enzimas), pois, no caso dos métodos baseados em SVM, ultrapassaram 64GB de memória RAM, ou, no caso do SITAR, o tempo total de execução do experimento foi superior a 120 horas. Os resultados neste contexto apresentaram menor significância estatística (calculada sobre os resultados nas bases de dados NR, GPCR e IC), conforme pode ser observado na Tabela 5.7.

Em relação ao tempo médio para realização de uma execução do experimento, os métodos `KRONRLS-KA` e `KRONRLS-MEAN` foram os mais rápidos, precisando de menos de 1 minuto para treinar e testar o modelo nas bases NR, GPCR e IC, e aproximadamente 24 minutos para a base Enzima. O

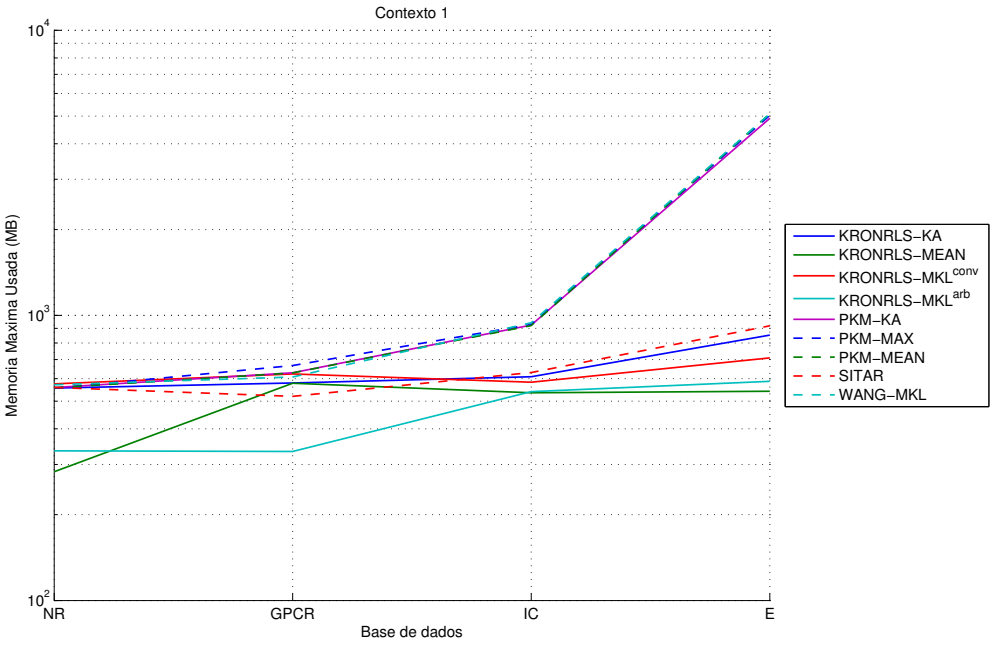

Figura 5.6: Utilização de memória (valor máximo requisitado) pelos métodos avaliados nos experimentos dos métodos avaliados nas bases de receptores nucleares (NR), GPCRs, canais de íons (IC) e Enzimas (Enzyme).

algoritmo SITAR levou em média 31, 779 e 2881 minutos para as bases NR, GPCR e IC, respectivamente. O algoritmo proposto por WANG et al. (2011) levou menos de 1 minuto para a base de dados NR, aproximadamente 50 minutos para a base GPCR e 374 minutos para os dados de canais de íons. Os requisitos de tempo computacional para os algoritmos KRONRLS−MKL^{conv} e KRONRLS−MKL^{arb} foram os mesmos reportados no Contexto 1, de modo que o algoritmo KRONRLS−MKL^{arb} apresentou uma clara vantagem em termos de escalabilidade associada a um desempenho pouco inferior ao melhor método.

Podemos observar que tanto os métodos baseados no KronRLS quanto o SITAR também mantem um consumo estável de memória neste contexto de treinamento, independentemente do tamanho da base, sendo em média 600MB, enquanto os métodos baseados em SVM apresentaram crescimento exponencial em relação ao tamanho da matriz de interações (Figura 5.8).

5.2.3 Análise dos pesos dos kernels

Os pesos dos kernels atribuídos pelos algoritmos KRONRLS−MKL^{conv}, KRONRLS−MKL^{arb} e WANG−MKL, bem como pela heurística KA, podem ser utilizados para analisar a capacidade destes métodos identificarem corretamente as fontes de informação mais relevantes. A abordagem aqui adotada foi a comparação dos pesos obtidos para os kernels de drogas e proteínas nos experimentos do Contexto 1 (Figura 5.9) com o desempenho médio obtido por cada kernel nos experimentos com kernels simples (Figura 5.4). Podemos observar que os pesos atribuídos pela heurística KA nestas bases são muito próximos de uma seleção média dos kernels (0.10), o que seria equivalente à combinação média dos kernels. Na base de dados NR, o algoritmo KRONRLS−MKL^{conv} foi capaz de identificar corretamente os kernels com menor desempenho, tanto para drogas (LAMBDA, MARG, MINIMAX, SPEC e TAN) e

Tabela 5.6: Resultados dos experimentos (5 x 5 CV) utilizando a combinação de kernels. O melhor método em cada base/condição encontra-se destacado em negrito. O desvio padrão é apresentado entre parênteses. O treinamento dos métodos PKM−KA, PKM−MEAN, PKM−MAX, WANG−MKL e SITAR foi realizado com a base de dados completa, i.e., desbalanceada. Os resultados indicados com 'ND' indicam que os métodos excederam os limites de tempo e memória estabelecidos no experimento.

Contexto 2						
Dataset	Metodo	Teste		Teste (sub-amostragem)		Tempo Médio (seg.)
NR	SINGLE† ([SW] - [SIDER])	0.3416	(±0.0205)	0.7009	(±0.0062)	0.19
	SINGLE* ([SPEC-K4] - [GIP])	0.5187	(±0.0255)	0.8389	(±0.0176)	0.18
	KRONRLS-KA	0.4321	(±0.0147)	0.7950	(±0.0235)	0.31
	KRONRLS-MEAN	0.4078	(±0.0211)	0.7813	(±0.0240)	0.22
	KRONRLS-MKL^{conv}	**0.5368**	(±0.0137)	**0.8298**	(±0.0142)	72.08
	KRONRLS-MKL^{arb}	0.4942	(±0.0189)	0.8065	(±0.0117)	2.94
	PKM-KA	0.4517	(±0.0197)	0.7844	(±0.0273)	2.92
	PKM-MAX	0.0811	(±0.0087)	0.5057	(±0.0224)	2.45
	PKM-MEAN	0.4356	(±0.0215)	0.7726	(±0.0300)	3.02
	SITAR	0.5032	(±0.0287)	0.8513	(±0.0127)	373.63
	WANG-MKL	0.4899	(±0.0202)	0.8055	(±0.0178)	3.65
GPCR	SINGLE† ([MIS-K4M1] - [MINMAX])	0.2752	(±0.0132)	0.8486	(±0.0092)	3.99
	SINGLE* ([SW] - [GIP])	0.6188	(±0.0075)	0.9249	(±0.0028)	3.43
	KRONRLS-KA	0.6208	(±0.0081)	0.9437	(±0.0030)	3.71
	KRONRLS-MEAN	0.6213	(±0.0085)	0.9422	(±0.0021)	3.61
	KRONRLS-MKL^{conv}	0.6440	(±0.0052)	0.9475	(±0.0019)	402.29
	KRONRLS-MKL^{arb}	0.6472	(±0.0090)	0.9391	(±0.0048)	62.39
	PKM-KA	0.6782	(±0.0096)	0.9396	(±0.0026)	555.26
	PKM-MAX	0.1572	(±0.0093)	0.7121	(±0.0085)	557.78
	PKM-MEAN	0.6732	(±0.0113)	0.9359	(±0.0023)	568.97
	SITAR	0.6042	(±0.0094)	0.9587	(±0.0079)	9348.15
	WANG-MKL	**0.7064**	(±0.0088)	**0.9501**	(±0.0027)	606.55
IC	SINGLE† ([PPI] - [GIP])	0.6789	(±0.0078)	0.9372	(±0.0049)	9.07
	SINGLE* ([SW] - [GIP])	0.8679	(±0.0056)	0.9815	(±0.0006)	9.94
	KRONRLS-KA	0.8533	(±0.0019)	0.9765	(±0.0023)	11.21
	KRONRLS-MEAN	0.8677	(±0.0021)	0.9782	(±0.0019)	10.12
	KRONRLS-MKL^{conv}	0.8761	(±0.0016)	**0.9807**	(±0.0016)	1009.01
	KRONRLS-MKL^{arb}	0.8726	(±0.0031)	0.9791	(±0.0007)	114.08
	PKM-KA	0.9047	(±0.0017)	0.9791	(±0.0003)	2769.24
	PKM-MAX	0.2535	(±0.0079)	0.7651	(±0.0099)	2620.33
	PKM-MEAN	**0.9065**	(±0.0020)	0.9797	(±0.0004)	2693.35
	SITAR	0.8065	(±0.0036)	0.9802	(±0.0025)	34583.82
	WANG-MKL	0.8788	(±0.0033)	0.9779	(±0.0007)	4486.51
E	SINGLE† ([GO] - [MINMAX])	0.3938	(±0.0019)	0.9244	(±0.0021)	267.03
	SINGLE* ([SW] - [GIP])	0.8429	(±0.0054)	0.9777	(±0.0012)	268.43
	KRONRLS-KA	0.8630	(±0.0013)	0.9775	(±0.0005)	287.83
	KRONRLS-MEAN	0.8667	(±0.0011)	0.9783	(±0.0005)	282.41
	KRONRLS-MKL^{conv}	0.8818	(±0.0025)	**0.9806**	(±0.0005)	15435.96
	KRONRLS-MKL^{arb}	**0.8837**	(±0.0017)	0.9780	(±0.0003)	4233.98
	PKM-KA	ND		ND		ND
	PKM-MAX	ND		ND		ND
	PKM-MEAN	ND		ND		ND
	SITAR	ND		ND		ND
	WANG-MKL	ND		ND		ND

†melhor combinação no conjunto de treinamento
*melhor combinação no conjunto de teste

proteínas (MIS−k3m2). O algoritmo KRONRLS−MKL^{arb} demonstrou ser mais sensível aos efeitos da regularização dos kernels, uma vez que os pesos atribuídos por ele muitas vezes se aproximam da média, assim como o KA. Resultados semelhantes podem ser observados nas outras bases.

A estratégia de estimação dos pesos adotada pelo método WANG−MKL também mostrou-se eficiente, sendo capaz de identificar corretamente a maioria dos kernels mais relevantes, com uma forte tendência a selecionar um vetor de pesos mais esparso. Também observamos que este método é aparentemente afetado pelo tamanho da base, o que pode ser observado nos pesos atribuídos aos kernels

Tabela 5.7: *p-values* dos resultados dos algoritmos propostos quando comparados com os métodos competidores, no teste de Friedman (Contexto 2). Valores em negrito indicam diferença significativa ($p < 0.05$).

	KRONRLS-MKL^{conv}	KRONRLS-MKL^{arb}
KRONRLS-KA	**4.10E-04**	**9.32E-03**
KRONRLS-MEAN	**5.27E-04**	**1.13E-02**
KRONRLS-MKL^{conv}	-	3.51E-01
KRONRLS-MKL^{arb}	3.51E-01	-
PKM-KA	7.90E-01	2.30E-01
PKM-MAX	**9.85E-09**	**1.59E-06**
PKM-MEAN	8.41E-01	4.63E-01
SITAR	**2.70E-03**	**3.88E-02**
WANG-MKL	3.17E-01	**5.32E-02**

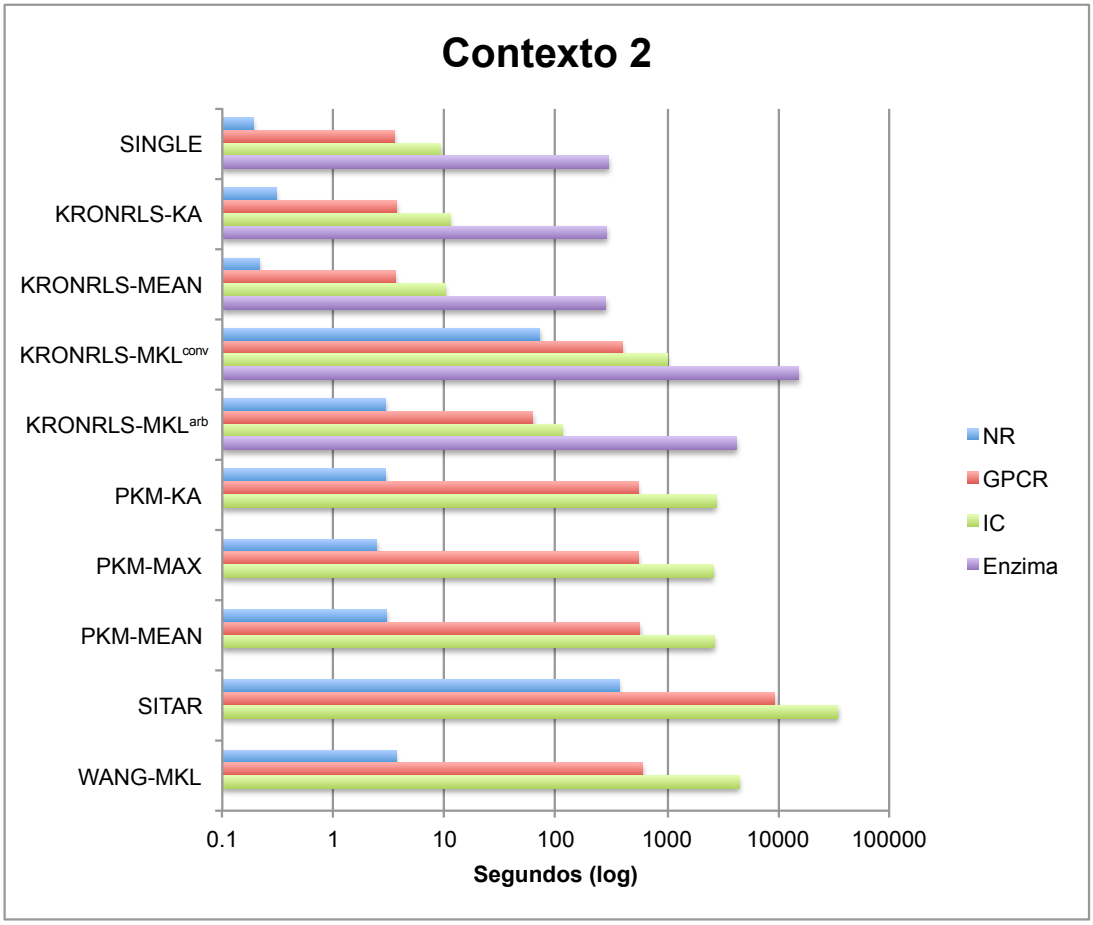

Figura 5.7: Tempo médio de execução dos experimentos nas bases NR, GPCRs, IC e Enzima.

`MIS-k3m1` e `MIS-k4m1`, cujos valores descrescem a medida que o tamanho das bases aumenta, embora estes tenham apresentado bom desempenho nos experimentos com kernels simples em todas as bases. A fim de quantificar o grau de concordância entre os pesos encontrados e o desempenho médio obtido por cada kernel, adotamos uma abordagem aproximada como critério de avaliação. Foi calculado o coeficiente de correlação entre os pesos encontrados e a AUPR média obtida por cada kernel de um conjunto (i.e., drogas), quando utilizado em conjunto com todos os kernels do outro conjunto (i.e., proteínas) (Tabela 5.8). Apesar desta análise não ser totalmente conclusiva, dado que a AUPR média é apenas uma aproximação dos pesos ideais, ela permite traçar algumas observações importantes. O algoritmo `KRONRLS-MKL`^{conv}

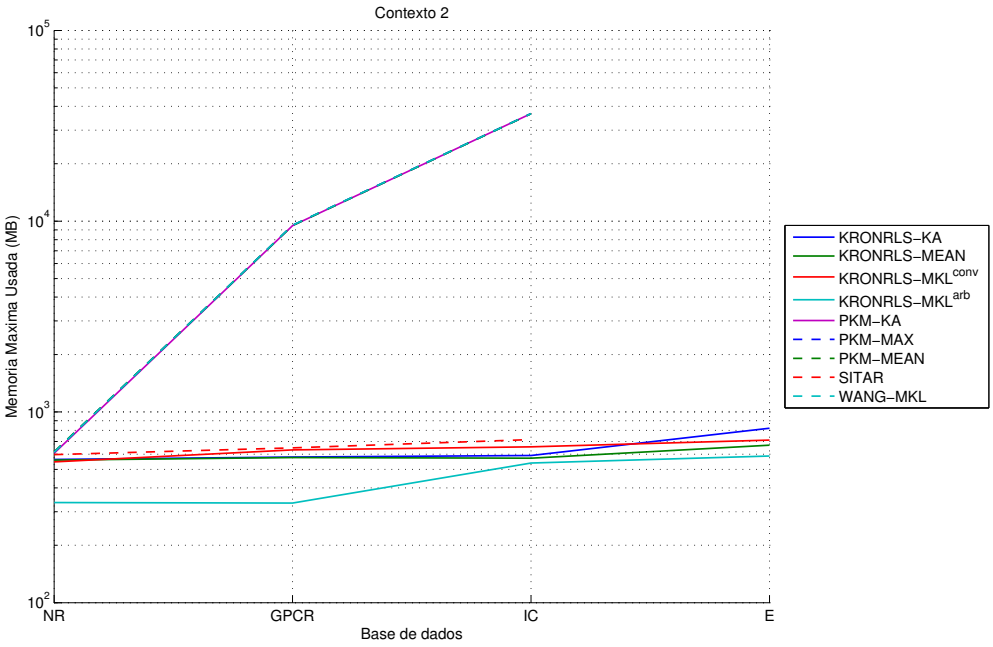

Figura 5.8: Utilização de memória (valor máximo requisitado) nos experimentos dos métodos avaliados nas bases de receptores nucleares (NR), GPCRs e canais de íons (IC).

Tabela 5.8: Coeficiente de Correlação entre desempenho médio de cada kernel nos experimentos de pares de kernels simples e os pesos médios a eles atribuídos pelos métodos KA, KRONRLS−MKL^{conv}, KRONRLS−MKL^{arb} e WANG−MKL.

Dataset	Entidade	KA	KRONRLS-MKL^{conv}	KRONRLS-MKL^{arb}	WANG-MKL
NR	Proteínas	0.7662	**0.8408**	0.3620	0.2135
	Drogas	0.7561	**0.8215**	0.4329	0.5741
GPCR	Proteínas	**0.5436**	0.2173	-0.3197	0.5243
	Drogas	-0.5135	**0.8086**	-0.3342	0.6252
IC	Proteínas	0.3994	0.1736	0.4216	**0.4713**
	Drogas	-0.6140	**0.8382**	-0.2395	0.4072
E	Proteínas	**0.3766**	0.2463	0.1793	0.3115
	Drogas	-0.1003	**0.7742**	0.0764	0.3754

apresentou maior correlação com o desempenho médio de todos os kernels de drogas, e obteve maior correlação com o desempenho dos kernels de proteínas apenas na base NR. O KA demonstrou maior correlação nos conjuntos GPCR e Enzimas, enquanto o WANG−MKL foi superior no conjunto de kernels de proteínas da base IC. A maior correlação dos pesos indicados pelo KA com os kernels de proteínas pode ser justificado pela maior homogeneidade na qualidade destes kernels, e, uma vez que os pesos atribuídos pelo KA foram mais próximos da média, apresentaram maior correlação. Entretanto, em um contexto de maior heterogeneidade da qualidade dos kernels (e.g., drogas), o algoritmo KRONRLS−MKL^{conv} demonstrou maior capacidade de selecionar kernels mais relevantes.

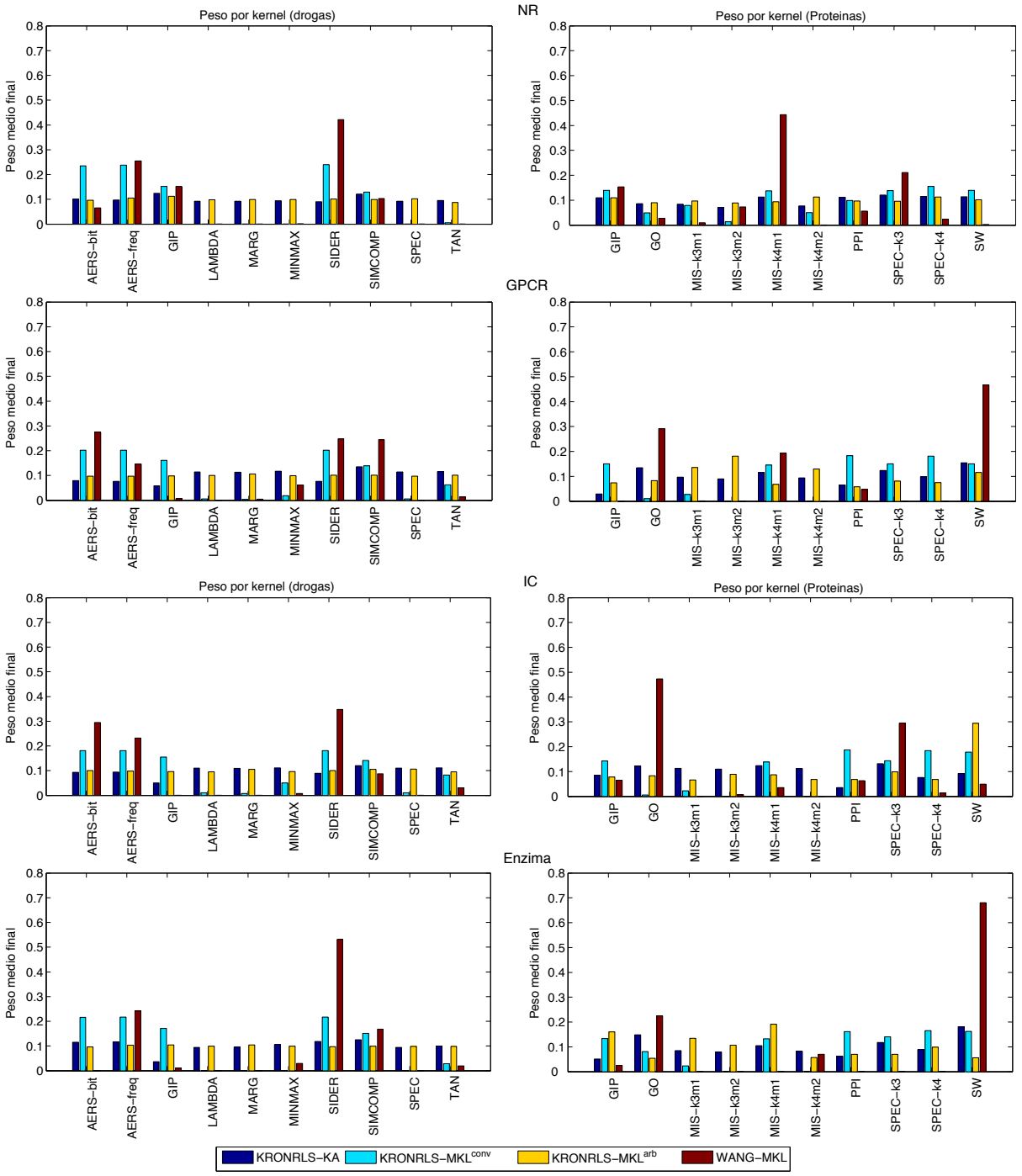

Figura 5.9: Pesos atribuídos aos kernels estudados pela heurística `KA` e algoritmos `KRONRLS-MKL` e `WANG-MKL`. Como pode-se observar, o `KA` apresentou pesos próximos da combinação média, enquanto os métodos `KRONRLS-MKL`^{conv} e `WANG-MKL` efetivamente foram capazes de descartar os kernels menos relevantes.

5.2.4 Predições em bases de dados atualizadas

Dado que os conjuntos de dados utilizados neste estudo foram gerados há aproximadamente oito anos atrás, e uma vez que novas interações são continuamente identificadas e reportadas, é esperado que novas interações, originalmente não consideradas, estejam agora presentes nestas bases. Sendo assim, com o objetivo de avaliar a qualidade das predições obtidas pelos métodos analisados em um

cenário mais atual, realizamos uma busca por novas interações depositadas nestas bases. Esta análise distingue-se da adotada em trabalhos anteriores (LAARHOVEN; NABUURS; MARCHIORI, 2011; BLEAKLEY; YAMANISHI, 2009), pois ao invés de apenas contabilizar a quantidade de novas interações corretamente identificadas, optamos por calcular o AUPR sobre toda a base de dados, a fim de ter uma visão da real dificuldade desta tarefa. Os métodos competidores foram treinados com o conjunto de dados utilizado nos experimentos descritos na seção anterior, e os pares droga-proteína com maiores scores foram considerados como interações mais prováveis. As predições obtidas foram então comparadas com as interações confirmadas na versão mais recente de quatro grandes bases de dados de interações biológicas: DrugBank (WISHART et al., 2008), MATADOR (GüNTHER et al., 2008), KEGG (KANEHISA; GOTO, 2000)) and ChEMBL (BENTO et al., 2014). A quantidade de novas interações encontradas em cada uma dessas bases é reportada na tabela Tabela 5.9.

Tabela 5.9: Total de novas interações encontradas nas versões atuais das bases KEGG, Matador, Drugbank e ChEMBL.

Dataset	KEGG	MATADOR	DRUGBANK	ChEMBL
NR	8	3	11	168
GPCR	142	58	394	2306
IC	268	379	398	441
E	46	264	929	4764

As interações presentes no conjunto de dados original foram removidas, e a AUPR para cada base de dados foi calculada separadamente (Tabela 5.10). O baixos valores de AUPR obtidos por todos os métodos indicam a dificuldade de predizer interações neste espaço de buscas. Trabalhos recentes indicam que um dos possíveis motivos para o baixo desempenho neste caso pode ter sido provocada pela ausência de informação sobre pares droga-proteína que não interagem (padrões efetivamente negativos) (LIU et al., 2015; CHEN et al., 2015).

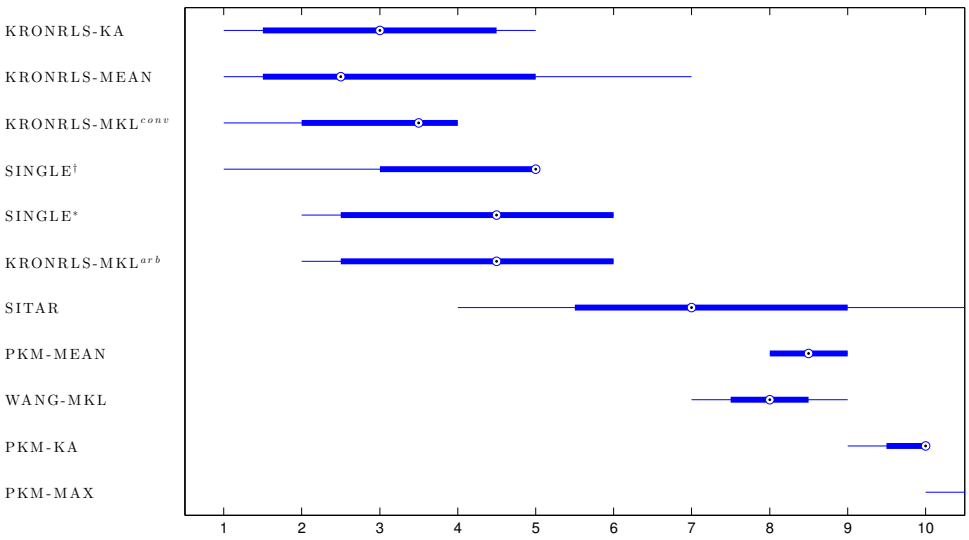

Figura 5.10: Ranking médio de cada método quando nos experimentos com bases de dados atualizadas. Os métodos baseados em KronRLS obtiveram desempenho superior quando comparados com outras estratégias de integração.

Ainda assim, o ranking do desempenho médio de cada método em todas as bases, indicou que os métodos baseados no KronRLS foram superiores, mesmo quando comparado às abordagens de kernel simples (Figura 5.10). Apesar de pequena, a diferença é significativa, como pode ser observado na Figura 5.11, na qual os resultados do teste de Friedman-Nemenyi são apresentados na forma de gráfico CD (*critical difference*) (DEMSAR, 2006). Este teste consiste em uma abordagem não paramétrica e conservadora para comparação do desempenho entre classificadores, de modo que utilizamos um critério menos restritivo para avaliação ($p < 0.10$). O teste revela que os métodos KRONRLS-MKL^{conv}, KRONRLS-KA, SINGLE e KRONRLS-MEAN tiveram desempenho significativamente melhor que os algoritmos baseados em SVM. Por outro lado, os dados não são suficientes para concluir se algoritmo KRONRLS-MKL^{arb} é superior aos métodos baseados em SVM.

Tabela 5.10: Valores da AUPR quando comparados os scores preditos e novas interações encontradas nas versões atuais das bases KEGG, Matador, Drugbank e ChEMBL.

Dataset	Método	KEGG	MATADOR	DRUGBANK	ChEMBL
NR	SINGLE†	**0.4122**	0.0149	0.0442	0.2537
	SINGLE*	0.2533	**0.0227**	0.0389	0.2245
	KRONRLS-KA	0.0599	0.0095	0.0499	**0.3246**
	KRONRLS-MEAN	0.0410	0.0097	0.0469	0.3238
	KRONRLS-MKL^{conv}	0.1791	0.0085	**0.0568**	0.3089
	KRONRLS-MKL^{arb}	0.0414	0.0081	0.0454	0.3115
	PKM-MEAN	0.0618	0.0057	0.0182	0.1932
	PKM-KA	0.0753	0.0052	0.0203	0.1868
	PKM-MAX	0.0282	0.0067	0.0105	0.1424
	WANG-MKL	0.0598	0.0034	0.0312	0.2430
	SITAR	0.0960	0.0011	0.0042	0.0639
GPCR	SINGLE†	0.0912	0.0179	0.1474	0.2121
	SINGLE*	0.0912	0.0179	0.1474	0.2121
	KRONRLS-KA	0.0867	0.0230	0.1770	0.2489
	KRONRLS-MEAN	0.0746	0.0249	0.1880	0.2544
	KRONRLS-MKL^{conv}	**0.1133**	**0.0281**	**0.2268**	0.2228
	KRONRLS-MKL^{arb}	0.0687	0.0219	0.1760	**0.2564**
	PKM-MEAN	0.0730	0.0076	0.0603	0.1867
	PKM-KA	0.0474	0.0062	0.0464	0.1780
	PKM-MAX	0.0087	0.0025	0.0210	0.1262
	WANG-MKL	0.0740	0.0125	0.0742	0.1969
	SITAR	0.0536	0.0218	0.0841	0.1602
IC	SINGLE†	0.0905	0.0234	0.0557	0.0409
	SINGLE*	0.0905	0.0234	0.0557	0.0409
	KRONRLS-KA	0.0714	**0.0348**	0.0472	0.0399
	KRONRLS-MEAN	0.0618	0.0275	0.0369	0.0363
	KRONRLS-MKL^{conv}	0.0714	0.0272	0.0491	0.0387
	KRONRLS-MKL^{arb}	0.0722	0.0243	0.0449	0.0360
	PKM-MEAN	0.0275	0.0154	0.0534	**0.0665**
	PKM-KA	0.0278	0.0149	0.0545	0.0572
	PKM-MAX	0.0094	0.0147	0.0101	0.0104
	WANG-MKL	0.0315	0.0115	0.0460	0.0366
	SITAR	**0.1051**	0.0266	**0.0867**	0.0314
E	SINGLE†	0.0058	0.0621	**0.0518**	0.0422
	SINGLE*	0.0058	0.0621	**0.0518**	0.0422
	KRONRLS-KA	**0.0084**	**0.0666**	0.0323	0.0385
	KRONRLS-MEAN	0.0063	0.0652	0.0369	0.0432
	KRONRLS-MKL^{conv}	0.0074	0.0596	0.0384	0.0332
	KRONRLS-MKL^{arb}	0.0070	0.0592	0.0432	**0.0477**
	PKM-MEAN	0.0001	0.0013	0.0024	0.0155
	PKM-KA	0.0001	0.0012	0.0028	0.0156
	PKM-MAX	0.0001	0.0009	0.0028	0.0180
	WANG-MKL	0.0001	0.0011	0.0025	0.0137
	SITAR	0.0006	0.0045	0.0322	0.0403

†melhor combinação no conjunto de treinamento *melhor combinação no conjunto de teste

Em seguida, foi realizada uma análise mais prática do potencial de predição do algoritmo KRONRLS-MKL^{conv}, observando-se as 5 interações com maior score predito (Tabela Tabela 5.11). Pode-

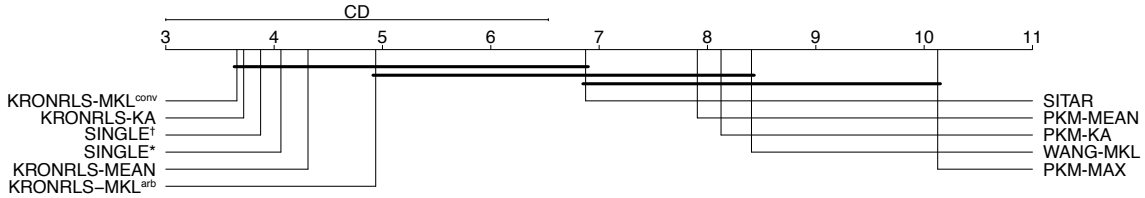

Figura 5.11: Comparação do desempenho na predição de novas interações em bases de dados atualizadas de todos os classificadores entre si com o teste de Friedman-Nemenyi. Grupos de classificadores que podem ser iguais em desempenho ($p = 0.10$) estão conectados.

mos notar que a grande maioria das interações (14 das 20) já haviam sido descritas nas bases de dados ChEMBL, DrugBank e Matador. Concentraremos a discussão em uma seleção das novas interações identificadas. Por exemplo, na base de dados de Receptores Nucleares, a predição na posição 5 indica a associação da droga Tretinoin com a proteína RORa (*nuclear factor RAR-related orphan receptor A*). Tretinoin é uma droga utilizada atualmente no tratamento de acnes (WEBSTER, 1998). É interessante notar que sua atividade molecular está associada com a ativação de receptores nucleares da família de proteínas próximas RAR.

Tabela 5.11: Cinco predições com maior score preditas pelo algoritmo KRONRLS-MKL^{conv}.

Droga		Proteína		
Nuclear Receptors				
D00951	*Medroxyprogesterone acetate*	hsa2099	*estrogen receptor 1*	(D,C)
D00585	*Mifepristone*	hsa2099	*estrogen receptor 1*	(C)
D00182	*Norethisterone*	hsa2099	*estrogen receptor 1*	(C)
D00105	*Estradiol*	hsa5241	*progesterone receptor*	(C)
D00094	*Tretinoin*	hsa6095	*RAR-related orphan receptor A*	
GPCR				
D02358	*Metoprolol*	hsa154	*adrenoceptor beta 2, surface*	(D,C)
D00283	*Clozapine*	hsa1814	*dopamine receptor D3*	(D,C,M)
D00371	*Theophylline*	hsa135	*adenosine A2a receptor*	(K,D,C)
D00371	*Theophylline*	hsa134	*adenosine A1 receptor*	(K,D,C)
D00095	*Adrenaline*	hsa155	*adrenoceptor beta 3*	(K,D,C)
Ion Channel				
D00775	*Riluzole*	hsa2898	*glutamate receptor, ionotropic, kainate 2*	(M)
D02356	*Verapamil*	hsa6833	*ATP-binding cassette, sub-family C (CFTR*	
D00294	*Diazoxide*	hsa10060	*ATP-binding cassette, sub-family C (CFTR*	
D02356	*Verapamil*	hsa56660	*potassium channel, two pore domain subfamily K, member 12*	
D00524	*Carbachol*	hsa1134	*cholinergic receptor, nicotinic, alpha 1 (muscle)*	
Enzyme				
D00542	*Halothane*	hsa1571	*cytochrome P450, family 2, subfamily E, polypeptide 1*	(D,C,M)
D00437	*Nifedipine*	hsa1559	*cytochrome P450, family 2, subfamily C, polypeptide 9*	(D,C,M)
D00528	*Anhydrous caffeine*	hsa1549	*cytochrome P450, family 2, subfamily A, polypeptide 7*	(M)
D03670	*Deferoxamine*	hsa1579	*cytochrome P450, family 4, subfamily A, polypeptide 11*	
D00139	*Methoxsalen*	hsa1543	*cytochrome P450, family 1, subfamily A, polypeptide 1*	(D,M)

Interações encontradas no KEGG, DrugBank, ChEMBL e Matador são marcadas com K, D, C e M respectivamente.

Este também é um bom exemplo para ilustrar os benefícios da incorporação de múltiplas fontes de dados. Ambos RORa e Tretinoin não compartilham nós no conjunto de treinamento. Todos os alvos farmacológicos do Tretinoin tem uma alta similaridade funcional (GO) a proteína RORa (com valor médio 0.8368), apesar da sua baixa similaridade de sequência (o valor médio do kernel SW para estas proteínas é de 0.1563). Adicionalmente, um dos alvos do Tretinoin é a proteína NR0B1 (*nuclear receptor subfamily 0, group B, member 1*). Esta proteína é muito próxima a RORa na rede PPI (score de similaridade 0.90).

Com relação ao conjunto de dados de Canais de Ions (IC), as predições 2 e 3 indicam a interação do Verapamil e Diazoxide com a proteína ABBCC8 (*ATP-binding cassete sub-family C*). ABBCC8 é uma das proteínas codificantes do receptor SUR1 (*sulfonylurea receptor*) e está associado a regulação de cálcio e diabetes tipo I REIS; VELHO (2002). É importante ressaltar que existem estudos com ratos que indicam que o Diazoxide pode ser usado em tratamentos de prevenção de diabetes HUANG et al. (2007).

5.3 Considerações finais

Os resultados obtidos nos experimentos realizados indicaram que os métodos propostos são capazes de identificar e selecionar as fontes de informação mais relevantes para o problema de predição de interações droga-proteína. Os métodos MKL presentes na literatura concentram-se principalmente no problema de aprendizagem de kernels quando estes são construídos sobre o mesmo conjunto de objetos, o que não é o caso no problema da predição de interações em redes bipartidas. Além disso, muitos destes métodos tem o algoritmo SVM como aprendiz base, o que introduz limitações no cálculo do kernel de pares.

A combinação de múltiplas fontes de informação tem se mostrado promissora, trazendo ganhos na qualidade das predições de novas interações em redes droga-proteína. Entretanto, a maior parte destes trabalhos utiliza a combinação simples de kernels, como a soma ou a média dos kernels base, além de também serem limitadas em relação ao tamanho da matriz de kernel de pares produzida. O método proposto preenche a lacuna deixada pelas abordagens anteriores, uma vez que substitui o algoritmo SVM pelo KronRLS, um método capaz de realizar eficientemente a predição de interações em bases bipartidas de tamanhos arbitrários.

Além disso, os pesos dos kernels base indicados pelo algoritmo `KRONRLS-MKL` apresentaram-se capazes de identificar os kernels de desempenho mais baixo, de acordo com a acurácia obtida por experimentos com todas as combinações de kernels simples. A combinação não esparsa destes kernels possibilita um aumento no poder de generalização do modelo, reduzindo o viés para um determinado tipo de kernel. Isto possivelmente leva a um melhor desempenho, uma vez que o modelo pode se beneficiar de fontes de informação heterogêneas de maneira sistemática (KLOFT et al., 2008). O desempenho do algoritmo não foi influenciado pelo desbalanceamento das classes, podendo ser treinado sobre todo o espaço de interações sem sacrificar o desempenho. De modo geral, o algoritmo proposto obteve melhores resultados quando comparado com outras estratégias *baseline* de combinação de informações e outras estratégias integrativas PERLMAN et al. (2011); WANG et al. (2011); QIU; LANE (2009). O algoritmo $\texttt{KRONRLS-MKL}^{arb}$ apresentou desempenho comparável à abordagem com combinação convexa dos pesos, com custo computacional inferior. Entretanto, a seleção dos kernels feitas por ele demonstrou ser mais sensível a regularização, de modo que acreditamos que este possa ser aprimorado, possivelmente com a introdução de restrições no procedimento de otimização adotado (combinação convexa).

6

Conclusão

Redes droga-proteína têm recebido bastante atenção nos últimos anos, dada sua relevância para a inovação farmacêutica e produção de novos fármacos (BARABáSI; GULBAHCE; LOSCALZO, 2011; YAMANISHI, 2013). Este tipo de rede consiste em uma modelagem matemática na qual a existência de interação entre uma droga (composto químico) e uma proteína (alvo farmacológico) é representada através de uma aresta ligando estes nós em um grafo.

O recente crescimento de bases de dados abertas (ERTL; JELFS, 2007), tem motivado o desenvolvimento de métodos computacionais com foco na mineração da informação disponível sobre drogas e alvos farmacológicos. Uma abordagem que tem se destacado é a chamada mineração de dados no espaço quimiogenômico através do uso de uma classe especial de algoritmos de aprendizagem de máquina, os chamados métodos de kernel. O desempenho obtido por um método de kernel em um determinado problema depende intrínsecamente da qualidade da medida de similaridade adotada, i.e., o kernel (SCHOLKOPF; SMOLA, 2001). Uma forma de superar esta limitação consiste em utilizar um conjunto de kernels, ao invés de se restringir a uma única medida de similaridade, e aplicar estratégias de otimização para seleção dos melhores kernels para a tarefa em questão.

Com o objetivo de combinar a alta disponibilidade de dados biológicos na internet e o potencial preditivo dos métodos de kernel, propomos neste trabalho um algoritmo de combinação de kernels capaz de selecionar e combinar fontes de informação heterogêneas relevantes para a predição de interações droga-proteína. A técnica proposta incorpora a aprendizagem de múltiplos kernels ao problema da predição de links em redes bipartidas, com a escalabilidade necessária para realizar predições em bases com centenas de drogas e proteínas combinando dezenas de bases de dados distintas. Duas abordagens de otimização dos pesos utilizados na combinação dos kernels foram propostas, e comparadas com outras estratégias de combinação de kernels em quatro bases de dados de interações conhecidas.

Os experimentos foram realizados sobre um total de 10 kernels de drogas e 10 kernels de proteínas, calculados sobre compostos e seus respectivos alvos biológicos em quatro classes de proteínas distintas: receptores nucleares, receptores acoplados a proteína G, canais de íons e enzimas. Um total de 100 combinações possíveis de kernels tiveram sua qualidade preditiva avaliada tanto individualmente, quanto também sob 9 abordagens de combinação distintas. Os resultados demonstraram a vantagem dos métodos propostos em termos de qualidade das predições, bem como em relação ao custo computacional e consumo de memória. A combinação não convexa dos kernels mostrou-se interessante por conciliar qualidade das predições a menor tempo de convergência, entretanto, o método demonstrou maior sensibi-

lidade a regularização dos pesos dos kernels, encontrando maior dificuldade em descartar kernels pouco informativos.

6.1 Contribuições da tese

As contribuições desta tese se dão em três linhas principais:

- **Relevância de kernels para a predição de interações farmacológicas**: foram investigados os efeitos de diferentes descritores (i.e., medidas de similaridade) no problema de predição de interações droga-proteína em espaços quimiogenômicos. Foi identificado que além do tipo de kernel utilizado para cada tipo de entidade, a parametrização adotada têm grande influência no desempenho do modelo preditivo construído. Com relação aos kernels de drogas, observamos que medidas de similaridade baseadas em características farmacológicas e efeitos colaterais são capazes de identificar características relevantes do perfil de interação de cada composto, enquanto medidas baseadas apenas no grafo da estrutura química, como o `SPEC, LAMBDA` e `MARG`, não conseguem bons resultados, com exceção do kernel `SIMCOMP`, que além do grafo da estrutura considera também aspectos fisico-químicos da molécula. Diversos tipos de kernels baseados em sequências de proteínas foram considerados, os quais em geral demonstraram grande habilidade em identificar os aspectos relevantes para o problema, mas que podem variar drasticamente com a parametrização adotada (e.g., o kernel `MIS-k3-m2`).

- **Aprendizagem de múltiplos kernels**: demonstramos que as abordagens de MKL propostas anteriormente não eram suficientemente adequadas ao problema da predição de links em redes bipartidas, que requer abordagens específicas. Isto se dá especialmente pelo fato de que os vértices do grafo subjacente pertencem a conjuntos disjuntos, de maneira que não podem ser unificados em um único conjunto de kernels base. Associado a isto, a alta dimensionalidade alcançada pela matriz de kernel de pares impõe que classificadores base alternativos ao algoritmo SVM, utilizado na maior parte dos métodos propostos até então, sejam utilizados. Demonstramos que a utilização do algoritmo KronRLS como classificador base traz benefícios tanto em termos de desempenho como também de complexidade de tempo e memória.

- **Predição de interações droga-proteína**: propomos uma abordagem para a integração de múltiplas fontes de informação acerca de drogas e proteínas, capaz de selecionar e combinar os dados mais relevantes para a predição de novas interações. Os experimentos realizados demonstraram que a combinação de kernels base produz resultados superiores aos obtidos com cada kernel quando utilizado isoladamente. O desempenho do algoritmo proposto demonstrou robustez em relação ao desbalanceamento das classes, podendo ser treinado com o conjunto completo de interações possíveis associadas a dezenas de medidas de similaridade, sem grandes impactos no custo computacional.

Além destes aspectos, acreditamos que os resultados obtidos nesta tese abrem espaço para uma série de trabalhos na predição de *links*, não apenas em redes droga-proteína, mas em problemas de natureza bipartida em geral, algumas das quais são apresentadas na Seção 6.2.

6.2 Trabalhos futuros

Apesar de ter sido desenvolvido com foco na predição de interações em redes droga-proteína, o algoritmo proposto não limita-se a este contexto, podendo ser aplicado em outros problemas de predição em redes de natureza bipartida, como por exemplo, sistemas de recomendação e redes de documentos e características. A aplicação do KronRLS-MKL nestes contextos deve ocorrer de forma direta, sem a necessidade de maiores ajustes no algoritmo. Outras estratégias de construção do kernel de pares também podem ser exploradas, em especial, a soma de Kronecker (KASHIMA et al., 2009).

Neste trabalho, foi adotada uma estratégia ingênua de tratamento de dados faltosos nas matrizes de kernels base, onde foi atribuída similaridade zero a elementos cuja informação não foi encontrada em determinada fonte de informação. O tratamento de dados faltosos nas matrizes de kernel base pode ser largamente melhorado, sob a luz dos recentes avanços na pesquisa de métodos de imputação de valores faltosos em problemas MKL (KUMAR et al., 2013; LIU et al., 2013).

A concepção dos experimentos realizados neste trabalho é relevante para um contexto em que se deseja investigar a polifarmacologia ou o reposicionamento de drogas, uma vez que foi avaliada apenas a capacidade de predição de interação entre pares de drogas e proteínas que possuem ao menos uma interação conhecida. Outros experimentos podem ser realizados no contexto em que é apresentada uma nova droga (ou proteína), para a qual nenhuma interação é conhecida (ver Capítulo 2).

Outros tipos de kernel também podem ser introduzidos, como por exemplo kernels baseados em perfis de expressão gênica (LAMB et al., 2006), além da introdução de diferentes kernels de grafos (KONDOR; LAFFERTY, 2002; KUNEGIS; LUCA; ALBAYRAK, 2010). Diferentes parametrizações dos kernels base também podem ser avaliados, cuja combinação pode ser realizada em um passo único ou em camadas separadas para cada tipo de kernel.

A estratégia utilizada na adaptação do algoritmo KronRLS para o contexto da aprendizagem de múltiplos kernels pode ser utilizada para adaptar outros métodos para predição em redes bipartidas, mas que até então se restringem ao uso de um único kernel. Outra característica que pode ser re-utilizada é a introdução de critérios de regularização dos pesos, o que pode melhorar consideravelmente métodos com tendência a uma seleção mais esparsa dos pesos, como a heurística `WANG-MKL`.

Referências

A.K. Jain; R.P.W. Duin; MAO, J. Statistical Pattern Recognition: a review. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, [S.l.], v.22, n.1, p.4–37, 2000.

ARANY, a. et al. Multi-Aspect Candidates for Repositioning: data fusion methods using heterogeneous information sources. **Current Medicinal Chemistry**, [S.l.], v.20, n.1, p.95–107, 2012.

ARRELL, D. K.; TERZIC, a. Network systems biology for drug discovery. **Clinical pharmacology and therapeutics**, [S.l.], v.88, n.1, p.120–5, July 2010.

BARABáSI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. **Nature Reviews Genetics**, [S.l.], v.12, n.1, p.56–68, Jan. 2011.

BEN-HUR, a.; BRUTLAG, D. Remote homology detection: a motif based approach. **Bioinformatics**, [S.l.], v.19, n.Suppl 1, p.i26–i33, July 2003.

BEN-HUR, A.; NOBLE, W. S. Kernel methods for predicting protein-protein interactions. **Bioinformatics (Oxford, England)**, [S.l.], v.21 Suppl 1, p.i38–46, June 2005.

BENTO, A. P. et al. The ChEMBL bioactivity database: an update. **Nucleic Acids Research**, [S.l.], v.42, n.D1, p.D1083–D1090, 2014.

BLEAKLEY, K.; YAMANISHI, Y. Supervised prediction of drug-target interactions using bipartite local models. **Bioinformatics (Oxford, England)**, [S.l.], v.25, n.18, p.2397–403, Sept. 2009.

BUTCHER, E. Can cell systems biology rescue drug discovery? **Drug Discovery**, [S.l.], v.4, n.June, p.461–467, 2005.

BUTINA, D.; SEGALL, M. D.; FRANKCOMBE, K. Predicting ADME properties in silico: methods and models. **Drug discovery today**, [S.l.], v.7, n.11, p.S83–8, June 2002.

BYRD, R. H.; HRIBAR, M. E.; NOCEDAL, J. An Interior Point Algorithm for Large-Scale Nonlinear Programming. **SIAM Journal on Optimization**, [S.l.], v.9, n.4, p.877–900, 1999.

CHAWLA, N. V.; JAPKOWICZ, N.; DRIVE, P. Editorial : special issue on learning from imbalanced data sets. **ACM SIGKDD Explorations Newsletter**, [S.l.], v.6, n.1, p.1–6, 2004.

CHEN, X. et al. Drug – target interaction prediction : databases , web servers and computational models. **Briefings in bioinformatics**, [S.l.], n.May, p.1–17, 2015.

CHEN, X.; LIU, M.-X.; YAN, G.-Y. Drug-target interaction prediction by random walk on the heterogeneous network. **Molecular bioSystems**, [S.l.], v.8, n.7, p.1970–8, July 2012.

CHENG, F. et al. Prediction of chemical-protein interactions network with weighted network-based inference method. **PloS one**, [S.l.], v.7, n.7, p.e41064, Jan. 2012.

CONSORTIUM, G. O.; OTHERS. The Gene Ontology (GO) database and informatics resource. **Nucleic acids research**, [S.l.], v.32, n.suppl 1, p.D258—-D261, 2004.

CRISTIANINI, N. et al. On kernel-target alignment. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 14. **Anais...** MIT Press, 2002. p.367–373.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines: and other kernel-based learning methods**. New York, NY, USA: Cambridge University Press, 2000.

CSERMELY, P. et al. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. **Pharmacology & therapeutics**, [S.l.], v.138, n.3, p.333–408, June 2013.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. **Proceedings of the 23rd international conference on Machine learning - ICML '06**, New York, New York, USA, p.233–240, 2006.

DEMSAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. **Journal of Machine Learning Research**, [S.l.], v.7, p.1–30, 2006.

DIEGO, I. M.; MUNOZ, A.; MOGUERZA, J. M. Methods for the combination of kernel matrices within a support vector framework. **Machine Learning**, [S.l.], v.78, n.1-2, p.137–174, Aug. 2010.

DING, H. et al. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. **Briefings in bioinformatics**, [S.l.], Aug. 2013.

DINUZZO, F. **Learning Functions with Kernel Methods**. 2011. Tese (Doutorado em Ciência da Computação) — University of Pavia.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012.

DUDEK, A. Z.; ARODZ, T.; GáLVEZ, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. **Combinatorial chemistry & high throughput screening**, [S.l.], v.9, n.3, p.213–28, Mar. 2006.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: ncbi gene expression and hybridization array data repository., [S.l.], v.Vol. 30, p.207–210, 2002.

EKINS, S. et al. In silico repositioning of approved drugs for rare and neglected diseases. **Drug discovery today**, [S.l.], v.16, n.7-8, p.298–310, 2011.

ERTL, P.; JELFS, S. Designing drugs on the internet? Free web tools and services supporting medicinal chemistry. **Current topics in medicinal chemistry**, [S.l.], v.7, n.15, p.1491–501, Jan. 2007.

FOUSS, F.; PIROTTE, A. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. **Knowledge and Data ...**, [S.l.], p.1–31, 2007.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the American Statistical Association**, [S.l.], v.32, n.200, p.675–701, 1937.

GOMEZ, S. M.; NOBLE, W. S.; RZHETSKY, a. Learning to predict protein-protein interactions from protein sequences. **Bioinformatics**, [S.l.], v.19, n.15, p.1875–1881, Oct. 2003.

GöNEN, M.; ALPAYDıN, E. Multiple kernel learning algorithms. **The Journal of Machine Learning Research**, [S.l.], v.12, p.2211–2268, 2011.

GöNEN, M.; ALPAYDıN, E. Localized algorithms for multiple kernel learning. **Pattern Recognition**, [S.l.], v.46, n.3, p.795–807, Mar. 2013.

GüNTHER, S. et al. SuperTarget and Matador: resources for exploring drug-target relationships. **Nucleic acids research**, [S.l.], v.36, n.suppl 1, p.D919—-D922, 2008.

GUZZI, P. H. et al. Semantic similarity analysis of protein data: assessment with biological features and issues. **Briefings in Bioinformatics**, [S.l.], v.13, n.5, p.569–585, 2012.

HATTORI, M. et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. **Journal of the American Chemical Society**, [S.l.], v.125, n.39, p.11853–11865, 2003.

HE, J.; CHANG, S. F.; XIE, L. Fast kernel learning for spatial pyramid matching. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR, 26. **Anais...** [S.l.: s.n.], 2008.

HUANG, Q. et al. Diazoxide prevents diabetes through inhibiting pancreatic β-cells from apoptosis via Bcl-2/Bax rate and p38-β mitogen-activated protein kinase. **Endocrinology**, [S.l.], v.148, n.1, p.81–91, 2007.

HUANG, Z. H. Z.; LI, X. L. X.; CHEN, H. C. H. Link prediction approach to collaborative filtering. **Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)**, [S.l.], p.0–1, 2005.

IORIO, F. et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. **PNAS**, [S.l.], v.33, n.107, p.14621–14626, Nov. 2010.

IORIO, F.; TAGLIAFERRI, R.; BERNARDO, D. di. Identifying network of drug mode of action by gene expression profiling. **Journal of computational biology : a journal of computational molecular cell biology**, [S.l.], v.16, n.2, p.241–51, Feb. 2009.

JACOB, L.; VERT, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. **Bioinformatics (Oxford, England)**, [S.l.], v.24, n.19, p.2149–56, Oct. 2008.

JORGENSEN, W. L. The many roles of computation in drug discovery. **Science (New York, N.Y.)**, [S.l.], v.303, n.5665, p.1813–8, Mar. 2004.

JUNKER, B. H.; SCHREIBER, F. **Analysis of biological networks**. [S.l.]: John Wiley & Sons, 2008. v.2.

KANDOLA, J.; SHAWE-TAYLOR, J.; CRISTIANINI, N. Optimizing kernel alignment over combinations of kernel., [S.l.], 2002.

KANEHISA, M. et al. KEGG for linking genomes to life and the environment. **Nucleic acids research**, [S.l.], v.36, n.suppl 1, p.D480—-D484, 2008.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, [S.l.], v.28, n.1, p.27–30, 2000.

KASHIMA, H. et al. On pairwise kernels: an efficient alternative and generalization analysis. **Advances in Knowledge Discovery and Data Mining**, [S.l.], 2009.

KASHIMA, H.; TSUDA, K.; INOKUCHI, A. Marginalized kernels between labeled graphs. **ICML**, [S.l.], n.2002, p.321–328, 2003.

KAWANABE, M.; NAKAJIMA, S.; BINDER, A. A Procedure of Adaptive Kernel Combination with Kernel-Target Alignment for Object Classification. In: ACM INTERNATIONAL CONFERENCE ON IMAGE AND VIDEO RETRIEVAL. **Proceedings...** [S.l.: s.n.], 2009. n.c.

KEISER, M. et al. Predicting new molecular targets for known drugs. **Nature**, [S.l.], v.462, n.7270, p.175–181, 2009.

KEISER, M. J. et al. Relating protein pharmacology by ligand chemistry. **Nature biotechnology**, [S.l.], v.25, n.2, p.197–206, Feb. 2007.

KIMELDORF, G.; WAHBA, G. Some results on Tchebycheffian spline functions. **Journal of Mathematical Analysis and Applications**, [S.l.], v.33, n.1, p.82–95, 1971.

KLOFT, M. et al. Non-sparse multiple kernel learning., [S.l.], p.1–4, 2008.

KLOFT, M.; LASKOV, P.; ZIEN, A. Efficient and Accurate p-Norm Multiple Kernel Learning. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. **Anais...** [S.l.: s.n.], 2009. p.997—-1005.

KONDOR, R.; LAFFERTY, J. Diffusion kernels on graphs and other discrete input spaces. In: ICML. **Anais...** [S.l.: s.n.], 2002.

KUHN, M. et al. Large-scale prediction of drug-target relationships. **FEBS letters**, [S.l.], v.582, n.8, p.1283–90, Apr. 2008.

KUHN, M. et al. A side effect resource to capture phenotypic effects of drugs. **Molecular systems biology**, [S.l.], v.6, n.1, p.343, 2010.

KUMAR, R. et al. Multiple Kernel Completion and its application to cardiac disease discrimination. **Proceedings - International Symposium on Biomedical Imaging**, [S.l.], p.764–767, 2013.

KUNEGIS, J. **On the Spectral Evolution of Large Networks**. 2011. Tese (Doutorado em Ciência da Computação) — . (November).

KUNEGIS, J.; LUCA, E. D.; ALBAYRAK, S. The link prediction problem in bipartite networks. **Computational intelligence for ...**, [S.l.], p.380–389, 2010.

LAARHOVEN, T. van; MARCHIORI, E. A biased look at drug–target interaction data bias., [S.l.], p.1–9, 2013.

LAARHOVEN, T. van; MARCHIORI, E. Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. **PLoS ONE**, [S.l.], v.8, n.6, p.e66952, June 2013.

LAARHOVEN, T. van; NABUURS, S. B.; MARCHIORI, E. Gaussian interaction profile kernels for predicting drug-target interaction. **Bioinformatics (Oxford, England)**, [S.l.], v.27, n.21, p.3036–43, Nov. 2011.

LAMB, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. **Science (New York, N.Y.)**, [S.l.], v.313, n.5795, p.1929–35, Sept. 2006.

LANCKRIET, G. R. G. et al. Learning the Kernel Matrix with Semidefinite Programming. **J. Mach. Learn. Res.**, [S.l.], v.5, p.27–72, Dec. 2004.

LANCKRIET, G. R. G. et al. A statistical framework for genomic data fusion. **Bioinformatics**, [S.l.], v.20, n.16, p.2626–2635, 2004.

LAUB, A. J. **Matrix Analysis for Scientists and Engineers**. Davis, California: SIAM, 2005. 139–144p.

LESLIE, C.; ESKIN, E.; NOBLE, W. S. The spectrum kernel: a string kernel for svm protein classification. **Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**, [S.l.], p.564–75, Jan. 2002.

LESLIE, C.; WESTON, J.; NOBLE, W. S. Mismatch String Kernels for SVM Protein Classification. In: NIPS. **Anais...** [S.l.: s.n.], 2002. p.1441—-1448.

LI, X.; CHEN, H. Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach. **Decision Support Systems**, [S.l.], v.54, n.2, p.880–890, Jan. 2013.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **Journal of the American Society for Information Science and Technology**, [S.l.], v.58, n.7, p.1019–1031, May 2007.

LIU, H. et al. Improving compound-protein interaction prediction by building up highly credible negative samples. **Bioinformatics**, [S.l.], v.31, n.12, p.i221–i229, 2015.

LIU, X. et al. Absent Multiple Kernel Learning Algorithms., [S.l.], n.JANUARY, p.1–12, 2013.

LONGWORTH, C.; GALES, M. J. F. Multiple kernel learning for speaker verification. In: ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2008. ICASSP 2008. IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2008. p.1581–1584.

Lü, L.; ZHOU, T. Link prediction in complex networks: a survey. **Physica A: Statistical Mechanics and its Applications**, [S.l.], n.October 2010, 2011.

MAHR, M.; KLAMBAUER, G.; HOCHREITER, S. **Rchemcpp: an r package for computing the similarity of molecules**. [S.l.]: Unpublished, 2012.

MASON, O.; VERWOERD, M. Graph theory and networks in Biology. **IET systems biology**, [S.l.], v.1, n.2, p.89–119, Mar. 2007.

MATLAB. **version 8.1.0 (R2013a)**. Natick, Massachusetts: The MathWorks Inc., 2013.

MEI, J.-P. et al. Drug-target interaction prediction by learning from local information and neighbors. **Bioinformatics (Oxford, England)**, [S.l.], v.29, n.2, p.238–45, Jan. 2013.

MENON, A.; ELKAN, C. Link prediction via matrix factorization. In: EUROPEAN CONFERENCE, ECML PKDD. **Anais...** [S.l.: s.n.], 2011. p.437–452.

MOGUERZA, J. M.; MUñOZ, A.; DIEGO, I. M. de. Improving Support Vector Classification via the Combination of Multiple Sources of Information. In: SSPR/SPR. **Anais...** Springer, 2004. p.592–600. (Lecture Notes in Computer Science, v.3138).

MOUSAVIAN, Z.; MASOUDI-NEJAD, A. Drug-Target Interaction Prediction via Chemogenomic Space : learning-based methods. **Expert opinion on drug metabolism & toxicology**, [S.l.], v.10, n.9, p.1273–1287, 2014.

NGUYEN, C. H.; HO, T. B. Kernel matrix evaluation. In: IJCAI INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Anais...** [S.l.: s.n.], 2007. n.2, p.987–992.

NGUYEN, C. H.; MAMITSUKA, H. Latent Feature Kernels for Link Prediction on Sparse Graphs. **IEEE Transactions on Neural Networks and Learning Systems**, [S.l.], v.23, n.11, p.1793–1804, Nov. 2012.

OVASKA, K.; LAAKSO, M.; HAUTANIEMI, S. Fast Gene Ontology based clustering for microarray experiments. **BioData mining**, [S.l.], v.1, n.1, p.11, 2008.

PAHIKKALA, T. et al. Efficient Regularized Least-Squares Algorithms for Conditional Ranking on Relational Data. **Machine Learning**, [S.l.], n.93, p.321–356, 2013.

PAHIKKALA, T. et al. Toward more realistic drug-target interaction predictions. **Briefings in Bioinformatics**, [S.l.], Apr. 2014.

PAHIKKALA, T.; WAEGEMAN, W. Conditional ranking on relational data. In: ECML PKDD'10 PROCEEDINGS OF THE 2010 EUROPEAN CONFERENCE ON MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES: PART II, Barcelona, Spain. **Anais...** Springer-Verlag Berlin: Heidelberg, 2010. p.499–514.

PALME, J.; BODENHOFER, U. **KeBABS - An R Package for Kernel Based Analysis of Biological Sequences.** Linz, Austria: Institute of Bioinformatics, Johannes Kepler University, 2014.

PAVLIDIS, P. et al. Gene Functional Classification From Heterogeneous Data. In: IN PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL MOLECULAR BIOLOGY. **Anais...** [S.l.: s.n.], 2001. p.249–255.

PERLMAN, L. et al. Combining drug and gene similarity measures for drug-target elucidation. **Journal of Computational Biology**, [S.l.], v.18, n.2, p.133–45, Feb. 2011.

QIU, S.; LANE, T. A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction. **Computational Biology and Bioinformatics, IEEE/ACM Transactions on**, [S.l.], v.6, n.2, p.190–199, 2009.

RAKOTOMAMONJY, A.; BACH, F. SimpleMKL. **Journal of Machine ...**, [S.l.], v.9, p.2491–2521, 2008.

RALAIVOLA, L. et al. Graph kernels for chemical informatics. **Neural networks : the official journal of the International Neural Network Society**, [S.l.], v.18, n.8, p.1093–110, Oct. 2005.

RAYMOND, R.; KASHIMA, H. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In: **Machine Learning and Knowledge Discovery in Databases**. [S.l.: s.n.], 2010. v.3, p.131–147.

REIS, A.; VELHO, G. Sulfonylurea receptor-1 (SUR1): genetic and metabolic evidences for a role in the susceptibility to type 2 diabetes mellitus. **Diabetes & metabolism**, [S.l.], v.28, n.1, p.14–19, 2002.

RESNIK, P. Semantic Similarity in a Taxonomy: an information based measure and its application to problems of ambiguity in natural language. **Journal of Aritificial Intelligence Research**, [S.l.], v.11, p.95–130, 1999.

RIFKIN, R.; YEO, G.; POGGIO, T. Regularized Least-Squares Classification. **Nato Science Series Sub Series III Computer and Systems Sciences**, [S.l.], v.190, p.131—-154, 2003.

SALWINSKI, L. et al. The database of interacting proteins: 2004 update. **Nucleic acids research**, [S.l.], v.32, n.suppl 1, p.D449—-D451, 2004.

SAWADA, R.; KOTERA, M.; YAMANISHI, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. **Molecular Informatics**, [S.l.], v.33, n.11-12, p.719–731, 2014.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with Kernels: support vector machines, regularization, optimization, and beyond**. Cambridge, MA, USA: MIT Press, 2001.

SCHöLKOPF, B.; TSUDA, K.; VERT, J.-P. **Kernel methods in computational biology**. Cambridge, Mass.: MIT Press, 2004.

SMEDLEY, D. et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. **Nucleic Acids Research**, [S.l.], 2015.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of molecular biology**, [S.l.], v.147, n.1, p.195–197, 1981.

SMOLA, A.; KONDOR, R. Kernels and regularization on graphs. **Learning theory and kernel machines**, [S.l.], p.1–15, 2003.

STARK, C. et al. BioGRID: a general repository for interaction datasets. **Nucleic acids research**, [S.l.], v.34, n.suppl 1, p.D535—-D539, 2006.

TAKARABE, M. et al. Drug target prediction using adverse event report systems: a pharmacogenomic approach. **Bioinformatics**, [S.l.], v.28, n.18, p.611–618, 2012.

TANABE, H. et al. Simple but effective methods for combining kernels in computational biology. **2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies**, [S.l.], p.71–78, 2008.

TANG, Y.; ZHANG, Y. Q.; CHAWLA, N. V. SVMs modeling for highly imbalanced classification. **IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics**, [S.l.], v.39, n.1, p.281–288, 2009.

TUNCBAG, N. et al. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. **Briefings in bioinformatics**, [S.l.], v.10, n.3, p.217–32, May 2009.

VISHWANATHAN, S. V. N.; SCHRAUDOLPH, N. N. Graph Kernels. **Journal of Machine Learning Research**, [S.l.], v.11, p.1201–1242, 2010.

WANG, H. et al. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. **Journal of computational biology : a journal of computational molecular cell biology**, [S.l.], v.20, n.4, p.344–58, Apr. 2013.

WANG, K. et al. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. **PLoS computational biology**, [S.l.], v.9, n.11, p.e1003315, Nov. 2013.

WANG, T.; XIE, H.; HU, S. A Heuristic Kernel Combination Approach Based on Kernel Fisher Criterion. **Journal of Information and Computational Science**, [S.l.], v.10, n.9, p.2799–2806, 2013.

WANG, Y.-C. et al. Kernel-based data fusion improves the drug-protein interaction prediction. **Computational biology and chemistry**, [S.l.], v.35, n.6, p.353–62, Dec. 2011.

WANG, Y. et al. PubChem: a public information system for analyzing bioactivities of small molecules. **Nucleic acids research**, [S.l.], v.37, n.suppl 2, p.W623—-W633, 2009.

WANG, Y.; ZENG, J. Predicting drug-target interactions using restricted Boltzmann machines. **Bioinformatics (Oxford, England)**, [S.l.], v.29, n.13, p.i126–i134, July 2013.

WEBSTER, G. F. Topical tretinoin in acne therapy. **Journal of the American Academy of Dermatology**, [S.l.], v.39, n.2, p.S38–S44, 1998.

WISHART, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. **Nucleic acids research**, [S.l.], v.36, n.suppl 1, p.D901—-D906, 2008.

YAJIMA, Y. One-class support vector machines for recommendation tasks. In: **Advances in Knowledge Discovery and Data Mining**. [S.l.]: Springer, 2006. p.230–239.

YAMANISHI, Y. Chemogenomic Approaches to Infer Drug–Target Interaction Networks. **Data Mining for Systems Biology**, Totowa, NJ, v.939, p.97–113, 2013.

YAMANISHI, Y. et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. **Bioinformatics (Oxford, England)**, [S.l.], v.24, n.13, p.i232–40, July 2008.

YAMANISHI, Y. et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. **Bioinformatics (Oxford, England)**, [S.l.], v.26, n.12, p.i246–i254, June 2010.

YAN, F. et al. A comparison of L1 norm and L2 norm multiple kernel SVMs in image and video classification. In: CONTENT-BASED MULTIMEDIA INDEXING, 2009. CBMI'09. SEVENTH INTERNATIONAL WORKSHOP ON. **Anais...** [S.l.: s.n.], 2009. p.7–12.

YANG, F.; XU, J.; ZENG, J. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. **Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**, [S.l.], p.148–59, 2014.

YANG, J. et al. A new multiple kernel approach for visual concept learning. **Advances in Multimedia Modeling**, [S.l.], p.250–262, 2009.

YANG, J. et al. Group-sensitive multiple kernel learning for object recognition. **IEEE transactions on image processing : a publication of the IEEE Signal Processing Society**, [S.l.], v.21, n.5, p.2838–52, May 2012.

YILDIRIM, M. a. et al. Drug-target network. **Nature biotechnology**, [S.l.], v.25, n.10, p.1119–26, Oct. 2007.

YING, Y.; HUANG, K.; CAMPBELL, C. Enhanced protein fold recognition through a novel data integration approach. **BMC bioinformatics**, [S.l.], v.10, p.267, 2009.

YU, H. et al. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. **PloS one**, [S.l.], v.7, n.5, p.e37608, Jan. 2012.

YU, S. et al. L2-norm multiple kernel learning and its application to biomedical data fusion. **BMC bioinformatics**, [S.l.], v.11, p.309, 2010.

ZHAO, S.; LI, S. Network-based relating pharmacological and genomic spaces for drug target identification. **PloS one**, [S.l.], v.5, n.7, p.e11764, Jan. 2010.

Apêndice

.1 Notação e conceitos básicos de álgebra linear

Álgebra linear é a área da matemática cujo objetivo é o estudo de matrizes e vetores, estruturas que podem ser utilizadas para representar muitos tipos diferentes de objetos e estruturas. Este trabalho concentra-se no estudo de redes complexas, as quais podem ser representadas na forma de matrizes. Dessa forma, algumas noções básicas de álgebra linear serão apresentadas nesta seção. Entretanto, vale ressaltar que não consiste em uma introdução à área, apenas a definição de conceitos utilizados no decorrer deste trabalho.

Um vetor $\mathbf{x} \in \mathbb{R}^n$ de tamanho n é composto por n números reais. Neste trabalho, representaremos vetores em letra minúscula e negrito, por exemplo, $\mathbf{u}, \mathbf{v}$, etc. O i-ésimo elemento do vetor será referenciado por $\mathbf{x}_i$, com $i \in \{1, 2, ..., n\}$. A norma de um vetor $\mathbf{x} \in \mathbb{R}^n$ é definida por:

$$\| \mathbf{c} \| = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{.1}$$

Se $\| \mathbf{x} \| = 1$ o vetor é dito unitário. O produto interno de dois vetores $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ é definido por:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i \cdot y_i. \tag{.2}$$

Se $\mathbf{x} \cdot \mathbf{y} = 0$ os vetores são chamados de ortogonais.

Uma matriz $X \in \mathbb{R}^{m \times n}$ de tamanho $m \times n$ consiste de mn números reais organizados por índices $i \in \{1, ..., m\}$ e $j \in \{1, ..., n\}$. Os elementos na i-ésima linha e j-ésima coluna de X são referenciados por x_{ij} ou $(X)_{ij}$. Matrizes serão representadas por letras maiúsculas, tais como A, B, etc. Uma matriz é dita quadrada se $m = n$. A norma de Frobenius de uma matriz $X \in \mathbb{R}^{m \times n}$ é definida como:

$$\| X \| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2}. \tag{.3}$$

A matriz transposta X^T de X é definida por $(X^T)_{ij} = x_{ji}$. Uma matriz X é dita simétrica se $X = X^T$, e são necessariamente quadradas. Uma matriz é diagonal se $x_{ij} = 0$ sempre que $i \neq j$. Algumas matrizes possuem uma nomenclatura e notação especiais, como, por exemplo, a matriz unitária $I_{n \times n}$, a qual é uma matriz diagonal de tamanho $n \times n$ definida por $(I_{n \times n})_{ii} = 1$ para todo $i \in \{1, ..., n\}$. As matrizes $0_{m \times n}$ e $1_{m \times n}$ são matrizes compostas de apenas zeros e de uns, respectivamente, de dimensões $m \times n$.

Uma matriz simétrica $S \in \mathbb{R}^{n \times n}$ é chamada de positiva semidefinida, denotada $S \succeq 0$ se

$$\mathbf{p}^T S \mathbf{p} \geq 0, \tag{.4}$$

para todo $\mathbf{p} \in \mathbb{R}^n$.

Traço de uma matriz

O traço de uma matriz $A \in \mathbb{R}^{n \times n}$ é definido como a soma dos elementos da diagnoal principal de A, i.e.,

$$tr(A) = a_{11} + a_{22} + ... + a_{nn} = \sum_{i=1}^{n} a_{ii}. \tag{.5}$$

Decomposição de matrizes

Definição .1. Dizemos que um vetor $\mathbf{v}$ de dimensão n é um autovetor de uma matriz simétrica $A \in \mathbb{R}^{n\times n}$, se somente se a seguinte equação linear for satisfeita:

$$A\mathbf{v} = \lambda\mathbf{v} \tag{.6}$$

onde λ é um escalar chamado de autovalor correspondente de $\mathbf{v}$. Dizemos então que A pode ser fatorada como:

$$A = Q\Lambda Q^T \tag{.7}$$

onde Q é uma matriz quadrada $(n \times n)$, cuja i-ésima coluna corresponde ao autovetor q_i de A, e Λ é uma matriz diagonal cuja diagonal principal contém os autovalores correspondentes, $\Lambda_{ii} = \lambda_i$.

O Produto de Hadamard

Definição .2. Sejam A e B matrizes de mesma dimensão $m \times n$. O produto de Hadamard $A \circ B$ é a matriz $m \times n$ definida por:

$$(A \circ B)_{ij} = (A)_{ij}(B)_{ij} \tag{.8}$$

O produto de Hadamard é comutativo, associativo e distributivo sobre a adição. É importante ressaltar que o produto de Hadamard de duas matrizes positivas-semidefinidas é positiva-semidefinida.

O produto de Kronecker

Definição .3. Sejam A uma matriz $m \times n$ e B uma matriz $p \times q$. O produto de Kronecker $A \otimes B$ é a matriz $mp \times nq$:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{.9}$$

A soma de Kronecker

Definição .4. A soma de Kronecker de duas matrizes quadradas $A \in \mathbb{R}^{m\times m}$ e $B \in \mathbb{R}^{n\times n}$ é definida como:

$$A \oplus B = (I_n \otimes A) + (B \otimes I_m) \tag{.10}$$

O operador *vec*

Definição .5. Seja A uma matriz $m \times n$. O operador *vec* é definido como sendo o vetor resultante do empilhamento das colunas de A, ou seja:

$$vec(A) = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{mn} \end{bmatrix} \tag{.11}$$