MINERAÇÃO DE DADOS USANDO ÁLGEBRA LINEAR PARA A PREDIÇÃO DE ALVOS DROGÁVEIS

EDUARDO CAMPOS DOS SANTOS

MINERAÇÃO DE DADOS USANDO ÁLGEBRA LINEAR PARA A PREDIÇÃO DE ALVOS DROGÁVEIS

Tese apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Doutor em Bioinformática.

Orientador: Júlio César Dias Lopes
Co-orientador: Marcos Augusto dos Santos

Belo Horizonte
Agosto de 2012

dos Santos, Eduardo Campos
Mineração de dados usando álgebra linear para a predição de alvos drogáveis / Eduardo Campos dos Santos. — Belo Horizonte, 2012
xx, 104 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Júlio César Dias Lopes
Co-orientador: Marcos Augusto dos Santos

1. predição de alvos drogáveis. 2. mineração de dados. 3. SVD. 4. decomposição por valores singulares. 5. recuperação de informação. 6. regressão logística. I. Título.

CDU

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"Mineração de Dados usando Álgebra Linear para a Predição de Alvos "Drogáveis""

Eduardo Campos dos Santos

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Vasco Ariston de Carvalho Azevedo
UFMG

Prof. Marcelo Matos Santoro
UFMG

Prof. Lucas Bleicher
UFMG

Prof. Marcos Augusto dos Santos
UFMG

Prof. Roney Santos Coimbras
FIOCRUZ

Prof. Floriano Paes Silva Junior
FIOCRUZ

Belo Horizonte, 01 de agosto de 2012.

Dedico este trabalho ao meu filhinho - o Pedrinho.
Espero ser para você um exemplo de dedicação, trabalho e perseverança.
E que você se torne um rapaz culto e inteligente.

Agradecimentos

Agradeço ao Dr. Julio Lopes pelas orientações para a coleta de dados na etapa inicial do projeto; ao Dr. Bráulio Roberto Gonçalves Marinho Couto pela orientação sobre os métodos estatísticos; ao Dr. Marcos Augusto dos Santos pela orientação sobre os métodos algébricos e computacionais a partir do início da segunda etapa desta longa jornada e no fatídigo processo de produção e sucessivas revisões do texto na crucial etapa final; novamente ao Bráulio e ao Marcos por todo esforço investido para conseguirmos produzir, publicar e apresentar artigos no decorrer do processo; aos demais professores e colegas que, de alguma forma, apoiaram-me nos não raros momentos de dificuldade; em especial, ao amigo e colega Sérgio de Alencar pelas palavras de incentivo e motivação para vencer certos obstáculos; aos professores Dra. Glória Franco e Dr. Vasco Azevedo pelo apoio durante a época em que atuaram como coordenadores do programa; aos funcionários que passaram pela secretaria durante o tempo deste projeto desempenhando um eficiente e importante trabalho (em especial ao Carlos e à Sheila); à Fapemig pela bolsa de fomento à pesquisa a mim destinada; ao amigo Dr. Paulo Eduardo Behrens que, do início ao fim deste projeto, sempre ofereceu-me apoio essencial para dar-me condições para realizar o trabalho; aos meus pais, Marcos e Nely, pela educação e formação que me propiciaram; à minha querida esposa, Rejanni, pelo carinhoso apoio nos momentos difíceis e pela paciência para aturar meus momentos de agrura; ao meu filhinho Pedrinho, que com menos de dois anos de idade é capaz de levantar minha moral e revigorar meus ânimos simplesmente com seu humor e carinho.

Resumo

Apresenta-se o desenvolvimento de um método para recuperar proteínas que são alvos drogáveis. A partir da representação desses alvos como vetores definidos a partir das anotações do InterPro, instrumentos da álgebra linear relacionados com a decomposição por valores singulares são utilizados para organizar semanticamente o espaço vetorial e permitir a recuperação eficiente das proteínas similares a uma dada consulta. Relações não observadas *prima facie* são descortinadas indicando, oportunidades para reposicionamento de fármacos conhecidos, estratégias para o desenvolvimento racional de novos compostos e a predição de potenciais alvos drogáveis e de efeitos colaterais latentes. As assinaturas do InterPro mais relevantes para discriminar alvos drogáveis e não-drogáveis foram determinadas por regressão logística. Os resultados são avaliados estatisticamente por análise de curvas ROC e dados corroborados em outros trabalhos.

Palavras-chave: descoberta de medicamento, desenvolvimento de medicamento, predição de alvos drogáveis, reposicionamento de medicamento, mineração de dados, álgebra linear, recuperação de informação latente, decomposição por valores singulares, indexação semântica latente, regressão logística, estudo de caso-controle.

Abstract

This work presents the development of a method for recovering target proteins that are druggable. From the representation of drug targets defined as vectors by using InterPro annotations, tools of linear algebra related to singular value decomposition are used to organize the semantic vector space and allow the efficient recovery of proteins related to a given query. Not *prima facie* relationships arise and indicate drug repositioning opportunities, rational development strategies and, the prediction of potential druggable targets and latent side-effects. The InterPro signatures which are most relevant to drug target/non-drug target discriminating were selected by logistic regression. The results are statistically evaluated by ROC curves analysis and data corroborated in the literature.

Keywords: drug discovery, drug development, druggable target prediction, drug repositioning, data mining, linear algebra, latent information retrieval, singular value decomposition, latent semantic indexing, logistic regression, case-control study.

Lista de Figuras

1.1 O processo de desenvolvimento de um novo fármaco 2
1.2 Mudança de paradigma no desenvolvimento de um novo fármaco 3

4.1 Coleta dos dados em bases públicas e armazenamento no banco de dados local 16
4.2 Arquivo de dados do TTD 20
4.3 Exemplo de arquivo do UniProtKB 26
4.4 Lista de tabelas do InterPro 28
4.5 Parte do esquema da base de dados do InterPro 29
4.6 Dados sobre alguns termos do InterPro 29
4.7 Dados de algumas anotações do InterPro para sequências proteicas humanas 30
4.8 Decomposição por valores singulares e a posterior redução de posto 33
4.9 Melhoria da recuperação de informação através da redução de posto 34
4.10 O teste de *scree* 35
4.11 Exemplo de dendrograma 39
4.12 Mapas de calor 40
4.13 Redes 42
4.14 Fluxograma para a costrução do modelo vetorial para alvos drogáveis . . . 44
4.15 O teste de *scree test* aplicado à matriz reduzida de alvos humanos 46
4.16 Fluxograma para a construção do modelo probabilístico 56
4.17 Fluxograma: integração de dados químicos, biológicos e farmacológicos . . 57
4.18 Correlação entre as métricas obtidas por SVD/redução de posto e o coeficiente de Tanimoto 58

5.1 Correlação entre o *bitscore* do BLAST e o coeficiente de dissimilaridade obtido pelo modelo. 61
5.2 Resultado do algoritmo de agrupamento hierárquico – conjunto completo de dados 62

5.3 Representação dos dados em espaço tridimensional 63
5.4 Resultado do HCL para amostra 42 proteínas – SVD 64
5.5 Resultado do HCL para amostra 42 proteínas – BLAST 65
5.6 Resultado do agrupamento por NMF sobre a matriz de similaridade vetorial 66
5.7 Resultado do agrupamento por NMF sobre a matriz de similaridade sequência 66
5.8 Visão expandida de uma região da Figura 5.2-a 67
5.9 Visão expandida de outra região da Figura 5.2-a 68
5.10 Visão expandida de uma terceira região da Figura 5.2-a 69
5.11 Visualização dos dados no espaço tridimensional 70
5.12 Comparação do *boxplot* para drogáveis e não-drogáveis 72
5.13 Análise de curva ROC para o modelo vetorial 73
5.14 Exemplo de recuperação de informação implícita 75
5.15 Visualização dos dados em uma rede. 78
5.16 Curva ROC para o método probabilístico 80
5.17 Rede de relacionamento entre fármacos 86
5.18 Rede de relacionamento entre fármacos (acrescentados vasodilatadores) . . 87

Lista de Tabelas

4.1 Cobertura da versão 22.0 do InterPro 28
4.2 Tabela de contingência 2x2 para duas variáveis dicotômicas 48
4.3 Análise univariada para o InterPro IPR001828 53
4.4 Análise univariada para o InterPro IPR016175 53

5.1 Dados da análise de curva ROC para o modelo vetorial 72
5.2 Potenciais alvos de *P. falciparum* e *T. gondii* para o agente anti-obesidade orlistat 76
5.3 Modelo probabilístico para predição da drogabilidade de proteínas humanas 79
5.4 Termos do InterPro (positivos) mantidos no modelo probabiístico 81
5.5 Termos do InterPro (negativos) mantidos no modelo probabiístico 82
5.6 Similaridade entre fármacos pela abordagem usando SVD sobre o espaço químico 84
5.7 Similaridade entre fármacos pela abordagem usando SVD sobre o espaço químico e biológico 85

Sumário

Agradecimentos **ix**

Resumo **xi**

Abstract **xiii**

Lista de Figuras **xv**

Lista de Tabelas **xvii**

1 Introdução **1**

1.1 O proteoma drogável e o índice de drogabilidade 4

1.2 Trabalhos relacionados 5

2 Justificativa **9**

3 Objetivos **11**

3.1 Objetivo geral 11

3.2 Objetivos específicos 11

4 Materiais e Métodos **13**

4.1 Coleta, armazenamento e integração de dados 15

4.2 Bancos de dados de públicos 17

4.2.1 DrugBank 17

4.2.2 *Therapeutical Target Database* (TTD) 19

4.2.3 *Kyoto Encyclopedia of Genes and Genomes* (KEGG) 22

4.2.4 *UniProt Knowledgebase* (UniProtKB) 24

4.2.5 InterPro 27

4.2.6 DEG - Database of Essential Genes 30

4.3 Técnicas para recuperação de informação 30

4.3.1 Decomposição por valores singulares 31
4.3.2 Métricas de similaridade 36
4.4 Agrupamento e visualização de dados 36
4.4.1 Algoritmo de agrupamento hierárquico 37
4.4.2 Algoritmo de agrupamento por fatoração de matriz não-negativa 38
4.4.3 Dendrograma e coeficiente de correlação cofenética 39
4.4.4 Mapas de calor (*heatmaps*) 40
4.4.5 Redes 41
4.4.6 Projeção em espaço de duas ou três dimensões 43
4.5 Representação vetorial dos alvos 43
4.6 Regressão logística 46
4.7 Construção do modelo probabilístico para a predição da drogabilidade de um alvo proteico humano 52
4.8 Outros modelos vetoriais: integração de dados químicos, farmacológicos e biológicos 54

5 Resultados e discussão **59**
5.1 Predição de alvos drogáveis usando regressão logística 78
5.1.1 Classificação de alvos usando regressão logística 80
5.2 Modelos para a representação de fármacos integrando dados químicos e biológicos 82

6 Conclusões **89**
6.1 Perspectivas 91

Referências Bibliográficas **93**

Capítulo 1

Introdução

Estudos têm estimado que o custo para lançar um novo medicamento varia entre 500 milhões e 2 bilhões de dólares [DiMasi et al., 2003]. Ainda que este valor seja um assunto controverso, considerado muitas vezes superestimado [Light & Warburton, 2011; Goozner, 2004; Riggs, 2004], o certo é que a necessidade de altos investimentos desestimulam a pesquisa de agentes para novos alvos terapêuticos. A situação é ainda mais preocupante ao considerarmos as doenças tropicais negligenciadas [Trouiller et al., 2002].

Essa realidade sugere a necessidade de criar estratégias que tirem maior proveito do conhecimento atual: alvos similares aos já conhecidos podem representar novas oportunidades isentas de patentes; descobertas de novas aplicações para fármacos conhecidos representam também uma boa abordagem.

No processo tradicional de desenvolvimento de um novo fármaco, inicialmente, testes *in vitro* ou *in silico* são utilizados para selecionar compostos que modulam a função biológica do alvo de maneira desejada. Posteriormente, passa-se à fase dos testes *in vivo* em animais modelo, quando experimentos são realizados para avaliar as características ADMET (Administração, Distribuição, Metabolismo, Excreção e Toxicidade). Depois é que o fármaco é administrado em pequenas dosagens em um grupo relativamente pequeno (de 10 a 100) de voluntários saudáveis, para determinar uma dosagem segura e estudar a farmacocinética e a farmacodinâmica do composto selecionado em humanos. O padrão de efeitos colaterais (quando aceitáveis) pode ser obtido nesta etapa. Sendo bem sucedido, ele é testado em pacientes que não apresentam complicações extras em seu estado de saúde; eles estão, em geral, no estágio inicial da doença e não sofrem ainda de implicações secundárias causadas por ela. Além disso, restringe-se ao máximo a relação de medicamentos que podem ser administrados concomitantemente. Por exemplo, pacientes com diabetes diagnosticada em estágio inicial

e ainda não tratada e que não apresentam evidências de dano em algum órgão, podem fazer parte do grupo para testar um novo agente antidiabético. As interações entre fármacos são criteriosamente observadas nesta fase. Se o medicamento não for descartado até este ponto, é porque ele apresentou evidências suficientes de eficácia e não apresentou complicações. É quando o composto é testado na escala de milhares de indivíduos, distribuídos em uma população diversificada e em condições mais complexas quanto ao estado de saúde e uso de outras terapias concomitantes. Finalmente, após o lançamento no mercado, o produto é acompanhado para vigilância e eventual revisão (Figura 1.1).

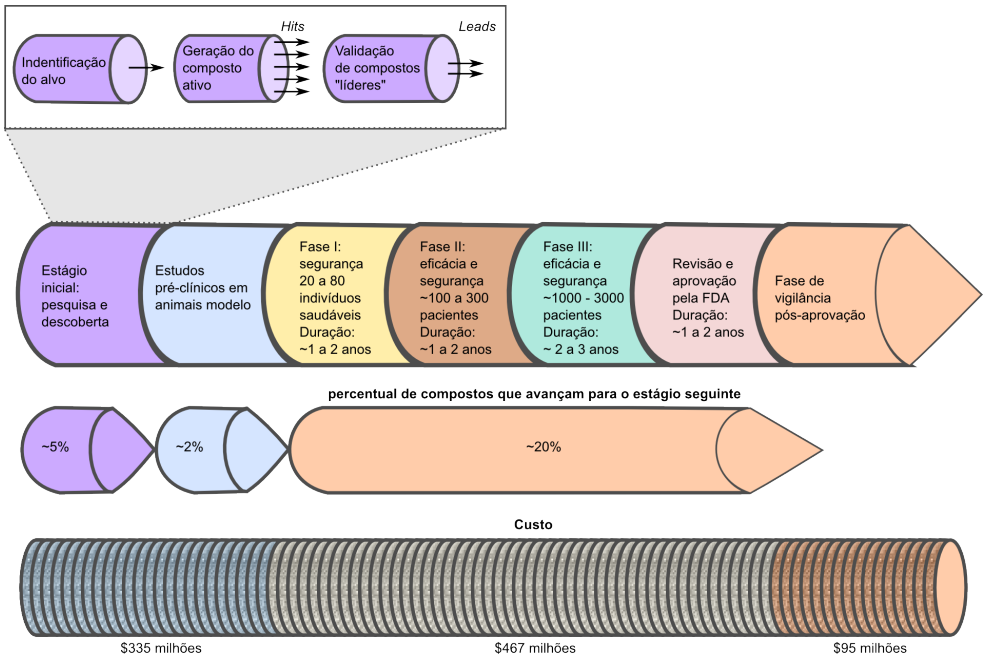


Figura 1.1. As etapas do processo de desenvolvimento de um novo fármaco. Exibe-se as estimativas, no decorrer do processo, do tempo decorrido, do percentual de casos de insucesso e dos custos do investimento segundo DiMasi et al. [2003]. O processo é demorado e dispendioso. Além disso, é estatisticamente mais propício ao fracasso do que ao sucesso. O tempo médio para um novo fármaco obter aprovação no mercado é entre 12 e 15 anos e somente um em 5000 compostos experimentados no estágio inicial torna-se um caso de sucesso no mercado. A maioria das falhas ocorrem na fases iniciais de pesquisas *in vitro* e estudos pré-clínicos em animais modelos vivos. Mas dado que apenas 20% dos compostos que entram nas fases de estudos em humanos recebem aprovação para serem lançados no mercado e ainda que os custos aumentam significativamente nas etapas mais avançadas, é importante tentar predizer os potenciais casos de insucesso nas fases clínicas. Abordagens computacionais têm sido desenvolvidas para este fim. Essas técnicas procuram auxiliar na seleção de compostos e de alvos. Figura adaptada de [O'Driscoll, 2004]

Avanços na biologia molecular, no que tange aos processos de anotação funcional

de genes (proteínas), no conhecimento dos processos metabólicos e patológicos, provocaram uma mudança no paradigma de desenvolvimento de novos fármacos, passando do "foco no ligante" para o "foco no alvo" [Raman et al., 2008; Harland & Gaulton, 2009]. A Figura 1.2 ilustra este fato, onde o processo anterior, muitas vezes encarado como uma "caixa preta", tinha como ponto de partida a análise da sintomatologia e conhecimento dos efeitos moleculares da doença [Raman et al., 2008]. Esse novo paradigma implicou em uma crescente busca por novas abordagens para auxiliar os processos de descoberta e validação de novos mecanismos relacionados às doenças. As mais recentes sugerem o desenvolvimento de compostos que interajam com diversos alvos, ao invés de otimizar a especificação da ação [Hopkins, 2008; Hopkins et al., 2006; Frantz, 2005], determinando uma outra mudança de paradigma, onde a ideia é desenvolver compostos "promíscuos" que modulem diversos alvos.

O objetivo deste trabalho é desenvolver ferramentas e modelos para explorar dados referentes aos alvos. Segundo Harland & Gaulton [2009], essa deverá ser a prioridade da indústria nos próximos anos.

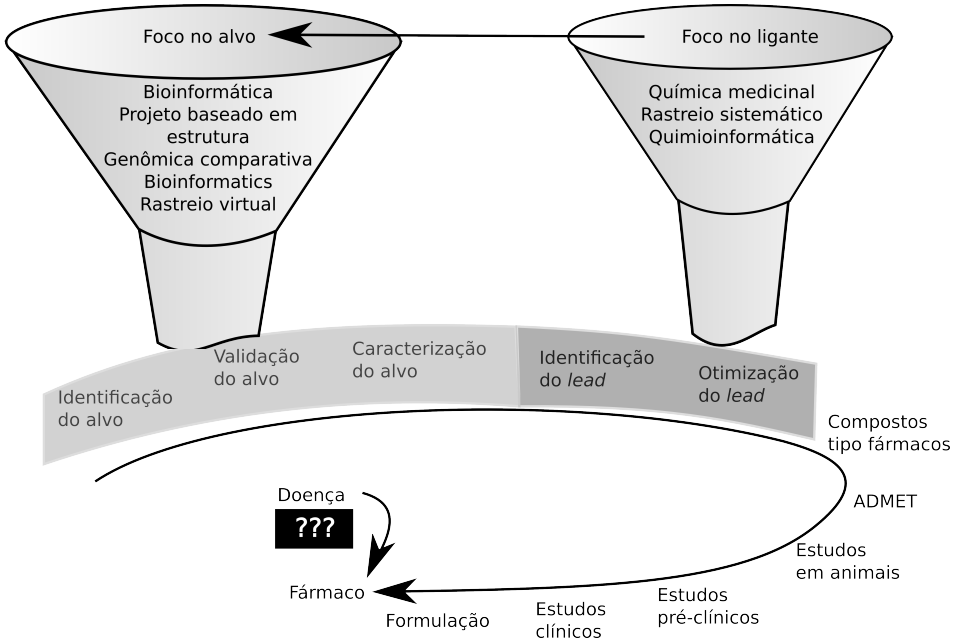


Figura 1.2. Mudança de paradigma no desenvolvimento de um novo fármaco. O processo de desenvolvimento de um novo fármaco e a mudança de paradigma – do modelo "doença a fármaco" inicial para o baseado no "foco no ligante" e deste para o baseado no "foco no alvo". Figura adaptada de [Raman et al., 2008].

1.1 O proteoma drogável e o índice de drogabilidade

Fármacos administrados por via oral permanecem como terapias preferenciais em tratamentos não-tópicos porque apresentam custos mais baixos de produção e armazenamento. Além disso, são mais facilmente administrados – em geral, o próprio paciente pode manipular e ingerir um medicamento por via oral. Medicamentos injetáveis requerem usualmente a atuação de um farmacêutico, enfermeiro ou pessoa com habilidade para realizar o procedimento. Esse panorama determina um conjunto de características desejáveis para compostos com propriedades terapêuticas e essas catacterísticas definem os chamados compostos drogáveis. Os alvos de interesse são aqueles (geralmente protéicos) que podem sofrer influência dessas moléculas. O emprego do termo "drogável", cunhado originalmente para designar moléculas pequenas de interesse em pesquisas farmacológicas, foi estendido para descrever as proteínas às quais essas moléculas podem se ligar [Harland & Gaulton, 2009].

Deve-se enfatizar que nem todo alvo drogável pode ser identificado como alvo para fins terapêuticos, uma vez que nem todo alvo drogável tem associação com algum processo relativo a alguma doença ou condição de saúde. Na verdade, a característica "drogável" pode até ser considerada, em alguns casos, como indesejável, visto que ela pode determinar a ocorrência de um efeito adverso.

Alvos drogáveis que apresentam características para serem abordados em aplicações terapêuticas são denominados "alvos terapêuticos". Outras características, como especificidade e promiscuidade, são atrativas conforme o propósito da aplicação terapêutica em desenvolvimento.

A predição sobre a drogabilidade de proteínas pode ser implementada a partir de um modelo baseado em suas propriedades bioquímicas ou a partir de um modelo baseado na estrutura tridimensional de toda a proteína ou de seus sítios de ligação.

A partir das proteínas drogáveis conhecidas pode-se tentar identificar outras proteínas com propriedades similares que podem constituir um conjunto de potenciais alvos drogáveis. Com efeito, tem-se observado diferentes esforços que buscam identificar completamente o "genoma drogável" ou o "proteoma drogável" [Hopkins & Groom, 2002]. Além de estudos para identificação das proteínas drogáveis em um dado proteoma, encontram-se na literatura iniciativas para se quantificar a "drogabilidade" das proteínas, levando ao conceito do "índice de drogabilidade". Harland & Gaulton [2009] cita a importância desse tipo de índice para comparar o grau de drogabilidade entre todas as proteínas identificadas como drogáveis.

Diversos trabalhos têm surgerido formas de se determinar um índice para quantificar a drogabilidade em um dado proteoma. Em geral, as abordagens propostas estimam a drogabilidade a partir da análise da estrutura tridimensional das proteínas ou, mais especificamente, de seus sítios ativos [Hajduk et al., 2005; Cheng et al., 2007; Schmidtke & Barril, 2010]. A identificação das proteínas drogáveis pode auxiliar também no estudo sobres as propriedades estruturais ou biológicas que caracterizam os alvos de interesse e, a partir dessas informações, propor um desenvolvimento racional de um novo fármaco [Raman et al., 2008].

1.2 Trabalhos relacionados

Ao iniciar este projeto, não foram encontrados modelos para predição da drogabilidade de alvos a partir de suas anotações funcionais. Os modelos existentes são baseados em cálculos estruturais e se restringem aos alvos com estrutura resolvida.

Diferentes abordagens podem ser empregadas no processo de seleção de um bom candidato a alvo terapêutico. Destacam-se nesta seção os modelos baseados em sequência-a-função; genômica comparativa; vias metabólicas; análise baseada em rede; estrutura-a-função; similaridade estrutural; mineração de dados e combinação de diferentes abordagens.

Os modelos baseados sequência-a-função procuram aproveitar o fato de haver muito mais dados sobre sequências do que sobre estruturas tridimensionais. A função de uma proteína pode ser inferida a partir do alinhamento de sua sequência ou de parte dela (relacionada a algum sítio de ligação) com as sequências peptídicas identificadas em outras proteínas cuja função é conhecida. Assim, esta abordagem pode ser empregada para inferir sobre possíveis alvos que devem desempenhar papel fundamental (pró ou contra) no desenvolvimento de determinada doença. Como exemplo desta abordagem, pode-se citar o trabalho de Geyer et al. [2005] onde identificou-se quatro sequências "cyclin-like" sendo que três destas apresentam propriedades bioquímicas típicas que permitiram associá-las à "kinase activity" que, por sua vez, é considerada uma importante característica de atrativos alvos terapêuticos.

A comparação de genomas de organismos patogênicos com o genoma humano permite inferir sobre proteínas muito conservadas nos patógenos sem similares no ser humano que podem ser experimentadas como alvos terapêuticos com uma menor probabilidade de ocorrência de efeitos adversos [Kramer & Cohen, 2004]. Realizando um estudo de genômica comparativa de oito fungos (*C. albicans*, *A. fumigatus*, *Blastomyces dermatitidis*, *Paracoccidioides brasiliensis*, *Paracoccidioides lutzii*, *Coccidioides immi-*

tis, *Cryptococcus neoformans* e *Histoplasma capsulatum*), Abadio et al. [2011] identificaram dez genes presentes em todos eles e ausentes no ser humano e classificaram quatro desses genes como potenciais alvos terapêuticos.

Nos modelos baseados em vias metabólicas, o objetivo é encontrar proteínas-alvos que atuam em determinado metabolismo essencial para a sobrevivência de um dado parasita ou que seja relacionado com o desenvolvimento de determinada doença ou mal-estar [Cornish-Bowden & Cardenas, 2003]. A identificação de *chokepoints* em redes metabólicas específicas foi usada para sugerir potenciais alvos de Plamodium falciparum, causador da malária, [Huthmacher et al., 2010] e de Schistosoma mansoni, causador da esquistossomose, [Berriman et al., 2009].

Uma rede de dados permite detectar relacionamentos interessantes entre determinadas entidades (proteínas, famílias proteicas, fármacos, doenças etc.). O modelo baseado em mapas metabólicos pode ser considerado como um exemplo particular de um modelo de análise de rede. Os vértices representam proteínas e cada aresta representa algum tipo de interação entre as proteínas por ela conectadas. Esses modelos podem conter, por exemplo, proteínas e fármacos e suas respectivas interações. O STITCH [Kuhn et al., 2008] permite a análise de redes composto-composto, proteína-proteína e composto-proteína. Janga & Tzakos [2009] partiram de uma rede fármaco-alvo contendo interações conhecidas validadas experimentalmente e a decomporam em duas outras redes: uma fármaco-fármaco e outra alvo-alvo. Essas decomposições foram obtidas, respectivamente, interconectando-se os fármacos que compartilham um número significativo de alvos e interconectando-se alvos que compartilham um número significativo de fármacos. Na análise das redes de fármaco-fármaco e de alvo-alvo obtidas, os autores identificaram a tendência da indústria farmacêutica em explorar alvos validados experimentalmente caracterizando uma tendência de lançamento de "*follow-on drugs*". Os autores também apontaram *clusters* que correlacionam as similaridades encontradas com a classificação ATC (Anatomical Therapeutical Chemical). Zhao & Li [2010] desenvolveram um método baseado em rede que combina dados sobre a estrutura química e a indicação dos fármacos com a rede de interação proteína-proteína. A análise de rede pode ser empregada também na pesquisa sobre o desenvolvimento de fármacos "promíscuos" ou na formulação de coquetéis a serem aplicados no combate a doenças infecciosas onde se observa a evolução dos organismos para mutações resistentes às substâncias empregadas [Janga & Tzakos, 2009].

Em alguns modelos, a função das proteínas candidatas a alvo têm sua função predita a partir de sua similaridade com outras proteínas que constituem alvos validados ou com famílias proteicas especialmente destacadas como atrativos alvos terapêuticos (como as "cinases"). Como exemplo de aplicação desse método para a descoberta de

novos fármacos, pode-se citar o desenvolvimento do Captopril [Ondetti et al., 1977]. Como trabalhos mais recentes, pode-se citar as pesquisas de Darapaneni et al. [2009] e Heikkinen et al. [2008] que empregaram a modelagem por homologia na pesquisa sobre potenciais sítios alvos em proteínas não-estruturais do vírus influenza A. Vaidehi et al. [2002] utilizaram a modelagem por homologia para predizer a estrutura e a função das GPCRs (G protein-coupled receptors) uma vez que, no caso dessas proteínas havia apenas uma estrutura elucidada e somente para a espécie *bovine rhodopsin*. Os autores motivaram-se na pesquisa sobre essa família de proteínas por saber-se que elas mediam nossos sentidos de visão, olfato, tato e nas dores. O estudo estrutural das GPCRs continua sendo especialmente interessante para a pesquisa de alvos terapêuticos e ainda há, relativamente, poucas com estrutura resolvida. O *Potential Drug Target Database* (PDTD) [Gao et al., 2008; Li et al., 2006] é um banco de dados de potenciais alvos drogáveis determinados por experimentos *in silico* de *docking* reverso. Keiser et al. [2007, 2009] propõem um algoritmo, chamado *Similarity Ensemble Approach* (SEA), que associa fármacos a categorias de atividade biológica, levando em consideração a similaridade bidimensional da estrutura química de cada composto com um compêndio de ligantes. Luo et al. [2011] desenvolveram um método que combina propriedades estruturais de fármacos e alvos avaliando as energias de ligação calculadas por programas de *docking*. A principal desvantagem desta abordagem é que ela requer o conhecimento da estrutura tridimensional do proteína e isso não acontece para a maioria dos potenciais alvos terapêuticos.

Em outras abordagens, combina-se técnicas computacionais para "mineração textual" com curagens manuais realizadas por especialistas [Agarwal & Searls, 2008]. O *Manually Annotated Targets and Drugs Online Resource* (MATADOR) [Gunther et al., 2008] é uma base de dados de interações proteína-composto químico anotadas manualmente que inclui não apenas interações diretas entre o composto e a proteína, mas também interações indiretas fundamentadas em diferentes mecanismos - interação da proteína com um metabólito de um fármaco, alteração de uma dada expressão gênica etc. As interações são definidas a partir de mineração textual da literatura buscando termos das ontologias do MeSH, do OMIM e de outros nos *Abstracts* publicados no PubMed. Campillos et al. [2008] estudaram as similaridades entre os efeitos colaterais para associar fármacos bem estabelecidos a novos alvos.

Diferentes metodologias são combinadas a fim de construir-se uma abordagem mais abrangente ou mais eficiente. Pode-se por exemplo, combinar o modelo de análise de via metabólica com o modelo de identificação, por homologia, de proteínas ou domínios bem conservados para aumentar a credibilidade na identificação de bons candidatos a alvos. São selecionados aqueles que aparecem como "essenciais" tanto no

contexto de sobrevivência do organismo como no contexto das vias metabólicas conhecidas para o organismo. Gajria et al. [2008] constuíram um sistema contendo um banco de dados de "genoma e funções genômicas" do parasita *Toxoplasma gondii* e ferramentas de mineração de dados integradas para combinar diferentes tipos de informação nas consultas ao repositório. Perlman et al. [2011] avaliaram diferentes métricas para indicar a similaridade entre fármacos (usando ATC e o método SEA) e genes usando anotações do Gene Ontology (GO) e de bases de dados de interação proteína-proteína (PPI). Os autores avaliam duas possibilidades para combinar as métricas de similaridade entre fármacos S(d,d') com a métrica de similaridade entre alvos S(t,t'): uma baseada em média aritmética e outra baseada em média geométrica ponderada. A principal desvantagem do método é que a integração dos índices de similaridade é um problema não garantidamente convexo, demorado e caro computacionalmente.

Este trabalho está organizado em seis capítulos dos quais este é o primeiro. No que se segue, tem-se uma breve justificativa do trabalho. Os objetivos encontram-se no Capítulo 3. Na seção de Materiais e Métodos (Capítulo 4), está descrito como os dados foram extraídos das bases públicas assim como os elementos de álgebra linear necessários para construir o modelo vetorial proposto. Como não foram realizados experimentos de bancada, descreve-se também vários instrumentos para agrupar e visualizar os dados. A regressão logística, usada para caracterizar a importância dos atributos para a detecção dos alvos, é descrita com algum detalhe. No Capítulo 5, os resultados são apresentados junto de alguns exemplos de utilização da metodologia corroborados na literatura. Seguem-se as Conclusões (Capítulo 6), as Referências Bibliográficas e os Apêndices.

Capítulo 2

Justificativa

No desenvolvimento de novos medicamentos, o primeiro passo no paradigma atual, e que deve se manter ainda por alguns anos, é tentar identificar os alvos que podem ser modulados por um composto do tipo farmacológico [Harland & Gaulton, 2009]. Os investimentos nesse processo são elevados e aumentam a cada ano [DiMasi et al., 2003], devido, principalmente, ao declínio da eficiência na predição dos bons alvos terapêuticos [Projan, 2003; Scannell et al., 2012; Ruffolo, 2006].

Muitos esforços têm sido feitos na construção de bases de dados públicas contendo informações cuidadosamente curadas por especialistas [Zhu et al., 2010; Wishart et al., 2008; Kanehisa et al., 2010; Gunther et al., 2008]. A diversidade de fontes, escopos e tipos de dados, traz o desafio de integrar as informações e tentar encontrar novas associações. Há dados biológicos sobre alvos conhecidos e potenciais candidatos para que novas descobertas possam aflorar: dados sobre a estrutura química, sobre a bioatividade e sobre o mecanismo de ação de fármacos conhecidos e de outros produtos (compostos naturais, drogas ilícitas etc.); dados sobre vias metabólicas e outros processos associados a doenças ou condições que se deseja tratar ou prevenir. Seja em pesquisas que tratem cada um desses contextos separadamente, seja em pesquisas que procurem combiná-los, faz-se necessário o uso de técnicas computacionais capazes de descobrir e extrair conhecimento de uma dada coletânea de dados – as chamadas técnicas de recuperação de informação.

A simples integração dos dados em uma base única pode não ser suficiente para encontrar novas associações. O fundamental não é construir apenas um banco de dados relacional, mas sim um sistema onde se possa imbuir inteligência ou capacidade de raciocínio [Zygmunt, 2010]. A forma usual para abordar esse desafio é procurar organizar os dados semanticamente, seja pelo emprego de alguma ontologia, seja pelo desenvolvimento de uma métrica de similaridade "semântica". De modo geral, diz-se

que "dois objetos são semanticamente similares se eles se relacionam a objetos similares" [Jeh & Widom, 2002]. A partir de uma consulta direta em uma base de dados relacional, dificilmente emergeria objetos que não compartilham nenhum termo entre si, mas que podem ser considerados associados quando são observadas relações desses objetos com outros intermediários na massa de dados.

Esse tipo de pesquisa em uma massa de dados de alvos drogáveis permite ainda encontrar eventuais correlações entre alvos que podem sugerir possibilidades para terapias adjuntivas ou prever efeitos adversos.

Capítulo 3

Objetivos

3.1 Objetivo geral

Desenvolver uma metodologia para análise e predição de alvos proteicos drogáveis a partir de dados de anotação de proteínas reconhecidas como drogáveis. Além de indicar potenciais alvos, essa metodologia deverá identificar similaridades e/ou correlações entre alvos dificilmente detectadas por alinhamento de sequências ou pela busca direta em bases de dados.

3.2 Objetivos específicos

- Representar proteínas alvo como vetores formados por conjuntos de descritores binários relativos a anotações biológicas.

- Representar fármacos como vetores formados por conjuntos de descritores binários relativos às propriedades estruturais químicas do composto, às classes terapêuticas do fármaco, às anotações biológicas relativas aos seus alvos associados ou a combinações desses diferentes tipos de descritores.

- Desenvolver uma métrica de similaridade semântica entre alvos ou entre fármacos aplicando a decomposição por valores singulares.

- Analisar a representatividade química e/ou biológica dos modelos vetoriais por estudos de correlação com outras métricas consagradas e de uso comum (coeficiente de Tanimoto, alinhamento sequencial usando BLAST e outras métricas recentemente publicadas e que têm recebido destaque).

- Fornecer listas de potenciais candidatos a alvos drogáveis ordenados segundo o coeficiente de similaridade semântica.
- Fornecer listas, ordenadas segundo o critério de similaridade semântica, de pares alvo-alvo, fármaco-fármaco bem como contendo mapeamentos entre fármaco e indicação ou classe de atividade farmacológica quando possível.
- Avaliar diferentes técnicas de análise e visualização (técnicas de agrupamento e optimização) dos dados reorganizados semanticamente.
- Fornecer um modelo para construção de grupo controle (diante da inexistência de um) a ser usado em estudos estatísticos de caso-controle.
- Fornecer um modelo probabilístico para predição de potenciais alvos drogáveis utilizando regressões univariada e multivariada.
- Discutir alguns casos de estudo aplicando-se as diferentes técnicas de análise empregadas.

Capítulo 4

Materiais e Métodos

A seguir, com o propósito de garantir e facilitar a reprodutibilidade, arrolamos as atividades relacionadas à coleta, seleção e integração dos dados, ao lado dos procedimentos para validação da metodologia.

- Coleta de informações farmacológicas e biológicas do DrugBank, TTD e KEGG.
- Análise da qualidade dos dados coletados.
- Identificação dos campos de interesse.
- Construção de *parsers* para extrair os dados e associações desejados.
- Integração das informações em um banco local.
- Obtenção dos dados do InterPro.
- Análise da base de dados do InterPro, seleção das informações desejadas e integração destas ao banco local.
- Construção de uma rotina para baixar do UniProt os *flat files* relativos aos alvos conhecidos.
- Obtenção dos *flat files* do UniProt de outras sequências que compartilham alguma assinatura do InterPro com o conjunto de alvos.
- Estudo dos dados do UniProt e seleção dos campos de interesse.
- Construção de um *script* para extrair e integrar os dados do UniProt.
- Escrita de *queries* SQL explorando os diferentes tipos de informação contidos na base de dados.

- Representação de proteínas alvo como vetores formados por conjuntos de descritores binários relativos às suas anotações biológicas.
- Construção de um modelo vetorial para os alvos proteicos humanos aplicando a decomposição por valores singulares e redução de posto.
- Desenvolvimento de uma métrica de similaridade entre alvos.
- Demonstração do ganho de informação devido à redução de posto.
- Pesquisa na literatura e seleção de publicações que corroborem novas associações extraídas do modelo proposto.
- Análise qualitativa da representatividade biológica e farmacológica do modelo vetorial.
- Estudo quantitativo de correlação com o alinhamento de sequências.
- Extensão das comparações quantitativas por análise de curvas ROC.
- Discussão de alguns casos de estudo aplicando-se as diferentes técnicas de análise e visualização empregadas.
- Projeção de sequências proteicas de *Plasmodium falciparum* e de *Toxoplasma gondii* no espaço vetorial construído com alvos humanos para avaliar potenciais oportunidades de reposicionamento de fármacos para o combate desses parasitas.
- Escolha e discussão de estudos de caso associados às doenças tropicais negligenciadas.
- Proposta de um método para construção de grupo controle (diante da inexistência de um) a ser usado em estudos estatísticos de caso-controle.
- Desenvolvimento de um modelo probabilístico para predição de potenciais alvos drogáveis utilizando regressões univariada e multivariada.
- Análise do conjunto de assinaturas do InterPro obtidos pelo modelo de seleção de atributos.
- Comparação dos *motifs* relevantes propostos pelo modelo probabilístico com aqueles propostos por Hopkins & Groom [2002].
- Construção e comparação de modelos vetoriais para os medicamentos utilizando diferentes tipos de descritores.

- Comparação dos resultados obtidos com os diferentes modelos vetoriais representativos dos fármacos.

- Disponibilização das rotinas, da base de dados e das listas produzidas durante todo o trabalho.

4.1 Coleta, armazenamento e integração de dados

Todos os dados utilizados neste trabalho estão disponíveis em repositórios públicos. Eles foram extraídos do TTD [Zhu et al., 2010; Chen et al., 2002], do *DrugBank* [Wishart et al., 2008], do *KEGG-DRUG* [Kanehisa et al., 2010], do *UniProtKB* [Consortium, 2010], do *InterPro* [Hunter et al., 2012], do *Database of Essencial Genes* (DEG), do UniProt Gene Ontology Annotation (GOA) [Barrell et al., 2009] e do PubMed [Roberts, 2001].

A coleta, seleção e integração desses dados foi a etapa mais difícil deste trabalho. Descrevemos detalhadamente os procedimentos para facilitar a sua reprodutibilidade.

Os alvos drogáveis conhecidos foram extraídos do TTD, do DrugBank e do KEGG-DRUG. Esta relação inclui alvos terapêuticos e outros sabe-se que são modulados pela ação de algum medicamento mas que não estejam associados a alguma terapia. Para este trabalho, foram filtrados somente os alvos proteicos. O relacionamento de cada alvo drogável proteico com a base de dados do UniProtKB, ou foi obtido diretamente das bases de onde foram retirados, ou foi obtido por mapeamento secundário, em um cuidadoso processo envolvendo identificadores e nomes de genes e proteínas. A partir dos identificadores uniprot_AC dos alvos drogáveis, as respectivas anotações depositadas no InterPro foram obtidas. Os termos do InterPro associados ao conjunto de alvos drogáveis conhecidos foram usados como consulta à base do InterPro para obter a lista de todas as outras proteínas que compartilham alguma anotação com algum alvo conhecido. Dados adicionais (nome recomendado e sinônimos, gene associado, sequência da estrutura primária etc.) sobre as proteínas selecionadas foram obtidos do UniProtKB. Por fim, informações complementares foram obtidas de outras bases públicas. Do PubMed foram extraídos título, resumo e autores das publicações relacionadas em alguma das outras bases consultadas; do DEG foram extraídos a relação de genes considerados essenciais (informação que foi cruzada com um dos dados do DrugBank). Relacionamentos com o GO foram obtidos a partir do GOA (Figura 4.1).

Com relação aos dados referentes ao InterPro, foram extraídos o InterPro ID, o nome e o tipo da anotação, o intervalo da sequência correspondente à anotação e

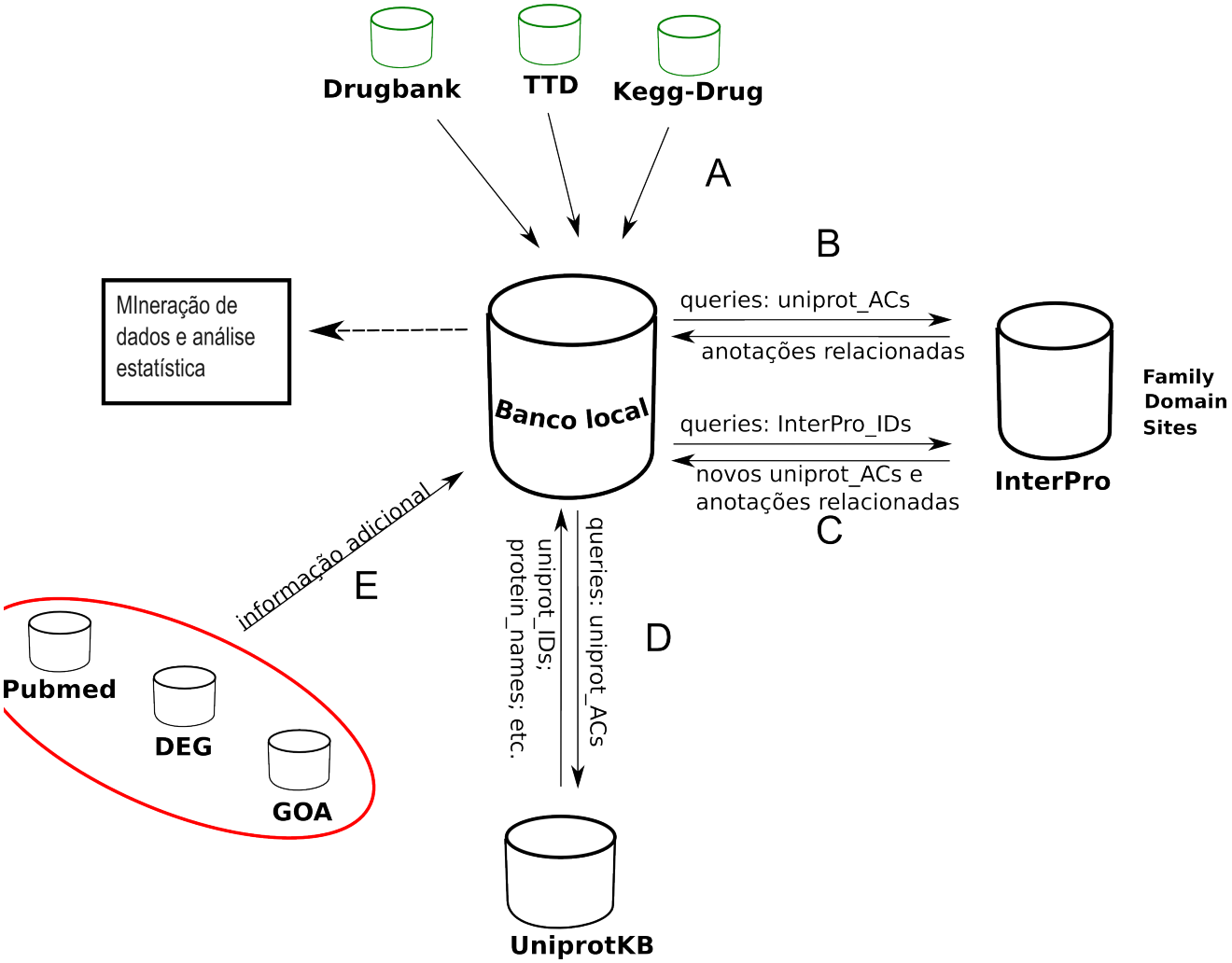


Figura 4.1. Coleta dos dados em bases públicas e armazenamento no banco de dados local. Inicialmente, foram extraídos do TTD, DrugBank e Kegg, os dados sobre fármacos e suas conhecidas associações com alvos proteicos (terapêuticas ou não) (**A**). Os identificadores uniprot_AC dos alvos proteicos foram usados em consultas aos dados do InterPro, obtendo-se as respectivas anotações (**B**). Os identificadores InterPro_ID obtidos foram usados em nova consulta aos dados do InterPro obtendo-se novas proteínas identificadas pelos respectivos uniprot_AC (**C**). Dados complementares sobre as proteínas selecionadas (uniprot_ID, sequência de resíduos de aminoácidos etc.) foram extraídos do UniProtKB (**D**). Informações adicionais foram extraídas de outras bases de dados públicas (**E**).

o método empregado para definir a anotação. Quanto aos tipos e anotação, foram selecionadas aquelas dos tipos família (F), domínio (D), região (G), sítio de ligação (B) e sítio ativo (A).

Sobre os fármacos, os principais dados extraídos foram: os respectivos identificadores no TTD, DrugBank e KEGG-DRUG (sempre que possível), os SMILES (*Simpli-*

fied Molecular Input Line Entry System) relativos à estrutura química, os alvos proteicos associados, o mecanismo de ação do fármaco sobre o alvo (quando conhecido), a indicação ou classe terapêutica e o estado do fármaco ou da associação fármaco-alvo enquanto aprovado ou experimental.

Para facilitar o gerenciamento e a integração dos dados, eles foram armazenados em uma base de dados relacional usando o MySQL como sistema gerenciador de banco de dados (SGBD).

O MySQL foi, inicialmente, o SGBD escolhido por apresentar os recursos necessários além de ser prático e possuir vasta documentação. Apesar de ser um SGBD de código aberto, o MySQL é de propriedade particular. Por isso a base de dados foi exportada e carregada em outro SGBD de código aberto, o PostgreSQL, este sim classificado como *software livre*, para garantir a reprodutibilidade do armazenamento dos dados a partir do arquivo de *dump*.

4.2 Bancos de dados de públicos

4.2.1 DrugBank

O DrugBank tem o foco principal nos fármacos aprovados pela *US FDA* ou em experimentação nos EUA [Wishart et al., 2008, 2006]. Além de informações sobre os fármacos, fornece dados sobre seus respectivos alvos. Quando os alvos são proteicos, o DrugBank fornece o correspondente código *AC number* do UniProtKB.

É oferecida a possibilidade de busca por nomes (e sinônimos) de fármacos, alvos e doenças. Também podem ser feitas a buscas a partir de uma lista de classes terapêuticas, palavras e textos, por similaridade das sequências de aminoácidos ou de nucleotídeos dos alvos. Quanto às propriedades químicas do composto, oferece recurso para busca por similaridade de *fingerprint* baseada na métrica de Tanimoto. É possível desenhar a molécula desejada usando um *applet* java ou fornecer o SMILE correspondente.

Operadores *booleanos* simples (AND, OR e NOT), além de aspas, menos "-" e mais "+" são permitidos, mas apenas na busca por texto. Não permite a combinação de diferentes parâmetros de busca além do que é permitido pelo uso de operadores na busca textual.

Os modos de visualização fornecidos são HTML e CSV. No formato HTML, exibe primeiramente uma página com resumo dos registros encontrados. Cada gene essencial aparece relacionado em uma tabela e é um *link* para uma página específica com maiores informações.

É possível visualizar a estrutura da molécula diretamente na página referente ao composto por meio de figura bidimensional. Há também um recurso de visualização da estrutura tridimensional a partir do clique em um botão que carrega um *applet* java.

Dentre outras informações sobre o alvo, há um parâmetro que indica a essencialidade do mesmo para a sobrevivência do organismo. Há três valores possíveis para este campo: "Essential", "Non-Essential" e "Not available".

Quanto a outras fontes, além do link para o UniProt e para o PDB (quando houver), oferece links para o KEGG-Pathway, KEGG-Drug e KEGG-Compound.

Em versões anteriores (inclusive a versão usada inicialmente neste trabalho), o DrugBank utilizava identificadores com formato distinto para fármacos aprovados (APRD00001) e fármacos experimentais (EXPT00001). Utilizar identificadores que carregam em si uma característica sobre a entidade que representam – no caso, se o fármaco é experimental ou aprovado – leva invariavelmente à necessidade de alterações dos identificadores na medida em que a entidade tem sua característica alterada. Quando um fármaco passa do estado experimental para aprovado, era necessário alterar seu identificador para manter uma coerência com o propósito de se iniciar os identificadores com APRD ou EXPT. Na versão atual, utiliza-se identificadores no formato (DB00001) e a informação sobre seu estado enquanto aprovado ou experimental é fornecida a partir de um campo próprio.

Quanto às referências da literatura, o DrugBank as agrupa explicitando quais referem-se ao fármaco, quais fornecem informações gerais sobre o alvo e quais discutem a evidência da relação entre o fármaco e o alvo.

Na seção de *downloads*, disponibiliza os dados em formato ASCII (*flat-file*) onde o início de cada registro é marcado por uma linha do tipo:
#BEGIN_DRUGCARD DB00001
e seu fim é marcado por uma linha do tipo:
#END_DRUGCARD DB00001

Cada campo aparece com seu nome identificador em outras linhas iniciadas por "#" e seu conteúdo aparece nas linhas não vazias sub-sequentes.

Além dos arquivos de texto (*flat-files*), a seção de *downloads* fornece arquivos nos formatos MOL e SDF relativos a cada fármaco contido no DrugBank.

Os dados do DrugBank foram extraídos a partir do arquivo *drugcards.txt*. Foi desenvolvido um *parser* na linguagem *perl* para extrair os dados desejados.

4.2.2 *Therapeutical Target Database* (TTD)

O TTD, é um banco de dados de proteínas e RNA reconhecidos como alvos terapêuticos a partir de uma inspeção manual da literatura. O projeto é mantido pelo *Bioinformatics and Drug Design Group*, sediado no Departamento de Ciência da Computação da Universidade Nacional da Singapura [Chen et al., 2002; Zheng et al., 2006].

A interface não oferece flexibilidade para buscas mais complexas. Mas os dados estão disponíveis em um arquivo no formato CSV (Figura 4.2). Também são disponibilizados arquivos MOL e SDF dos fármacos e arquivos contendo as sequências dos alvos em formato FASTA. Algumas discrepâncias foram encontradas e resolvidas. O identificador *AC Number* do UniProtKB é fornecido nos casos em que o alvo é uma proteína. (Esse importante item facilitador, e próprio arquivo para *download*, não estavam disponíveis no início deste projeto.) Mesmo com o arquivo, a análise foi difícil em função de seu formato, de valores ausentes e de erros tipográficos. Por exemplo, o primeiro alvo da lista é uma sequência depositada no Swiss-Prot, mas a o arquivo não fornece o UniProt ID. O primeiro código em cada linha designa o alvo, mas em seguida não há padrão muito bem definido. Por exemplo, linhas que indicam as doenças associadas ao alvo são denotadas pela palavra-chave "Disease" na segunda coluna, mas as linhas que designam o mecanismo de ação são identificadas pelo próprio mecanismo em si (Antagonist, Agonist, Binder etc.).

Os parâmetros possíveis para a pesquisa no banco são:

- Target Name: campo textual sem possibilidade de uso de operadores lógicos;
- Type of the Target: botões de rádio com as opções All, Successful, Clinical Trial e Research;
- Drug Name: campo textual também sem permitir o uso de operadores lógicos;
- Type of the Drug: botões de rádio com as opções All, Approved e Clinical Trial;
- Disease Indication: caixa de seleção sem possibilidade de seleções múltiplas;
- Target BioChemical Class: como no caso do campo "Disease Indication";
- Drug Mode of Action: como no caso do campo "Disease Indication" e
- Drug Therapeutic Class: como no caso do campo "Disease Indication".

Nenhum campo é obrigatório, mas a simples consulta onde passa-se apenas os valores "All" tanto para o campo "Type of Target" como também para o campo "Type

```
TTDS00001  Name    Muscarinic acetylcholine receptor
TTDS00001  Type of target  Successful target
TTDS00001  Synonyms   (m)AChR
TTDS00001  Synonyms   MAChR
TTDS00001  Disease Alzheimer's disease
TTDS00001  Disease Bronchospasm (histamine induced)
TTDS00001  Disease Glaucoma
TTDS00001  Disease Motion sickness
...
TTDS00001  BioChemical Class  G-protein coupled receptor (rhodopsin family)
TTDS00001  Pathway Calcium signaling pathway
TTDS00001  Pathway Neuroactive ligand-receptor interaction
TTDS00001  Pathway Regulation of actin cytoskeleton
TTDS00001  Related US Patent  6211204
TTDS00001  Related US Patent  6323194
...
TTDS00001  Drug(s) Bethanechol     DAP000263  Urinary retention  Approved
TTDS00001  Drug(s) Trospium   DAP000342  Urge urinary incontinence  Approved
...
TTDS00001  Drug(s) Aclidinium      DCL000677  Chronic obstructive pulmonary disease   Phase III
TTDS00001  Drug(s) CHF 5407   DCL000750  Chronic obstructive pulmonary disease   Phase I
...
TTDS00001  Antagonist      Trospium   DAP000342
...
TTDS00001  Antagonist      Aclidinium      DCL000677
TTDS00001  Antagonist      CHF 5407   DCL000750
...
TTDS00002  UniProt ID      P11229
TTDS00002  Name    Muscarinic acetylcholine receptor M1
TTDS00002  Type of target  Successful target
TTDS00002  Synonyms   M1 receptor
TTDS00002  Disease Alzheimer's disease
TTDS00002  Disease Bronchospasm (histamine induced)
...
TTDS00002  Disease Cognitive deficits
TTDS00002  Disease Schizophrenia
TTDS00002  Function   The muscarinic acetylcholine receptor mediates various cellular responses,
TTDS00002  Sequence   MNTSAPPAVSPNITVLAPGKGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTELKTVNNYFLLSLACADLII
TTDS00002  BioChemical Class  G-protein coupled receptor (rhodopsin family)
...
...
TDS00012  Drug(s)  (S)-amisulpride  DPR000002  Schizophrenia  Discontinued
TTDS00012  Drug(s)  1192U90  DPR000003  Schizophrenia  Discontinued
TTDS00012  Drug(s)  SDZ-HDC-912  DPR000105  Schizophrenia  Discontinued
TTDS00013  Drug(s)  RGH-1756  DPR000095  Schizophrenia  Discontinued
TTDS00014  Drug(s)  Belaperidone  DPR000019  Schizophrenia  Discontinued
TTDS00014  Drug(s)  Sonepiprazole  DPR000110  Schizophrenia  Discontinued
TTDS00059  Drug(s)  FK778  DPR000045  Heart transplantation  Discontinued
...
...
TTDS00393  Drug(s)  Ramelteon  DAP000070  Circadian rhythm sleep disorder(CRSD)  Discontinued
TTDS00422  Drug(s)  Ramelteon  DAP000070  Circadian rhythm sleep disorder(CRSD)  Discontinued
```

Figura 4.2. Arquivo de dados do TTD. O próprio formato do código indica o tipo do alvo enquanto *Successful* (TTDSxxxxx), *Clinical trial* (TTDCxxxxx) ou *Pre-clinical trial* (TTDRxxxxx). O código TDS00012 (faltando um T) que aparece na figura é de um erro encontrado no arquivo original. O identificador do UniProt não é fornecido para todos os alvos proteicos (como no caso do primeiro da lista). As linhas que exibem a função e a sequência do *muscarinic acetylcholine receptor* estão truncadas à direita por questão de espaço.

of Drug", que supostamente deveria resultar em todos os alvos contidos na base de

dados, resulta em "Sorry, Nothing is found". Todos os campos atuam na busca como filtros positivos. Não há o equivalente ao operador lógico *NOT*.

A interface permite fazer buscas por similaridade de sequências da proteína usando BLAST, onde uma única proteína deve ser passada por vez em formato FASTA. Também é possível realizar buscas por similaridade estrutural do composto sintetizado baseada na métrica de Tanimoto. O composto deve ser fornecido submetendo-se um arquivo em formato MOL ou SDF que contenha a estrutura de um único composto.

A exibição dos resultados é no formato HTML com os dados referentes a cada alvo são exibidos em uma grande tabela. Quando uma consulta resulta em mais de um alvo, uma página intermediária é exibida contendo o indicador do alvo, seu nome e algumas informações resumidas. As referencias que corroboram a identificação do alvo aparecem no final da tabela.

A partir de sua nova versão, o TTD formalizou o uso de seus identificadores próprios para alvos e fármacos. Na versão anterior, nenhum identificador era citado. Mas o padrão adotado apresenta o mesmo defeito do padrão inicialmente adotado pelo DrugBank. Apresenta uma característica desinteressante para que o identificador possa ser estável e tornar-se universal. Tanto os identificadores usados para alvos como aqueles usados para fármacos carregam em si uma propriedade sobre um estado atual do objeto que eles representam. No caso dos alvos, aqueles classificados como "*Successful*" recebem um identificador que inicia com TTDS, os que são identificados como alvos em processo de experimentação na fase clínica são designados por identificadores iniciados por TTDC e aqueles identificados em fase de pesquisa pré-clinica são designados por identificadores iniciados por TTDR. Algo semelhante ocorre para os fármacos cujos identificadores iniciam-se por DAP, DCL ou DPR conforme o fármaco seja aprovado ou encontre-se nas fases clínica, pré-clinica (testes com animais) ou "experimental" (testes *in vitro*) com os primeiros alvos. Se um alvo classificado como experimental passa a ser reclassificado como "alvo de sucesso", é necessário alterar seu código e isso pode eventualmente invalidar *links* em outras bases. Além disso, um alvo pode ser classificado como caso de sucesso para determinada terapia, mas encontrar-se em fase experimental para outro medicamento ou terapia. Complicações semelhantes podem ocorrer no caso dos medicamentos que são reclassificados.

O mais atraente no TTD é que ele é focado no alvo, ao passo que o DrugBank e o KEGG-DRUG são focados no fármaco. Todos são repositórios de dados ricos com informações obtidas pela curagem manual da literatura científica. Mas nenhum desses ambientes oferece facilidade para selecionar determinado conjunto de alvos, fármacos ou referências para proceder outros procedimentos computacionais.

4.2.3 *Kyoto Encyclopedia of Genes and Genomes* (KEGG)

O KEGG é um conjunto de bases de dados de sistemas biológicos que integra informações sobre genes, compostos químicos, fármacos, reações, vias metabólicas, ontologias relacionadas a doenças e outras informações. O cerne dos relacionamentos está centrado no contexto de vias (*pathways*) que tanto podem representar sistemas de interações físicas como ontologias específicas.

Dentre as bases de dados contidas no KEGG, destacamos:

- KEGG PATHWAY (vias metabólicas),
- KEGG BRITE (Hierarquias funcionais),
- KEGG DISEASE (Ontologias relacionadas a doenças),
- KEGG ORTHOLOGY (grupos ortólogos segundo o contexto do KEGG),
- KEGG DRUG (fármacos),
- KEGG COMPOUND (metabólitos e outros compostos químicos),
- KEGG ENZYME (enzimas) e
- KEGG REACTION (reações enzimáticas).

Para este trabalho, foram extraídos dados do KEGG-Drug a partir do *flat-file* disponível para *download* no sítio oficial do projeto. Foi avaliada a possibilidade de utilizar um *parser* desenvolvido no projeto *BioWarehouse*. Entretanto, ocorreram vários problemas que não permitiram o funcionamento do *script*. Assim, construímos nosso próprio *perl script* para selecionar e extrair os dados desejados.

Os dados selecionados foram:

- o identificador interno para o fármaco;
- nomes e sinônimos do fármaco para fins de identificação com outras fontes;
- nome do alvo;
- mecanismo de ação do fármaco sobre o alvo e
- o alvo propriamente dito quando designado um gene ou um grupo ortólogo.

```
TARGET  nome_do_alvo mecanismo_de_ação [ID];
        nome_do_alvo mecanismo_de_ação [ID1] [ID2];
        nome_do_alvo mecanismo_de_ação [ID]
```

No arquivo disponibilizado no KEGG-Drug, os dados relativos ao alvo aparecem contidos em um único campo identificado como TARGET e que apresenta o formato:

O KEGG-Drug aponta "alvos" no contexto do *pathways* direcionando a análise da ação do fármaco como perturbações sobre sistemas moleculares. Esses sistemas podem estar estruturados em diferentes categorias e sub-categorias, como, por exemplo:

- metabolismo (metabolismo do piruvato, glicólise etc.);
- processos celulares (transporte e catabolismo, crescimento celular etc.) e
- doenças humanas (cânceres, distúrbios imunológicos, distúrbios metabólicos etc.).

Dentro dessa variedade de categorias e sub-categorias, os identificadores que aparecem no campo TARGET apresentam-se em formatos bem variados podendo indicar um gene (e.g. [HSA:185]), um *pathway* em algum contexto do KEGG-PATHWAY, baseado em interações e reações moleculares (e.g. [PATH:hsa04080(185)]) ou em algum contexto do KEGG-BRITE que pode incluir outros tipos de associações (e.g. [BR:hsa04000("Group A: estrogen")]). O identificador pode ainda ser um *EC number* indicando uma enzima ou uma classe enzimática (e.g. [EC:3.4.23.15]).

O contexto de *pathways* do KEGG pode auxiliar no processo de expandir a análise das interações entre fármaco-alvo considerando-se uma categorização de alvos mais complexos (e realísticos) ao focalizar não apenas nas proteínas em si e dando pistas que podem auxiliar na construção de uma "fármaco-ontologia" – um ponto importante segundo Harland & Gaulton [2009]. Entretanto, esta mesma estrutura dificulta o processo de combinação de informação contida em outras fontes de associações fármaco-alvo (e.g. DrugBank e TTD).

Uma outra dificuldade ocorre em alguns casos, em que não há nenhum código identificador no campo TARGET como, por exemplo:

```
TARGET  alpha-chymotrypsin inhibitor;
        elastase inhibitor;
        trypsin inhibitor
```

Estes casos requerem o uso de comparação textual de nomes e sinônimos de proteínas ou genes – uma tarefa nada trivial, como apontado por Harland & Gaulton [2009].

Introduzimos em nossa base de dados as associações entre fármaco e alvo associando o identificador do fármaco no KEGG com o UniProt obtido através do mapeamento do gene humano (identificado por HSA:xxx) ou do grupo ortólogo (KO:xxx). Ao armazenar essas associações na base de dados, tomamos o cuidado de manter a informação sobre a forma em que foi obtida (via associação direta com um dado gene ou via associação com grupo ortólogo e levando, a partir daí, a genes humanos). Os mapeamentos gene-uniprot e ko-gene foram obtidos diretamente do servidor FTP do KEGG em arquivos próprios para isso.

4.2.4 *UniProt Knowledgebase* (UniProtKB)

O UniProtKB é um banco de dados público de sequências proteicas composta por duas seções: o Swiss-Prot que é a parte do UniProtKB que é anotada e curada manualmente por especialistas e o TrEMBL que é constituído por outras sequências anotadas automaticamente e ainda não revisadas. Quando uma sequência do TrEMBL é revisada e aprovada, ela é assinalada como revisada e passa a fazer parte do Swiss-Prot.

A cada registro no UniProtKB é assinalado um identificador único denominado *ID*. No Swiss-Prot, esse identificador tem o formato **X_Y**, onde:

- **X** é um código mnemônico de, no máximo, cinco caracteres alfanuméricos, que geralmente indica o gene que codifica a sequência em questão;
- O símbolo '_' serve como separador;
- **Y** é um código mnemônico de no máximo, cinco caracteres alfanuméricos, que indica a espécie. Em geral, este código é formado pelas três primeiras letras do gênero e as duas primeiras letras da espécie. Para as espécies que mais aparecem no banco de dados, este código tem uma formato definido arbitrariamente, mas que facilita a memorização (como HUMAN para humano, BOVIN para bovino, CHICK para *Gallus gallus*, ECOLI para *Escherichia coli*, PIG para porco etc).

No TrEMBL, o formato do identificador único é bem parecido com o do Swiss-Prot. Também tem o formato **X_Y**, onde:

- **X** é idêntico ao chamado *accession number*, um código composto por seis caracteres alfanuméricos, um a mais que no caso do Swiss-Prot;
- O símbolo '_' serve como separador;

- **Y** é um código mnemônico de no máximo, cinco caracteres alfanuméricos, que indica a espécie. Em alguns casos, por falta de tempo de se atribuir um código para cada espécies no TrEMBL, um código dito "virtual" é usado. Esses códigos virtuais começam com o dígito "9" e, usualmente indicam um conjunto de organismos.

Além do identificador *ID*, cada registro tem um outro identificador denominado *accession number* (AC) que é usado para garantir a estabilidade dos registros frente às constantes atualizações. Por questão de consistência, pode ser necessário alterar o nome de um registro, ou até mesmo seu ID. Um *accession number* é sempre conservado. O que se faz é gerar outro AC para a sequência, mas todos os antigos são mantidos. Por isso, é mais comum usar o AC do que o ID em citações, ainda que aquele não seja único para uma dada sequência.

O *UniProt consortium* mantém versões antigas da base. Registros que tornaram-se "depreciados" ou que foram mesclados com outros permanecem acessíveis. Mas os arquivos dos registros trazem apenas os AC numbers antigos. Se um ID tiver sido alterado, o antigo não aparece no arquivo, ainda que a consulta no site do projeto redirecione corretamente para o novo ID.

Os dados do UniProKB podem ser obtidos no servidor FTP do projeto ou via *link* HTML. Pode-se baixar um arquivo único contendo todos os registros da base de dados, ou baixar cada registro em um arquivo separado. Para realizar este trabalho, os registros foram baixados em arquivos separados, já que não era necessário obter todos os registros da base. Foi utilizado uma linha de comando de *bash* (o interpretador de comandos padrão do *Linux*) contendo uma estrutura de repetição que lê um arquivo com todos os IDs desejados e usa o programa *wget* para baixar os correspondentes arquivos do UniProtKB:

```
for uid in ` cat uniprotIDs `;
  do wget http://uniprot.org/uniprot/$uid.txt –O files/$uid;
done
```

A Figura 4.3 ilustra um exemplo de arquivo do UniProtKB e como alguns dos principais campos para este trabalho foram extraídos. Cada linha do arquivo inicia com um código de dois caracteres que indica o tipo de conteúdo contido na linha. Somente as linhas com a sequência de resíduos de aminoácidos são iniciadas com o caracter de espaço. O início de cada registro é denotado pelo código ID e o final é denotado por duas barras "//". Isso vale tanto para a versão em um único arquivo como para a versão com os dados distribuídos em vários arquivos. A primeira linha contém ainda a

situação do registro (revisado ou não-revisado) e o tamanho da sequência em número de resíduos. que também está disponível na linha iniciada pelo código SQ no final do arquivo. A linha iniciada por AC contém todos os códigos *accession numbers* já atribuídos à sequencia em questão. Quando há muitos códigos desse tipo, ocasionado por diversas atualizações do registro, pode haver mais de uma linha iniciada por AC. A ordem dos códigos AC é a inversa da ordem cronológica. O primeiro código é o AC atual. As linhas iniciadas por DR contêm dados de referência cruzada com outras bases de dados (e.g. DrugBank, InterPro e outras). Neste trabalho, esses dados foram extraídos diretamente das fontes de interesse de modo a garantir maior completude e consistência das informações.

Foi desenvolvido um programa em linguagem *perl* que produz um arquivo SQL para ser usado com as ferramentas de importação (e.g. *mysqlimport*) de um SGBD.

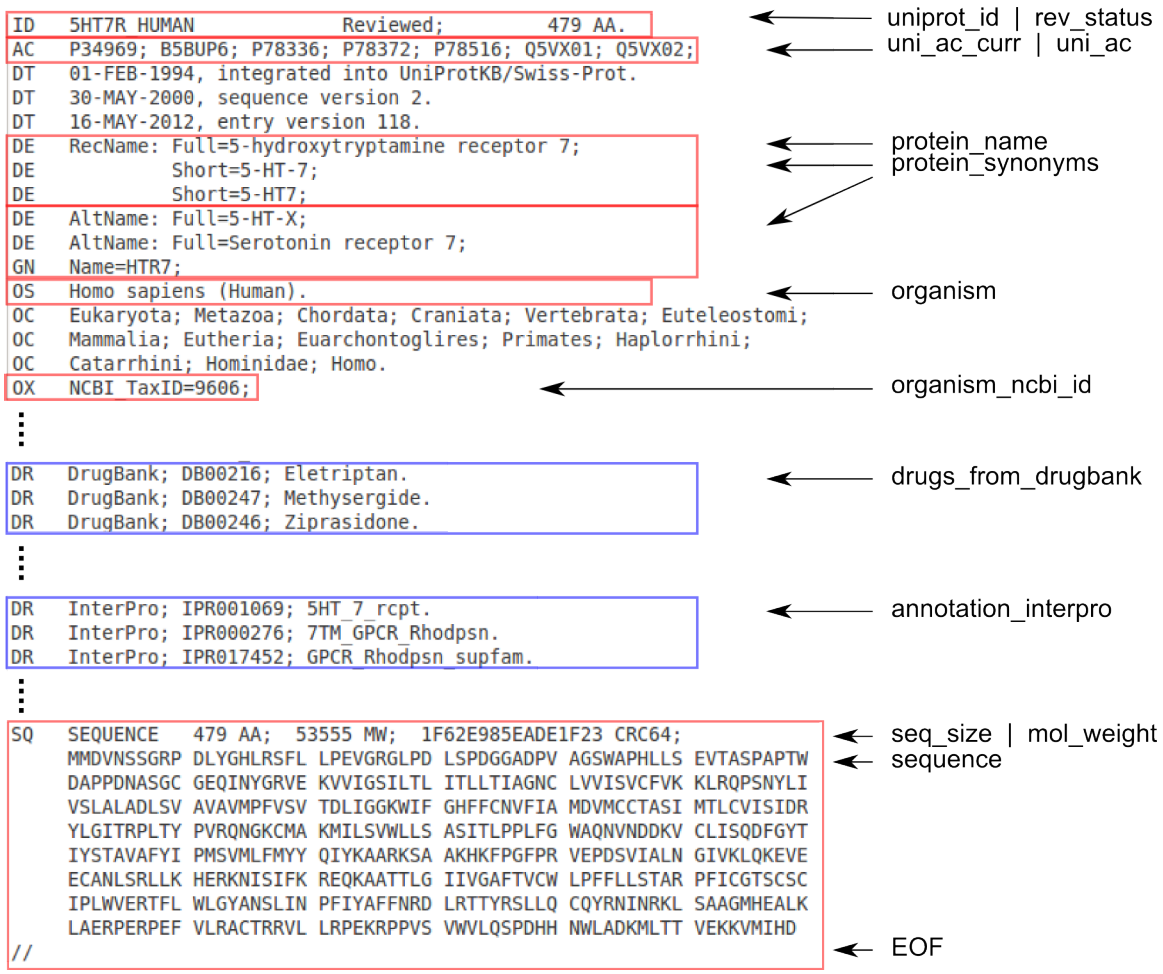


Figura 4.3. Exemplo de arquivo do UniProtKB. À direita, a indicação de alguns dos principais campos extraídos. As linhas iniciadas por DR fornecem referências cruzadas com outras fontes, que são dados que foram extraídos diretamente das fontes externas.

4.2.5 InterPro

O InterPro é um meta banco de dados que agrega informações dos seguintes repositórios públicos cujas anotações foram obtidas por diferentes metodologias e diferentes escopos:

- ProDom: contém sequências do UniProtKB agrupadas utilizando-se PSI-BLAST;
- PROSITE patterns: padrões nas sequências representados por expressões regulares;
- PROSITE and HAMAP profiles: perfis de sequências representados por matrizes;
- PRINTS: *fingerprints* que são grupos alinhados de Matrizes de Sequência de Posição Específica não-ponderada (PSSMs);
- PANTHER, PIRSF, Pfam, SMART, TIGRFAMs, Gene3D e SUPERFAMILY, que são baseados em modelos ocultos de Markov (HMMs).

Uma anotação associa uma "assinatura" a uma região da sequência proteica. Essa assinatura refere-se a algum termo que, por sua vez, denota uma família, domínio ou sítio funcional. Esse termo é designado por um código, denominado *InterPro ID* (ou simplesmente InterPro) que tem o formato IPRxxxxxx onde cada x é um dígito decimal. A região referente à anotação é designada citando-se as posições inicial e final na sequência.

Quando diferentes assinaturas ocorrem para um mesmo conjunto de proteínas em uma mesma região da sequência, com sobreposição maior do que 75%, presume-se que todas elas descrevem uma mesma família, domínio ou sítio funcional. Neste caso, um curador atribui um único termo do InterPro a essas assinaturas.

No caso de uma assinatura conferir com apenas um subconjunto de proteínas em comparação com outra assinatura, presume-se que aquela assinatura seja funcional ou taxonomicamente mais específica do que esta. Neste caso, considera-se a assinatura mais específica como "filha" da outra, resultando uma estrutura hierárquica para o InterPro. Esta estrutura também é definida criteriosamente pelos curadores.

Um termo do InterPro pode ser do tipo Família (F), Domínio (D), Região (G), Repetição (R) ou de Sítio (B, A, C ou P). A subclassificação entre os tipos de sítio refere-se a sítio de ligação (B), sítio ativo (A), sítio conservado (C) ou modificações pós-translacionais – PTMs (P).

O servidor FTP do projeto fornece um arquivo SQL para criação das tabelas e os dados de cada tabela são disponibilizados separadamente em arquivos ASCII com estrutura tabular.

A versão do InterPro utilizada neste trabalho foi a 22.0. A Tabela 4.2.5 sumariza o conteúdo e a cobertura das anotações. A base de dados SQL contém 62 tabelas (Figura 4.4). A Figura 4.5 ilustra, de forma sintetizada, como os dados de nosso interesse estão organizados no banco do InterPro.

Tabela 4.1. Cobertura da versão 22.0 do InterPro

Tipo de anotação	Número de anotações
Active site	79
Binding site	52
Conserved site	506
Domain	5428
Family	11379
Region	1123
PTM	23
Repeat	253
Total de assinaturas	18843

Os dados relativos ao InterPro de interesse para este trabalho foram incorporados ao banco de alvos. A Figura 4.6 exibe dados sobre alguns termos do InterPro e a Figura 4.7 exibe dados sobre algumas anotações. Uma anotação associa uma dada assinatura (perfil, padrão etc.) a determinada região da estrutura primária de uma dada proteína.

```
abstract.tab            entry_xref.tab              pdb_pub_additional.tab
author.tab              etaxi.tab                   pfam_clan_data.tab
book2author.tab         example_auto.tab            pfam_clan.tab
book.tab                example.tab                 protein2genome.tab
common_annotation.tab   intact_data.tab            protein_accpair.tab
cv_database.tab         interpro2go.tab             protein_ida.tab
cv_entry_type.tab       journal_syn.tab             protein.tab
cv_evidence.tab         journal.tab                 proteome_rank.tab
cv_rank.tab             matches.tab                 pub2author.tab
cv_relation.tab         match_struct.tab            pub.tab
cv_synonym.tab          method2pub.tab              struct_class.tab
db_version.tab          method.tab                  supermatch.tab
entry2common.tab        mv_entry2protein.tab        tax_entry_count.tab
entry2comp.tab          mv_entry2protein_true.tab   tax_name_to_id.tab
entry2entry.tab         mv_entry_match.tab          taxonomy2protein.tab
entry2method.tab        mv_method2protein.tab       text_index_entry.tab
entry2pub.tab           mv_method_match.tab         uniprot_taxonomy.tab
entry_accpair.tab       mv_proteome_count.tab       varsplic_master.tab
entry_deleted.tab       mv_secondary.tab            varsplic_match.tab
entry_friends.tab       mv_tax_entry_count.tab      varsplic_supermatch.tab
entry.tab               organism.tab
```

Figura 4.4. Lista de tabelas do InterPro (versão 22.0).

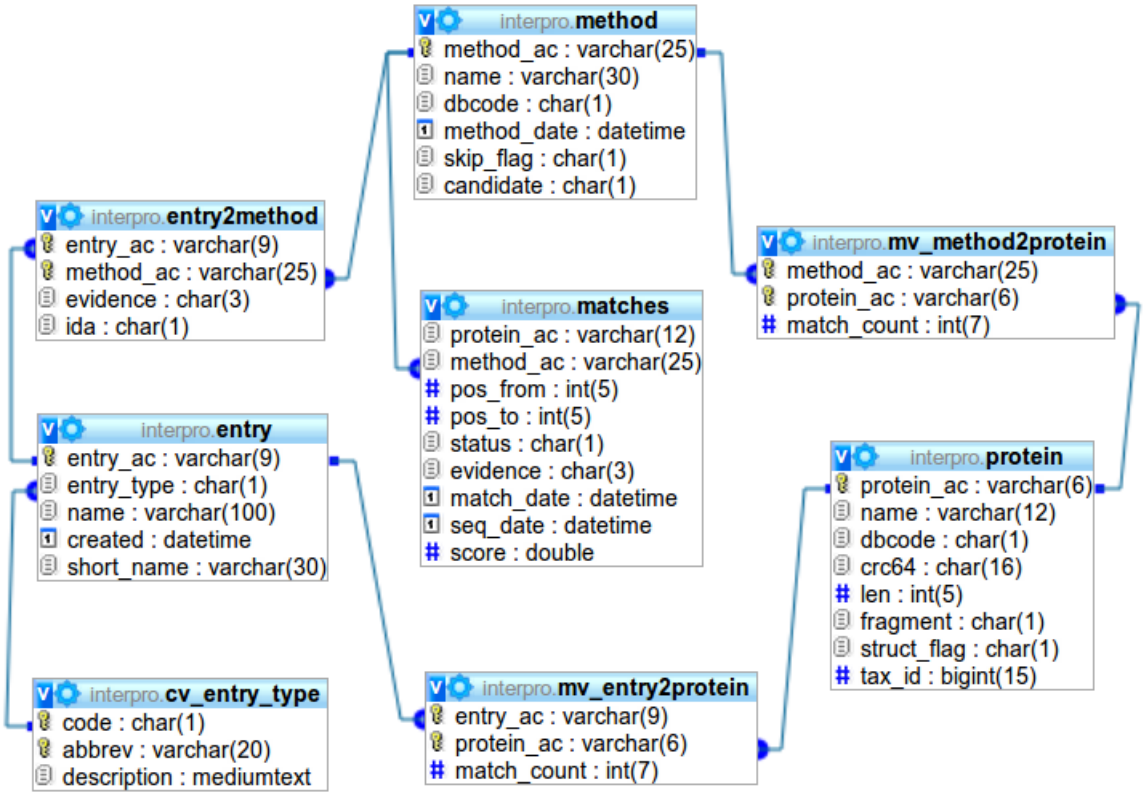


Figura 4.5. Parte do esquema da base de dados do InterPro. Uma dada assinatura (identificada pelo InterPro ID no campo entry_ac) é associada a uma dada região (demarcada pelas posições inicial e final) de uma dada sequência (identificada pelo UniProt AC no campo protein_ac). Cada registro explicita ainda o método (denotado pelo campo method_ac) empregado para definir a associação. Podem ocorrer associações idênticas (ou com diferenças sobre a região da sequência) obtidas por métodos distintos.

```
+-------------+---------+-----------------------------------------------+
| interpro_id | type_id | interpro_name                                 |
+-------------+---------+-----------------------------------------------+
| IPR000001   | D       | Kringle                                       |
| IPR000003   | F       | Retinoid X receptor                           |
| IPR000005   | D       | Helix-turn-helix, AraC type                   |
| IPR000006   | F       | Metallothionein, vertebrate                   |
| IPR000007   | D       | Tubby, C-terminal                             |
| IPR000008   | D       | C2 calcium-dependent membrane targeting       |
| IPR000009   | F       | Protein phosphatase 2A, regulatory subunit PR55 |
| IPR000010   | D       | Proteinase inhibitor I25, cystatin            |
| IPR000011   | G       | Ubiquitin-activating enzyme, E1-like          |
| IPR000012   | F       | Retroviral VpR/VpX protein                    |
+-------------+---------+-----------------------------------------------+
```

Figura 4.6. Dados sobre alguns termos do InterPro. Uma consulta que exibe o identificador, o tipo e o nome de algumas assinaturas.

uniprot_id	uniprot_ac	interpro_id	method_id	pos_ini	pos_end
A0A2G2_HUMAN	A0A2G2	IPR000421	PF00754	1	42
A0A2G2_HUMAN	A0A2G2	IPR000421	PS01286	29	45
A0A2G2_HUMAN	A0A2G2	IPR000421	PS50022	1	45
A0A2G2_HUMAN	A0A2G2	IPR008979	SSF49785	1	47
A0A2G3_HUMAN	A0A2G3	IPR000421	PF00754	1	42
A0A2G3_HUMAN	A0A2G3	IPR000421	PS01286	29	45
A0A2G3_HUMAN	A0A2G3	IPR000421	PS50022	1	45
A0A2G3_HUMAN	A0A2G3	IPR008979	SSF49785	1	47
A0A2G4_HUMAN	A0A2G4	IPR000421	PF00754	1	42
A0A2G4_HUMAN	A0A2G4	IPR000421	PS50022	1	45

Figura 4.7. Dados de algumas anotações do InterPro para algumas sequências proteicas humanas. Uma anotação consiste de uma assinatura descrita pelo termo *interpro_id*, determinada pelo método designado pelo código *method_id*, assinalada à região definida pelas posições *pos_ini* e *pos_end* da sequência identificada pelos códigos *uniprot_id* ou *uniprot_ac*.

4.2.6 DEG - Database of Essential Genes

O DEG é uma base de dados de genes essenciais para 21 organismos: 14 eucariotas e outros sete procariotas. Relaciona apenas genes e não proteínas. Neste trabalho, os genes foram mapeados para o correpondente UniProt ID a partir do gene ID (GI). Na versão atual do DEG, 6.8, os dados são disponibilizados para *download* no formato de arquivos de texto tabular (CSV). Na época em que o site do projeto foi consultado, a versão era a 5.4 e os dados eram disponibilizados apenas por meio da interface web da base de dados. Também aqui, foi construído um parser usando *perl* para extrair os dados dos arquivos HTML do DEG.

No total, foram obtidas 53.219 proteínas identificadas como essenciais na análise do DEG 5.4. Destas, 713 são proteínas humanas e o restante (52.506) está distribuído entre os outros 20 organismos.

Comparando com dados do DrugBank, foram encontradas 75 inconsistências – alvos assinalados como essenciais pelo DEG e como não-essenciais pelo DrugBank. O DrugBank classifica cada alvo como essencial (2724 alvos) ou não-essencial (1790 alvos).

Os dados sobre essencialidade não foram usados pelos métodos desenvolvidos neste trabalho. Mas foram mantidos na base de dados para estudos futuros.

4.3 Técnicas para recuperação de informação

Nesta seção apresentamos os fundamentos da metodologia empregada na construção do modelo vetorial apresentado na seção 4.5.

Aplicamos uma técnica de substituição dos vetores de dados originais em vetores correspondentes em um espaço de dimensão reduzida no qual relações de ordens superiores presentes na massa de dados original são abstraídas e incorporadas na base ortogonal. Essa transformação evidencia as similaridades e dissimilaridades mais extremas e faz aparecer relacionamentos entre dois membros que só poderiam ser detectados considerando-se associações não *prima facie* [Deerwester et al., 1990; Ampazis & Perantonis, 2004].

Quando diferentes termos podem ser usados para descrever um mesmo objeto ou conceito, diz-se que esses termos são sinônimos. Além disso, em um dado conjunto de informação, pode haver conceitos próximos ainda que distintos ou mesmo conceitos distintos mas correlacionados. E essas proximidades e correlações entre os conceitos nem sempre podem ser percebidas facilmente por estarem implícitas no volume de dados. Reciprocamente, um mesmo termo pode ser usado com significados bastante distintos quando aplicados em contextos diferentes (polissemia). O emprego de uma taxonomia específica para descrever os itens em determinada base de dados, como no caso do *Gene Ontology* (GO) ou do *InterPro*, pode minimizar a ocorrência de sinônimos e a polissemia. Mas mesmo nos casos em que uma taxonomia de termos tenha sido empregada, pode haver semelhanças e correlações implícitas, não vislumbradas *a priori*, que se formam na medida em que a base de dados aumenta utilizando-se os termos adotados inicialmente. Chagoyen et al. [2006] mostraram como os termos do GO que descrevem processos biológicos podem ser tratados em um espaço de dimensão reduzida para descobrir novas associações implícitas na literatura científica.

Depois de integrar as informações coletadas dos repositórios públicos, representamos os itens de interesse (alvos e fármacos) em subespaços definidos com o uso da decomposição por valores singulares [Eldén, 2007; Eldén, 2006].

4.3.1 Decomposição por valores singulares

A decomposição por valores singulares (SVD) é uma técnica oriunda da álgebra linear que permite reescrever uma matriz qualquer mxn ($m > n$), na forma da Equação 4.1;

$$A = USV^T \tag{4.1}$$

onde:

- A é a matriz mxn original;

- r é o chamado *posto* da matriz A e expressa o valor mínimo entre o número de linhas linearmente independentes ou o número de colunas linearmente independentes de A;
- U é uma matriz mxr e $U^TU = I$;
- S é uma matriz diagonal rxr;
- V é uma matriz nxr e $V^TV = I$.

A matriz U é a chamada matriz de autovetores à esquerda de A. Seus vetores representam a combinação linear das colunas de A. Por sua vez, a matriz V é a matriz de autovetores à direita de A, que são a combinação linear das linhas de A. A raiz quadrada dos autovalores de A^TA são dispostos em S em ordem decrescente e são conhecidos como valores singulares. Pode-se expressar A pelo somatório dos r termos formados pela multiplicação do valor singular s pelo produto dos autovalores u e v^T correspondentes (Equação 4.2).

$$A = \sum_{e=1}^{r} s_e u_e v_e^T \tag{4.2}$$

Os valores singulares diminuem a cada novo termo. Se essa queda de valores for acentuada, pode-se, então, obter uma aproximação A_k da matriz A tomando-se os k primeiros termos do somatório (Equação 4.3).

$$A \approx A_k = \sum_{e=1}^{k} s_e u_e v_e^T \tag{4.3}$$

Pelo teorema de Eckart-Young, a matriz A_k é uma matriz de posto reduzido associado à matriz A [Eckart & Young, 1936].

Estima-se que essa redução diminui o "ruído" associado a interelacionamentos menos significativos [Eldén, 2007; Berry et al., 1995; Dumais, 1992]. O efeito da decomposição por valores singulares e posterior redução de posto nas matrizes envolvidas está ilustrado na Figura 4.8.

Deerwester et al. [1990] usam a "forma truncada da decomposição por valores singulares" e formulam o termo "estrutura semântica" ao descrever o relacionamento entre palavras em um documento textual. Este conceito é estendido por Dumais [1992] ao apontar que ele se aplica não apenas a textos, mas também a outros objetos que

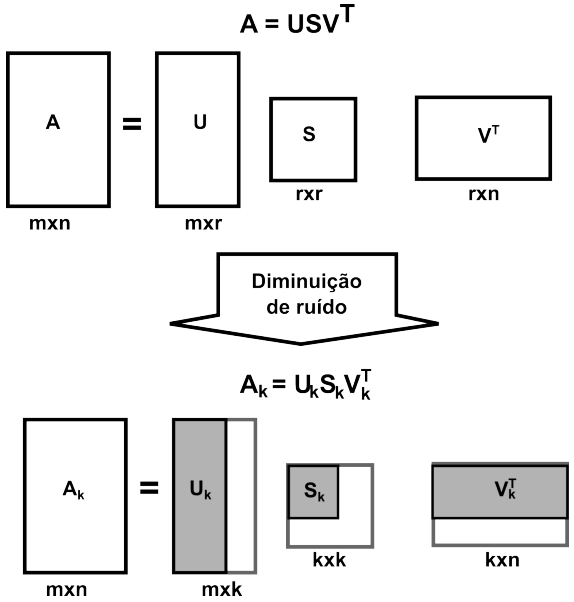


Figura 4.8. A decomposição por valores singulares e a posterior redução de posto. Baseado em Chen et al. [2008].

podem ser descritos textualmente. A hierarquia entre os termos de uma dada ontologia pode ser tratada da mesma maneira que a "estrutura semântica" das palavras em um texto. De modo geral, uma relação semântica entre dois atributos indica que, se o primeiro é incluído em uma regra, o segundo também é [Witten & Frank, 2011]. Uma vantagem da representação dos objetos no espaço obtido pela decomposição por valores singulares seguida da redução de posto é que dois objetos podem ser considerados muito similares mesmo quando não compartilham nenhum descritor [Dumais, 1992]. Outra vantagem da redução de posto é que ela favorece a discriminação entre os objetos. Isto pode ser vislumbrado comparando-se as versões *full* e reduzida do gráfico dos valores ordenados de uma métrica de similaridade calculados para uma dada proteína em relação às outras (Figura 4.9). A matriz original esparsa resulta em alguns poucos degraus referentes às associações diretas. Em um mesmo intervalo, a redução de posto ocasiona um maior número de patamares o que indica uma discriminação mais detalhada entre os membros do conjunto. A proteína selecionada aleatoriamente foi a 5HT7R_HUMAN.

O problema de se encontrar o melhor valor de k para a redução de posto permanece em aberto. Usualmente, emprega-se um método visual a partir da curva de decaimento dos valores singulares. Esse método é denominado *teste de scree* [Cattell, 1966]. O termo "*scree*" é derivado da geologia, onde refere-se à base rochosa de uma escarpa. Traçando-se o gráfico dos valores singulares obtidos com a fatoração por valores singulares, pode-se observar o decréscimo da variação de seus valores. O "teste

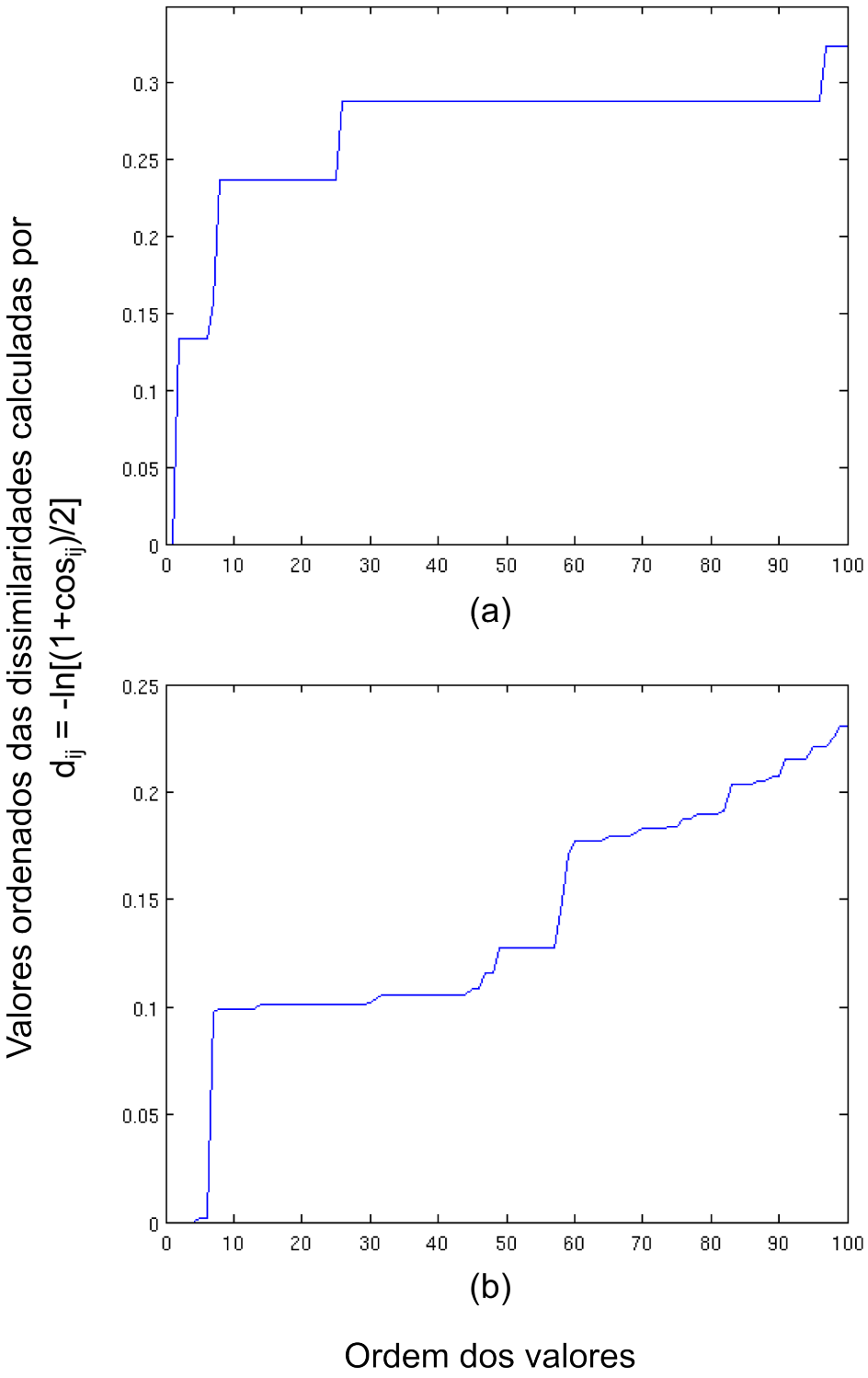


Figura 4.9. Melhoria da recuperação de informação através da redução de posto. A comparação entre os gráficos com os 100 primeiros valores ordenados das dissimilaridades entre a proteína 5HT7R_HUMAN e todas as outras do conjunto calculados sobre a matriz original **(a)** e sobre a matriz com posto reduzido **(b)**. Este exemplo ilustra a maior discriminação no caso do espaço reduzido. Esse comportamento favorece o desempenho dos algoritmos de agrupamento e a recuperação de informação.

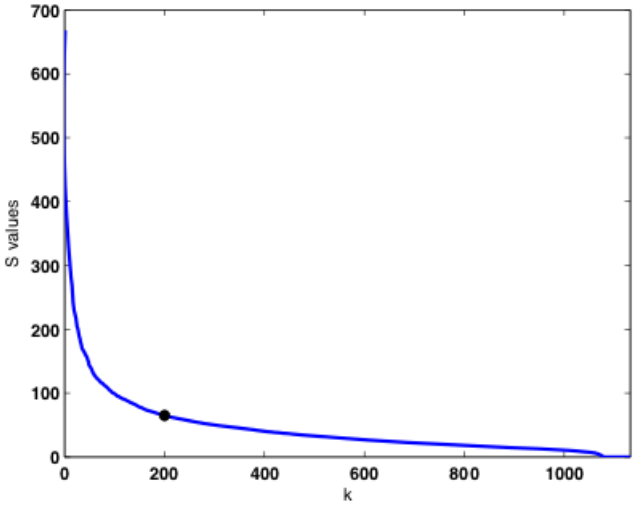


Figura 4.10. O teste de *scree*. Retirado de Chagoyen et al. [2006]. Aplicou-se SVD no desenvolvimento de uma métrica de similaridade entre publicações científicas a partir de uma matriz com as frequências de termos do GO citados nesses documentos. Na figura, mostra-se os valores singulares e a redução de posto $k = 200$ determinada pelo teste de "scree".

de *scree*" consiste em interpretar este gráfico como se fosse a ilustração de uma escarpa e tentar identificar visualmente a base dessa escarpa. Obviamente, esse é um teste bastante subjetivo mas que é largamente empregado nesse tipo de estudo (Figura 4.10).

Uma vez obtida a projeção dos dados no espaço vetorial reduzido, o próximo passo é encontrar uma métrica que permita comparar a similaridade entre os objetos. Novos membros podem ser projetados no espaço utilizando-se a matriz U conforme descrito na Equação 4.4;

$$q_k^* = q^T U_k; \tag{4.4}$$

onde:
q é o vetor do novo membro representando usando-se o mesmo conjunto de descritores iniciais; q_k^* é o correspondente vetor projetado no espaço vetorial reduzido (de posto k) e; U_k é a matriz de posto k de valores singulares da esquerda obtida pela fatoração da matriz original A por SVD.

4.3.2 Métricas de similaridade

A similaridade entre dois objetos no espaço vetorial reduzido obtido pela decomposição por valores singulares pode ser definida pelo coseno entre os vetores que representam os objetos (Equação 4.5).

$$sim\_cos_{ij} = \cos(\alpha_{ij}) = \frac{c_i c'_j}{\sqrt{c_i c'_i}\sqrt{c_j c'_j}} \tag{4.5}$$

onde:
- c_i representa a *i-ésima* linha da matriz $V_k S_k$;
- c'_i representa o vetor transposto de c_i
- sim_cos_{ij} é a métrica de similaridade baseada em coseno entre os objetos i e j;
- α_{ij} é o ângulo entre os vetores que representam os objetos i e j

Objetos muito similares tendem a apresentar um coseno do ângulo entre eles próximo do valor unitário. O coseno assim definido é equivalente ao produto vetorial entre duas linhas da matriz $V_k S_k$. Portanto, a medida de similaridade entre cada par de objetos do conjunto analisado é obtida a partir do cálculo do produto vetorial das respectivas linhas de $V_k S_k$.

Também, pode-se definir a métrica de similaridade entre dois objetos no espaço vetorial reduzido como a distância euclidiana entre dois pontos no espaço $V_k S_k$ (Equação 4.6).

$$sim\_dist_{ij} = \sqrt{\sum_{z=1}^{k}(x_{i,z} - x_{j,z})^2} \tag{4.6}$$

onde:
- sim_dist_{ij} é a distância euclidiana entre os objetos i e j;
- $x_{i,z}$ indica a *z-ésima* coordenada do vetor que representa o objeto i na matriz $V_k S_k$.

4.4 Agrupamento e visualização de dados

O rápido avanço de tecnologias de alto-desempenho apresentado em pesquisas de larga escala traz a necessidade de técnicas de tratamento, análise e visualização de grandes quantidades de informação. A identificação de novas classificações e seu relacionamento com o universo já conhecido auxiliam na descoberta de novos atributos ocultos no

grande volume de informação. A redução de posto e a capacidade de visualização dos dados são aspectos chaves para a análise desta composição em larga escala [Devarajan, 2008].

Há diversos algoritmos de agrupamento. A escolha do mais apropriado para analisar a base de dados em estudo depende de uma série de fatores. Alguns algoritmos requerem a definição *a priori* de algum parâmetro (e.g. o número estimado de grupos) ou ação do usuário iterativamente.

4.4.1 Algoritmo de agrupamento hierárquico

O algoritmo de agrupamento hierárquico é muito usado em bioinformática [Weinstein, 2008]. É um dos métodos mais simples. O algoritmo tem como entrada uma matriz de similaridade ou dissimilaridade. A partir desta matriz, o algoritimo aplica alguma métrica escolhida pelo usuário. O objetivo neste passo é escolher uma métrica que o próprio algoritmo poderá recalcular sempre que preciso. As métricas mais usuais são a correlação de Pearson e a distância euclidiana. Então, os dois pontos mais similares entre si no conjunto (segundo a métrica adotada pelo algoritmo) são associados a um mesmo grupo (este é o primeiro grupo criado). Depois disso, o algoritmo recalcula os valores da métrica entre cada par do novo conjunto, onde os dois primeiros objetos agrupados são agora tratados como um único objeto. Um outro parâmetro definido pelo usuário (além da métrica usada pelo algoritmo) é a forma de lidar com os cálculos relativos a cada novo agrupamento. Por exemplo, ao calcular a métrica em relação a um dado objeto fora de um grupo e o grupo em questão, pode-se usar a média dos valores considerando cada objeto no grupo (opcionalmente, ao invés da média, poderia-se usar o maior ou o menor valor etc.).

Uma das desvantagens do algoritmo hierárquico é que quando um objeto é colocado em algum grupo ele não mais o remove desse grupo. Entretanto, em casos em que a distribuição dos objetos não seja muito bem definida, pode ocorrer de um dado objeto ser associado a algum grupo "equivocadamente". Esse problema é menor para casos em que a distribuição dos grupos seja mais bem definida. E é nesse ponto que um pré-processamento da base de dados de modo a remover os ruídos residuais pode ajudar a melhorar a performance do algoritmo de agrupamento, qualquer que seja ele, sobretudo para um algoritmo tão simples quanto o hierárquico. O pré-processamento dos dados usando a decomposição por valores singulares e redução de posto pode favorecer a performance de um eventual método de agrupamento empregado.

Os algoritmos de agrupamento usados neste trabalho estão implementados no *Multi environment tool* (MeV) [Howe et al., 2010] e no Cytoscape [Cline et al., 2007]

(na visualização dos dados em redes). O MeV requer que a matriz de entrada seja uma matriz de similaridade. Para estes casos, a matriz de "distância ou dissimilaridade foi convertida para uma matriz de similaridade dividindo-se os valores pelo valor máximo e subtraindo de 1.

4.4.2 Algoritmo de agrupamento por fatoração de matriz não-negativa

A fatoração de matriz não-negativa (NMF) é uma técnica com aplicações em biologia computacional [Devarajan, 2008]. Dada uma matriz não-negativa A, ela é fatorada em duas matrizes não-negativas W e H e uma matriz residual U que pode conter valores negativos:

$$nmf(A) = WH + U \tag{4.7}$$

O problema de fatoração consiste em um problema de mínimos quadrados. Dada uma matriz não-negativa $A_{n,m}$ e um inteiro positivo $d < min(n, m)$, deseja-se encontrar as matrizes não-negativas $W_{n,d}$ e $H_{d,m}$ que minimizem a função:

$$f(W, H) = \frac{1}{2}||A - WH||_F^2, \tag{4.8}$$

onde $||X||_F = \sqrt{\sum_{i=1}^{m} \sum_{j}^{n} = 1|a_{ij}|^2}$ representa a *norma de Frobenius*.

O produto WH é chamado de fatoração não-negativa de A. Este produto representa uma aproximação de posto d da matriz original A.

A NMF constitui um problema de otimização onde obtem-se uma redução da dimensionalidade mas preservando a característica não-negativa da representação. O resultado é uma decomposição puramente aditiva: o "todo" é obtido pela soma das "partes".

Há várias formas de determinar W e H. Lee & Seung [2000] desenvolveram um método de atualização multiplicativa. Este é o método usado no algoritmo de agrupamento por NMF implementado no MeV. O usuário deve fornecer o número de *clusters* nos quais os elementos nas colunas da matriz de entrada serão distribuídos. O número de *clusters* é também o valor do posto desejado. O algoritmo executa múltiplas iterações variando o posto de 1 até o valor definido pelo usuário. Essas sucessivas iterações permitem definir uma árvore hierárquica entre os *clusters* (não exibida pelo

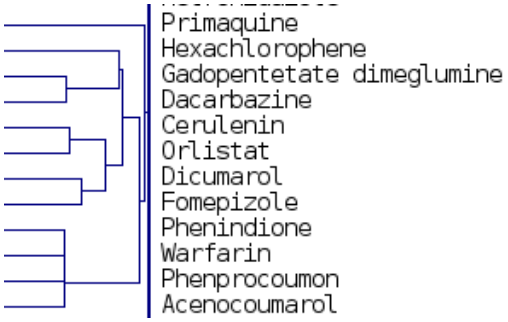


Figura 4.11. Um exemplo de dendrograma. Exibe-se o relacionamento entre alguns medicamentos.

programa) que lhe permite calcular o coeficiente de correlação cofenética.

4.4.3 Dendrograma e coeficiente de correlação cofenética

É comum representar o resultado de um algoritmo de agrupamento, sobretudo quando usado um algoritmo hierárquico, na forma de *dendrograma* (Figura 4.11). Como esta é uma forma simplificada de representar em duas dimensões uma relação contida em um espaço usualmente de dimensão muito superior, é natural que haja algumas distorções quanto à representatividade do dendrograma. Um método de agrupamento pode ser considerado melhor do que outro quando o dendrograma do primeiro fornece uma representação menos distorcida da realidade em comparação com o segundo. É possível avaliar o grau de deformação do dendrograma calculando o chamado coeficiente de correlação cofenética [Sokal & Rohlf, 1962]. Alternativamente, realiza-se apenas uma análise qualitativa dos resultados.

O coeficiente de correlação cofenética é calculado da seguinte forma: dada um conjunto de dados original A e sua representação em um dendrograma T, define-se as seguintes medidas de distâncias:

- $d(i, j)$ é a distância euclidiana entre os elementos i e j;
- $t(i, j)$ é a altura do nó ao qual esse dois pontos foram reunidos pela primeira vez.

Seja x a média de $x(i, j)$ e t a média de $t(i, j)$. O coeficiente de correlação cofenética c é dado por:

$$c = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(t(i,j) - t)^2]}} \tag{4.9}$$

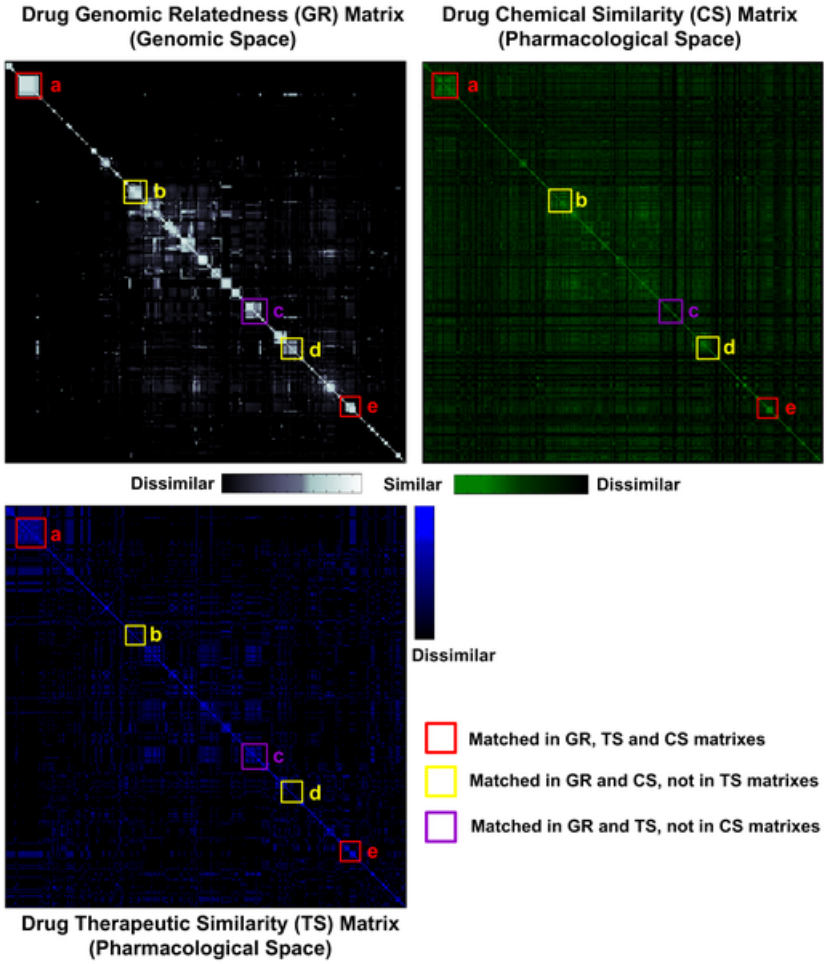


Figura 4.12. Mapas de calor. A representação dos dados em mapas de calor. Baseado em Chen et al. [2008]. Ilustra-se a similaridade entre fármacos de um dado conjunto de fármacos segundo seus relacionamentos no espaço genômico (GR); segundo as similaridades químicas (CS) e; segundo as similaridades terapêuticas (TS). A ordem dos elementos foi determinada pelo algoritmo de agrupamento aplicado ao primeiro caso (GR). Pode-se identificar visualmente a ocorrência de alguns grupos equivalentes nos três diferentes espaços.

4.4.4 Mapas de calor (*heatmaps*)

Um mapa de calor (*heat map* ou *heatmap*) é uma representação gráfica dos dados de uma matriz cujos elementos são valores numéricos. Os valores são representados de forma relativa utilizando-se uma escala de cor. Na representação mais comum, utiliza-se um gradiente entre duas cores. Por exemplo, valores de similaridade máxima (entre um elemento e ele mesmo) é denotado com a intensidade máxima de uma dada cor escolhida enquanto o menor valor de similaridade no conjunto é denotado pela total ausência dessa cor e na máxima intensidade da segunda cor escolhida (Figura 4.12).

A matriz de entrada pode apresentar dois tipos de entidades: uma representada

pelas linhas, a outra representada pelas colunas; ou o mesmo tipo de entidade tanto nas linhas como nas colunas. Os itens das linhas e colunas podem, inclusive, ser os mesmos, resultando em uma matriz quadrada que relaciona cada item com todos os outros do conjunto. Assim, a matriz original pode ser: uma matriz de expressão gênica, em que os genes são representados nas colunas e as amostras nas linhas; uma matriz de similaridade ou dissimilaridade, segundo alguma métrica, entre as proteínas de uma base de dados e as proteínas de um grupo de teste; uma matriz (quadrada) de similaridade ou dissimilaridade entre todos os pares de elementos de uma determinada base de dados; etc.

O mapa de calor é uma das formas preferidas de se representar os resultados de algoritmos de agrupamento em publicações no ramo das ciências biológicas [Weinstein, 2008]. Usualmente, acrescenta-se o uso de dendrogramas em uma ou duas margens do mapa de calor.

Neste trabalho, mapas de calor e dendrogramas são usados para ilustrar os resultados do algoritmo de agrupamento hierárquico sobre a matriz de dissimilaridade baseada no nosso modelo vetorial e também sobre a matriz de similaridade par-a-par baseada no *bitscore* do BLAST.

4.4.5 Redes

O uso de redes é comum para representar as interelações entre objetos em uma base de dados biológicos. Em geral, esses objetos são do mesmo tipo (e.g. todos indicando alguma proteína) ou de alguns poucos tipos diferentes (e.g. proteínas, fármacos e doenças).

A teoria das redes engloba conceitos não apenas relacionados com a representação dos dados, mas também com métricas de similaridade; nós críticos etc. Neste trabalho, utilizamos redes com o único objetivo de visualização dos dados da base em estudo.

Para visualizar em uma rede como os objetos se distribuem no contexto de um dado tipo de relação, pode-se utilizar um programa específico para este fim. Alguns pacotes estatísticos como o *R* [Meur & Gentleman, 2011]. contêm recursos para construção de redes. Mas o mais usado para a visualização final dos dados em pesquisas nas áreas biológica e afins talvez seja o *Cytoscape* [Cline et al., 2007]. E este é o programa utilizado em artigos recentes que consultamos para comparar nossos dados (Figura 4.13).

Na construção de uma rede, o programa primeiramente desenha todos os objetos do conjunto em estudo e as linhas que os interligam considerando-se a métrica de similaridade ou dissimilaridade adotada e o ponto de corte assumido.

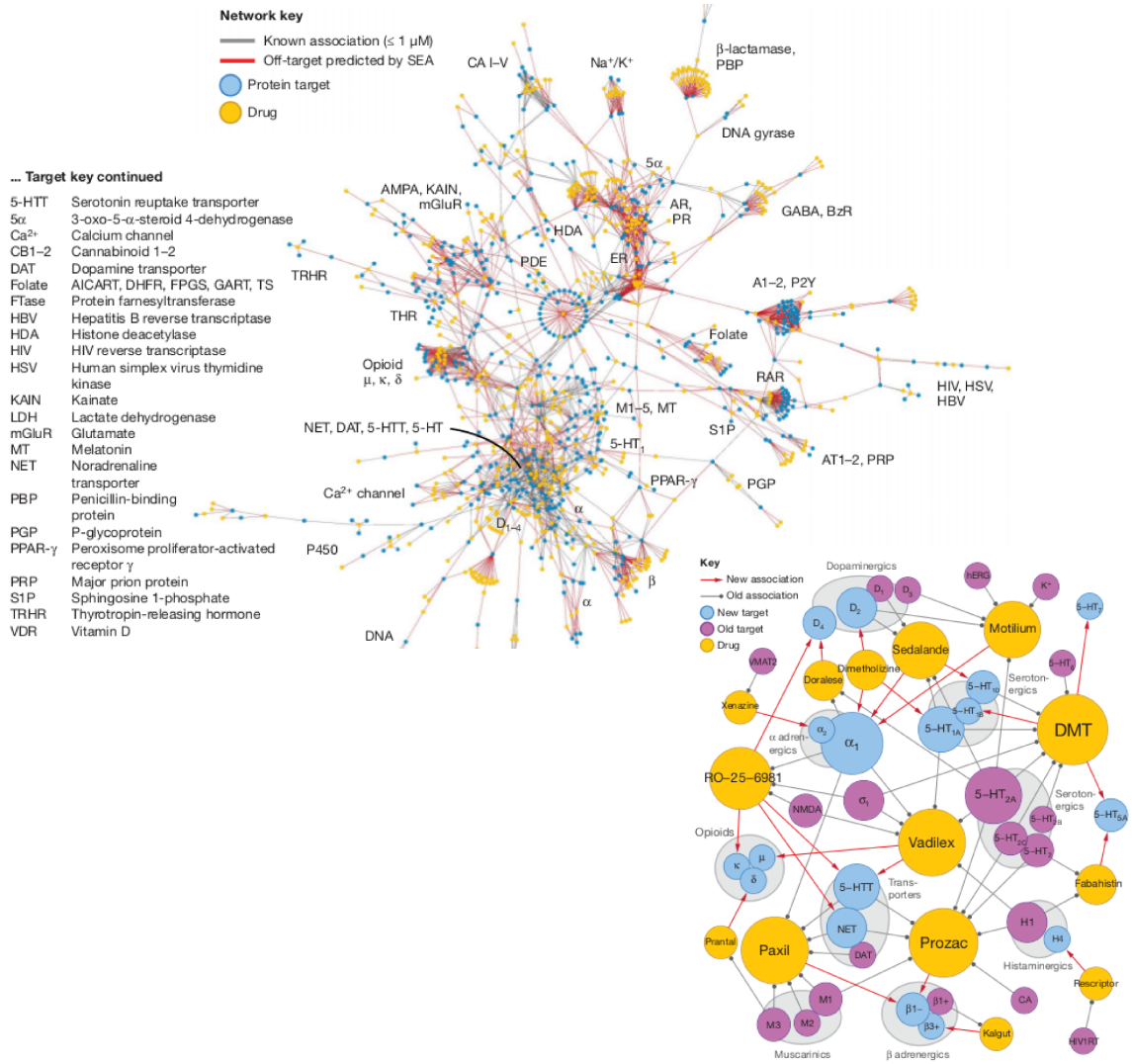


Figura 4.13. Visualização de dados em uma rede de interações. Usado por Keiser et al. [2009] para ilustrar a interação entre fármacos e alvos.

Uma vez desenhado todos os objetos e associações desejados, o posicionamento dos nós (que, no nosso caso, representam os objetos) pode ser reorganizado de várias maneiras. O modo mais usual para reorganizar os nós é utilizando um algoritmo "baseado em força". Nesse algoritmo, a posição dos nós é reorganizada iterativamente tratando a matriz de similaridade ou dissimilaridade como se ela representasse a força de atração ou repulsão entre os objetos. O algoritmo age de modo que, na configuração final, a proximidade entre cada objeto de um dado par de objetos esteja coerente com os valores das similaridades entre eles. Quanto maior a similaridade entre dois objetos, os nós que os representam são posicionados mais próximos entre si.

4.4.6 Projeção em espaço de duas ou três dimensões

Um outro método de visualização usado neste trabalho foi o método de otimização proposto por Marcolino et al. [2010]. Este método realiza uma projeção da base de dados em um espaço tri-dimensional. Nesse método, cada objeto da base de dados é tratado como um ponto no espaço R^m, onde m é o número de descritores contidos na base (ou seja, é a dimensão do espaço original). A distância euclidiana δ_{ij} nesse espaço multidimensional é calculada para cada par de objetos. Para representar a base de dados em um espaço reduzido (de duas ou três dimensões), o algoritmo procura ajustar um modelo onde os correspondentes valores da distância euclidiana γ_{ij} sejam próximos dos valores no espaço original. Matematicamente, estamos interessados em minimizar a função erro dada pela Equação 4.10.

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_{ij} - \gamma_{ij})^2, \tag{4.10}$$

onde δ_{ij} é a distância no espaço original e γ_{ij} é a distância no espaço reduzido.

Diferentes técnicas podem ser usadas para solucionar esse problema de otimização. Xie et al. [2000] usam o método de Newton truncado. O algoritmo de Marcolino et al. [2010] utiliza o método de Newton reflexivo [Coleman & Li, 1994] implementado no *MatLab otimization tool-box*™.

O método foi usado sobre a matriz de dissimilaridade baseada no modelo vetorial e também sobre a matriz de similaridade par-a-par baseada no *bitscore* do BLAST.

4.5 Representação vetorial dos alvos

A partir dos dados coletados das bases públicas, foi construído um modelo vetorial para representar os alvos drogáveis. A Figura 4.14 descreve o algoritmo empregado. Inicialmente, os alvos conhecidos foram obtidos do TTD, do DrugBank e do KEGG-DRUG. Foram obtidos 1906 alvos humanos drogáveis não redundantes. 365 deles foram reservados para testes de validação, restando 1541 para a construção do modelo. As anotações do InterPro para o conjunto de alvos foram usadas para obter outros candidatos do UniProtKB. Destes, aqueles que apresentaram pior alinhamento par-a-par usando BLAST, com cada alvo, foram selecionados para compor o grupo de não-alvos do conjunto de validação.

Cada um dos alvos é uma proteína identificada por um *UniProt ID* e representada por um vetor coluna de uma matriz, onde cada linha representa um determinado termo

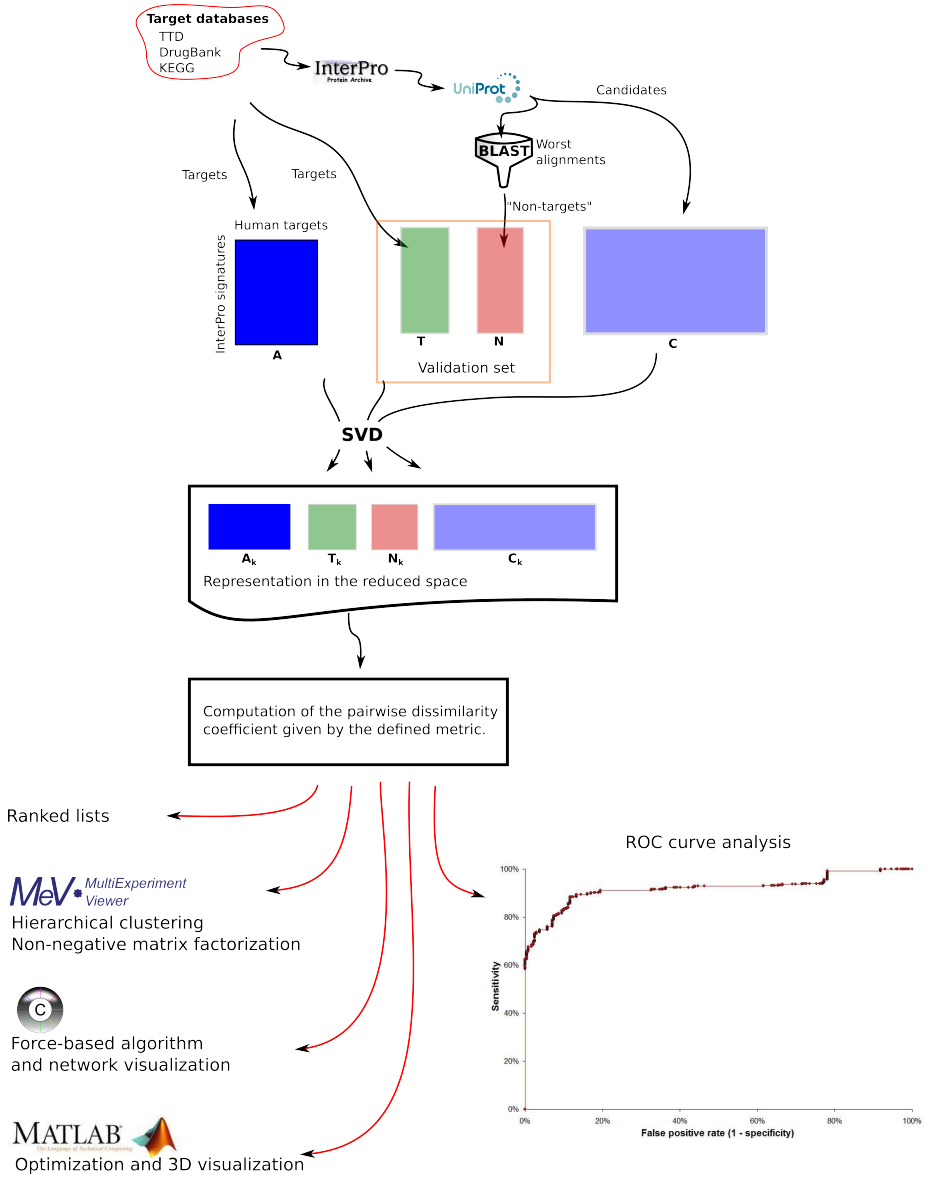


Figura 4.14. Fluxograma para a costrução do modelo vetorial para alvos drogáveis. Entre os alvos conhecidos obtidos em bases de dados públicas, alguns são selecionados para compor o grupo positivo do subconjunto de validação e o restante é usado para a construção do espaço vetorial. A partir das assinaturas do InterPro associadas aos alvos, outras sequências são selecionadas do UniProtKB. Aquelas com menores índices de similaridade de sequências com os alvos são classificadas como "não-alvos" e o restante é tratado como candidatos a alvos drogáveis. A decomposição por valores singulares e a redução de posto determinam a descrição dos alvos no espaço reduzido e a projeção das outras sequências nesse mesmo espaço. Tomando-se os vetores reduzidos, calcula-se os coeficientes de dissimilaridade entre cada entidade e cada um dos alvos iniciais. Diferentes métodos de análise e visualização são empregados. Os dados referentes ao grupo de validação são usados na construção da curva ROC.

do InterPro do tipo Família (F), Domínio (D) ou região (G) ou uma combinação entre um desses termos e outro de sítio ativo ou de ligação. Foram usados 2700 descritores binários sendo:

- 1069 termos do tipo F;
- 1244 termos do tipo D;
- 77 termos do tipo G;
- 310 descritores que indicam se uma dada assinatura de tipo F, D ou G contém uma dada assinatura de sítio ativo (A) ou de ligação (B).

Para definir os termos do InterPro para descrever os alvos, foi tomado inicialmente todos aqueles dos tipos F, D ou G associados a algum alvo conhecido. Depois, foram selecionados todos aqueles dos tipos A e B cuja anotação refere-se a uma região da sequência englobada por alguma anotação do tipo F, D ou G. Com exceção desse cuidado especial quanto às anotações relativas a sítios específicos, não foram levados em consideração a estrutura hierárquica dos termos do InterPro. É deixado para que a própria técnica trate essa característica *per se*.

A matriz $A_{2700x1541}$ inicial foi submetida à decomposição por valores singulares seguida da redução de posto para a eliminação de ruído conforme demonstrado por [Chen et al., 2008; Eldén, 2007; Berry et al., 1995; Dumais, 1992]:

$$A \approx U_k S_k V_k^T \tag{4.11}$$

A fatoração foi realizada usando o MATLAB [MATLAB, 2010] e a redução de posto foi definida pelo *teste de scree* (Figura 4.15). Foram experimentados diferentes valores para o posto k. Para valores menores que 320, alguns autovetores tornaram-se todos nulos indicando um corte impróprio. Valores maiores produziram pouca alteração na visualização dos dados. Observa-se que o valor definido para o posto é da ordem de gradeza daquele definido por Chagoyen et al. [2006] no estudo envolvendo a ocorrência de termos do GO na literatura científica. Isso era esperado uma vez que se sabe que o posto indica o número aproximado de agrupamentos no conjunto de dados e a taxonomia no InterPro é, em grande parte, definida diretamente a partir da taxonomia no GO [Burge et al., 2012; Zdobnov & Apweiler, 2001].

Após a redução de posto, o próximo passo foi determinar a similaridade entre as proteínas a partir de alguma métrica. Foram calculados as métricas de distância euclidiana (Equação 4.6) e de coseno (Equação 4.5).

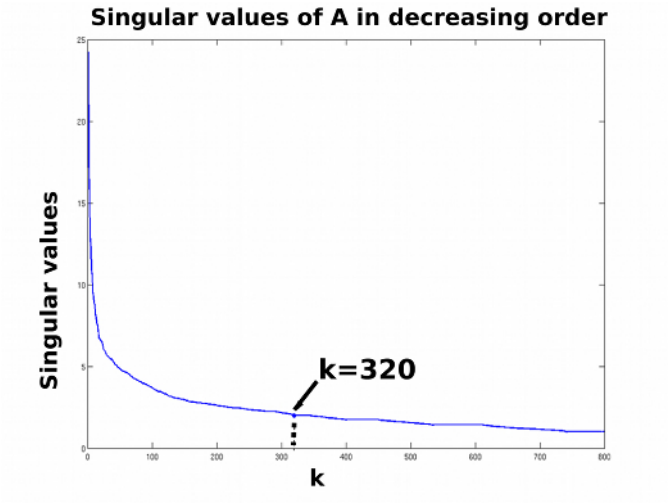


Figura 4.15. O teste de *scree test* aplicado à matriz reduzida de alvos humanos. Valores singulares de A (obtidos pela fatoração por SVD) estão plotados na ordem decrescente. O eixo x corresponde ao posto. O valor de corte $k = 320$ foi definido pelo *teste de scree*.

Como alguns programas de visualização requerem que os dados estejam relacionados por alguma métrica de similaridade ou dissimilaridade, a métrica do coseno foi convertida para uma métrica de dissimilaridade equivalente à métrica de distância evolutiva proposta por Stuart et al. [2002]:

$$d_{ij} = -ln((1 + cos_{ij})/2), \tag{4.12}$$

onde:
- d_{ij} representa a métrica de dissimilaridade entre os alvos i e j e;
- cos_{ij} representa o coseno entre os seus respectivos vetores no espaço reduzido.

4.6 Regressão logística

Nesta seção, são discutidos os fundamentos sobre regressão logística, que foi utilizada neste trabalho para discriminar as anotações relevantes para a drogabilidade de um alvo.

Estimando-se que haja uma relação entre uma variável explicativa X e uma variável resposta dicotômica Y, pode-se usar o valor da variável X para prever a probabilidade de "sucesso" para Y. Para fazê-lo, usa-se a técnica chamada de regressão logística univariada.

Para desenvolver um modelo de regressão nesse caso, parte-se da estratégia inicial

de ajustar um modelo em uma forma parecida com o modelo de regressão linear:

$$p = \beta_0 + \beta_1 x, \tag{4.13}$$

onde p representa a probabilidade de "sucesso" para Y e x representa o valor da variável explicativa X. Esse é simplesmente o modelo de regressão linear. Entretanto, esse modelo não atende à condição de que p seja uma probabilidade, porque ela deve estar restrita a assumir valores entre 0 e 1. Tenta-se, então, ajustar o modelo:

$$p = f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} \tag{4.14}$$

onde $f(x)$ é a função logística.

A *chance* de um dado evento é a razão entre a probabilidade desse evento ocorrer pela probabilidade dele não ocorrer (Equação 4.15).

$$r = \frac{p}{1-p}, \tag{4.15}$$

onde r é a chance a favor de um dado evento que tem probabilidade p de ocorrer.

Enquanto a probabilidade é uma medida que vai de 0 a 1, a chance pode ir de 0 a infinito. Geralmente, as chances são expressas como razões. Por exemplo, a chance a favor de um evento que tem probabilidade 0,80 de ocorrer é de 4 por 1.

No caso da regressão logística, falamos de uma chance em favor de sucesso dada por:

$$\frac{p}{1-p} = \frac{\dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}{\dfrac{1}{1 + e^{\beta_0+\beta_1 x}}} = e^{\beta_0+\beta_1 x} \tag{4.16}$$

Podemos dizer que modelar a probabilidade p de uma variável resposta dicotômica Y usando uma função logística equivale a ajustar um modelo de regressão linear para o logaritmo natural da chance de sucesso daquela variável. Em vez de assumir que a relação entre p e x seja linear, assume-se que a relação entre a chance de sucesso de Y e a variável explicativa X seja linear. A técnica de ajustar um modelo dessa forma é chamada de *regressão logística*.

Tabela 4.2. Tabela de contingência 2x2 para duas variáveis dicotômicas

Y	X		Total
	Sim	Não	
Sim	a_{11}	a_{12}	$R_1 = a_{11} + a_{12}$
Não	a_{21}	a_{22}	$R_2 = a_{21} + a_{22}$
Total	$C_1 = a_{11} + a_{21}$	$C_2 = a_{12} + a_{22}$	$\sum a_{ij}$

A técnica requer, então, ajustar o modelo

$$ln[\frac{\hat{p}}{1-\hat{p}}] = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{4.17}$$

Em um modelo de regressão linear, $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores dos coeficientes da população. Entretanto, não se pode usar o método dos mínimos quadrados (usual na regressão linear) para ajustar o modelo logístico. Isso porque o método dos mínimos quadrados assume que a resposta é contínua e com distribuição normal. O modelo logístico é ajustado por *estimação de verossimilhança*. Essa técnica usa a informação de uma amostra para encontrar as estimativas dos parâmetros mais prováveis de terem produzido os dados observados [Pagano & Gauvreau, 2004].

Após determinar os valores dos coeficientes β_0 e β_1 do modelo, é preciso avaliar se eles são adequados. Em geral, fala-se em avaliar a significância da variável explicativa na predição da probabilidade de sucesso da variável resposta.

Dentre os possíveis testes de significância, pode-se citar: a razão de máxima verossimilhança; o teste de Qui-quadrado; o teste de *Wald*; o teste de *score*; o teste exato de *Fisher*. Com exeção deste último, esses testes assumem que o tamanho da amostra seja grande para que os parâmetros estimados apresentem distribuição normal ou de Qui-quadrado. Há ocasiões em que o tamanho da amostra não é grande o suficiente para justificar essas suposições. Os chamados testes exatos não requerem suposição quanto à distribuição da amostra.

O *teste exato de Fisher* é mais apropriado para dados de uma amostra pequena e que podem ser expostos em uma tabela de contigência 2x2. Tivemos um especial interesse no teste exato de *Fisher* neste trabalho porque tratamos apenas variáveis explicativas que, também são dicotômicas. Para cada variável explicativa X poderemos representar sua relação com a variável reposta Y (ambas dicotômicas) em uma tabela 2x2 (Tabela 4.2).

Alternativamente, podemos representar esta tabela por uma matriz A_{2x2} em que

cada elemento a_{ij} é o número de observações em que $y = y_i$ e $x = x_i$. Calcula-se as somas dos valores por linha (R_i); por coluna (C_j); e a soma total $N = \sum R_i = \sum C_j$. Calcula-se, então, a probabilidade condicional p_{cutoff} para a situação representada por essa matriz (Equação 4.18).

$$p_{cutoff} = \frac{(R_1!R_2!)(C_1!C_2!)}{N! \prod a_{ij}!} \tag{4.18}$$

Depois disso, encontra-se todas as matrizes de valores inteiros não-negativos que resultam nos mesmos valores de R_i e C_j. Diz-se que essas matrizes são "consistentes" com a matriz dos valores observados. Para cada uma dessas matrizes, calcula-se a probabilidade condicional usando-se a Equação 4.18. A soma dessas probabilidades deve ser 1.

Para calcular o *p-value* do teste, as matrizes consistentes devem ser ordenadas segundo algum critério que meça a dependência. Esse critério pode ser a máxima verossimilhança (o mais usual); Qui-quadrado; ou algum outro método. O cálculo do *p-value* é realizado somando-se as probabilidades de um subconjunto dessas matrizes. O critério de seleção de quais matrizes devem entrar nesse cálculo é um problema aberto – não há uma fórmula fechada para resolvê-lo [Armitage et al., 2002]. Uma abordagem possível é somar as probabilidades das matrizes com probabilidade menor do que ou igual à probabilidade da matriz observada. O *p-value* calculado desta forma é chamado de *one-sided p-value* ou *one-tailed p-value* [Campbell et al., 2009]. Alguns autores criticam esse método de cálculo por ele ser muito conservativo [Hirji et al., 1991]. Armitage et al. [2002]; Hirji et al. [1991] defendem o uso de um cálculo chamado de *mid p-value*, onde o valor do *p-value* observado é dividido pela metade antes de ser somado aos outros valores. Este valor é menos conservativo do que o *one-sided p-value*. O chamado *two-sided p-value* ou *two-tailed p-value* pode ser calculado de diferentes formas, mas, em geral, pode ser aproximado simplesmente dobrando o valor do *one-sided p-value* ou do *mid p-value* [Campbell et al., 2009].

Por fim, para determinar se uma dada variável explicativa é significativa para predizer o resultado da variável resposta, é preciso definir o valor de corte para o *p-value*. Na falta de algum critério mais restritivo, assume-se o valor usual de 5%. Ou seja, uma variável explicativa é considerada significativa e mantida no modelo se o valor de *p-value* determinado pelo *teste exato de Fisher* for menor que 0,05. O *teste exato de Fisher* deve ser considerado no ajuste de modelos de regressão logística com amostra de tamanho pequeno [Hosmer & Lemeshow, 2004].

Vale ainda fazer algumas interpretações adicionais sobre o modelo logístico, so-

bretudo quando as variáveis são todas dicotômicas (que é nosso caso de interesse).

Pela Equação 4.17, pode-se concluir que, se ambos os coeficientes β_0 e β_1 forem negativos, a probabilidade de "sucesso" (p) da variável resposta é menor do que a de "falha" ($p-1$). Daí, poderíamos concluir que a variável explicativa (se significativa) colabora para reduzir a chance de sucesso da variável resposta. Além disso, segundo Pagano & Gauvreau [2004], se uma variável explicativa x_i é dicotômica, seu coeficiente β_i no modelo logístico (Equação 4.17) tem uma interpretação especial. Nesse caso, o antilogaritmo de $\hat{\beta}_i$, ou seja, o valor de $e^{\hat{\beta}_i}$, é a razão de chances estimadas da resposta para os dois casos possíveis para x_i. E o valor dessa razão pode ser calculado diretamente pelo produto cruzado dos valores na tabela de contingência $2x2$:

$$\hat{RC} = \frac{a_{11}a_{22}}{a_{12}a_{21}} \tag{4.19}$$

Frequentemente, conhece-se outras variáveis explicativas associadas com a mesma resposta. Naturalmente, surge a pergunta se a inclusão de outras variáveis explicativas melhoram o modelo. Tendo em mente o princípio da parsimônia para a seleção de um modelo, deseja-se construir um modelo com o menor número de variáveis possível Santner & Duffy [1989].

Para extender a análise univariada, o modelo multivariado pode ser implementado por um procedimendo de seleção passo-a-passo para frente ou para trás (*forward/-backward stepwise procedure*). Neste, as variáveis são removidas passo-a-passo do modelo, naquele, elas são adicionadas passo-a-passo.

O método de seleção passo-a-passo mais usual é o método "para a frente". E é este o método que escolhemos para usar neste trabalho e que passamos a discutir detalhadamente nesta seção. Métodos de seleção automática passo-a-passo já receberam várias críticas na literatura. Mas ainda são usados como ponto de partida para a construção de bons modelos, particularmente quando o número de variáveis envolvidas é muito grande ou quando o tempo para análise é limitado.

No *procedimento de seleção de variáveis passo-a-passo para frente*, inicia-se calculando o *p-value* de cada variável explicativa conforme o modelo univariado explicado na seção anterior. Ou seja, começamos ajustando o modelo

$$ln[\frac{\hat{p}}{1-\hat{p}}] = \hat{\beta}_0 + \hat{\beta}_k x_k \tag{4.20}$$

para cada variável explicativa x_k. Daí, seleciona-se a variável mais significativa (com o

menor *p-value*) para adicioná-la ao modelo. No próximo passo, ajusta-se o modelo

$$ln[\frac{\hat{p}}{1-\hat{p}}] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_k x_k \tag{4.21}$$

onde x_1 é a primeira variável explicativa selecionada por ser a que apresentou maior significância (menor *p-value*) e x_k designa cada uma das outras variáveis explicativas.

Recalcula-se, então, o *p-value* para as variáveis remanescentes e, se houver algum abaixo do valor de corte assumido, seleciona-se a variável que apresentou o menor *p-value* para ser adicionada ao modelo. Esses passos são repetidos até que as variáveis restantes não se mostrem significativas no senso aplicado, i.e., quando o teste de significância empregado resulte em valores de *p-value* acima do valor de corte.

O procedimento de seleção de variáveis passo-a-passo para frente tem uma característica que pode ser bastante prejudicial. Nesse método, quando uma variável é selecionada para fazer parte do modelo, ela não é mais retirada. Entretanto, pode ocorrer facilmente que uma dada variável já selecionada torne-se supérflua devido a interrelações com outras variáveis adicionadas posteriormente ao modelo. Para minimizar esse efeito, em geral, aplica-se um método ligeiramente diferente da versão "*seleção* para a frente" que é usualmente chamado de "procedimento de *regressão* passo-a-passo para frente".

O procedimento de regressão modifica o de seleção da seguinte forma: cada vez que uma nova variável é adicionada ao modelo, a significância de cada variável é recalculada e aquela que apresenta o maior valor de *p-value* é removida do modelo caso o valor do *p-value* seja maior do que um valor de corte superior assumido (por exemplo, 0,10). O modelo é então re-ajustado (tem seus coeficientes recalculados) antes do procedimento seguir para o próximo passo de seleção. O procedimento de regressão passo-a-passo para frente é finalizado quando não houver mais variáveis para serem selecionadas ou removidas segundo os critérios adotados.

Portanto, o modelo logístico univariado é extendido para um modelo multivariado ajustando-se um modelo representado pela Equação 4.22.

$$ln[\frac{\hat{p}}{1-\hat{p}}] = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i x_i, \tag{4.22}$$

onde n é o número de variáveis explicativas selecionadas; x_i designa cada uma dessas variáveis e; β_i seus respectivos coeficientes conforme determinados na etapa final do processo.

4.7 Construção do modelo probabilístico para a predição da drogabilidade de um alvo proteico humano

A Figura 4.16 ilustra o fluxograma para a construção do modelo. A partir da mesma matriz $A_{2700x1541}$ usada para construir o modelo vetorial, a regressão logística foi utilizada para selecionar os termos de InterPro relevantes para o cálculo da probabilidade de uma dada proteína ser drogável. Observa-se que a variável resposta (drogável ou não-drogável) é dicotômica. Particularmente, todas as variáveis explicativas são também dicotômicas tornando ainda mais atrativo o emprego desta metodologia, considerando a interpretação especial que os coeficientes passam a ter: representam a razão de chances estimada da resposta para os dois casos possíveis da respectiva variável Pagano & Gauvreau [2004].

O método requer um grupo com alvos e não-alvos. Foram extraídos 384 dos 1541 alvos no conjunto para compor o grupo positivo, restando 1157 alvos conhecidos para a construção do modelo. O grupo negativo foi construído, desta vez, utilizando a métrica de similaridade dada pelo modelo vetorial que desenvolvemos. Para selecionar os membros do grupo controle, partiu-se do grupo de todas as 29580 proteínas humanas não classificadas como drogáveis em qualquer das três bases de dados consultadas (TTD, DrugBank e KEGG-DRUG) e que compartilham ao menos um termo do InterPro com algum alvo no conjunto dos 1541 da matriz A. Cada uma dessas proteínas foi projetada no espaço vetorial reduzido (Equação 4.4) e teve calculada a dissimilaridade com relação a cada um dos 1157 alvos conhecidos (Equação 4.12). Daí, foram selecionadas as proteínas que apresentaram os maiores valores de dissimilaridade mínima em relação ao conjunto inicial de alvos. Foi estabelecido um corte usual de percentil 75 sobre os 29580 candidatos. Devido à ocorrência de muitos valores iguais, o corte resultou em um grupo controle com um pouco mais de $0,25$ do conjunto inicial. No total, 7830.

O próximo passo foi aplicar um estudo de caso-controle [Schlesselman, 1982] na construção de um modelo para predizer se uma dada proteína humana é drogável. Das 7830 proteínas, estabelecemos um valor de corte para extrair no mínimo quatro vezes o tamanho do grupo caso (4 x 1157). Isso resultou em um grupo controle com 5821 proteínas. O restante (2009) foi deixado para o grupo de validação. Assim, o tamanho final da amostra foi de 6978 ($5821 + 1157$).

Como desejávamos construir um modelo contendo descritores (ou atributos) mais facilmente empregados e interpretados, optamos por usar apenas os termos do InterPro dos tipos família (F), domínio (D) e região (G). Isso reduziu o número de descritores de

Tabela 4.3. Análise univariada para o InterPro IPR001828 – sua presença aumenta a chance da proteína ser drogável

InterPro IPR001828	Tamanho da amostra	Número de alvos	Percentual de alvos	p-value
Presença	21	19	90%	$< 0,001$
Ausência	6957	1138	16%	
Total	6978	1157	17%	

Tabela 4.4. Análise univariada para o InterPro IPR016175 – sua presença reduz a chance da proteína ser drogável

InterPro IPR016175	Tamanho da amostra	Número de alvos	Percentual de alvos	p-value
Presença	230	0	0%	$< 0,001$
Ausência	6748	1157	17%	
Total	6978	1157	17%	

2700 para 2390. Os 310 descartados não se referem diretamente a um dado InterPro, designam a ocorrência ou não de uma dada anotaçao de sítio dentro do intervalo de sequência atribuído a alguma anotação do tipo F, D ou G para a mesma proteína.

O número de descritores ainda era muito grande. Considerando a redução de posto de 2700 para 320 que já havíamos determinado com um bom resultado final, foi inferido que deveria ser possível reduzir bastante o número de descritores necessários para reproduzir um bom modelo.

O número de variáveis (2390) era ainda maior do que o número de equações (1157). Foi aplicado primeiramente a regressão logística univariada usando o corte de *p-value* igual a $0,05$ pelo *teste exato de Fisher* a cada uma das 2390 variáveis. Isso reduziu o número de variáveis para 587.

Entre os termos do InterPro mantidos no modelo, há aqueles que contribuem para aumentar a probabilidade da proteína ser drogável e aqueles que contribuem para diminuir esta probabilidade. Por exemplo, o termo do InterPro identificado por IPR001828 aumenta a chance de uma dada proteína ser drogável (Tabela 4.3). Enquanto o termo do InterPro identificado por IPR016175 reduz esta mesma chance (Tabela 4.4).

Os 587 descritores mantidos no conjunto após a análise univariada foram analisados por regressão logística multivariada empregando-se o método de regressão passo-a-passo para frente. Esse procedimento final determinou quais descritores deveriam ser mantidos no modelo além de fornecer o coeficiente que indica a contribuição de cada um no cálculo da probabilidade do alvo ser drogável. O resultado é representado por

uma fórmula para calcular a probabilidade de uma dada proteína ser drogável.

Faltava determinar o melhor valor de corte para a probabilidade calculada pela fórmula obtida de modo a maximizar a sensibilidade e a especificidade em um processo de classificação de proteínas como drogáveis ou não-drogáveis. Isso foi feito por uma análise de curva ROC.

A fórmula obtida para calcular a probabilidade, bem como o estudo para determinar o melhor valor de corte são apresentados na seção 5.1, no Capítulo de Resultados e Discussão.

4.8 Outros modelos vetoriais: integração de dados químicos, farmacológicos e biológicos

Em algoritmos de buscas na internet, não se faz restrições iniciais quanto aos termos usados. São palavras, frases, ou apenas sequências de caracteres tratados todos da mesma forma. Similarmente, foi construído um novo modelo de espaço vetorial para integrar dados de diferentes tipos, mas relacionados com os fármacos e seus alvos proteicos. Neste modelo, os alvos usados são humanos e não-humanos. Para garantir maior confiabilidade dos dados, foram carregadas somente informações relativas aos fármacos aprovados pela *FDA* e catalogados no DrugBank. Os dados extraídos foram: o SMILES canônico da estrutura química do fármaco; os alvos proteicos associados a cada fármaco e; as categorias terapêuticas (ou indicações) também associadas a cada fármaco.

Diferentes modelos foram construídos para comparar como as relações entre os fármacos são transformadas usando cada conjunto de descritores. No primeiro modelo, os fármacos são representados apenas por descritores relacionados com sua estrutura química. Esses descritores foram obtidos convertendo o SMILES a um *fingerprint* usando o programa OpenBabel [O'Boyle et al., 2011]. Os *fingerprints* gerados pelo OpenBabel contêm caracteres hexadecimais. Eles foram convertidos para correspondentes binários de modo que os descritores relacionados com a estrutura química do fármaco tenham o mesmo formato que os descritores relacionados com as propriedades biológicas de seus alvos.

Em outra abordagem, descritores biológicos foram acrescentados. São descritores binários que indicam a ocorrência ou não de determinadas anotações do InterPro do tipo F, D ou G para qualquer dos alvos associados ao fármaco.

Outro conjunto de descritores relacionados diretamente aos fármacos representam as diferentes categorias ou indicaões terapêuticas associadas. Estes também são

descritores binários.

Diferentes matrizes foram construídas representando diferentes combinações dos tipos de descritores. Cada uma foi usada na construção de um espaço vetorial de posto reduzido aplicando-se a mesma metodologia empregada na construção do modelo de espaço vetorial para alvos drogáveis.

Para avaliar a consistência dos modelos, foram feitos estudos de correlação entre a métrica de dissimilaridade no espaço vetorial reduzido e a métrica de Tanimoto fornecida pelo DrugBank (Figura 4.18).

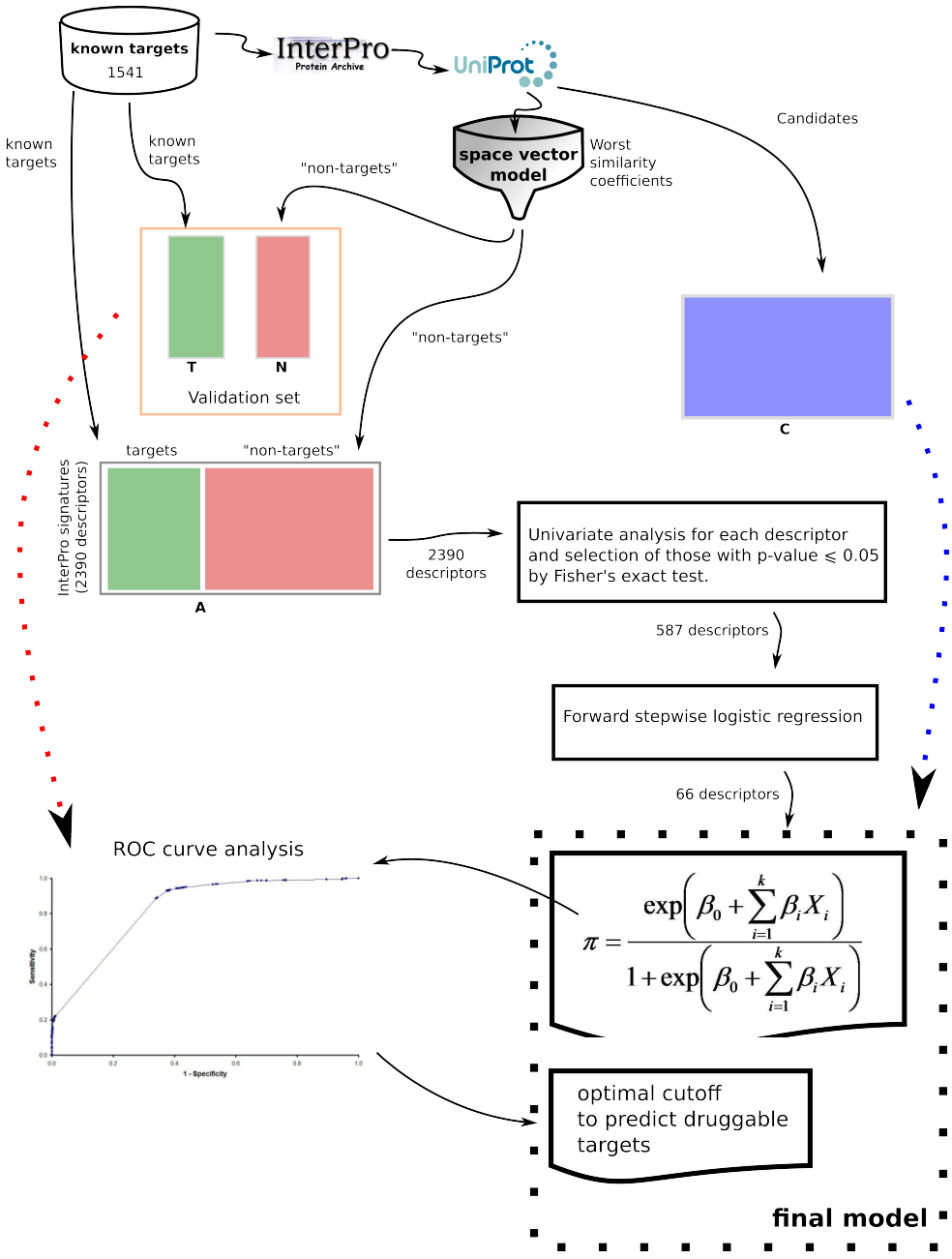


Figura 4.16. Fluxograma para a construção do modelo probabilístico baseado em regressão logística. Parte dos alvos conhecidos foram reservados para validação e o restante foi usado na matriz inicial. Os termos do InterPro presentes no conjunto de alvos foram usados para selecionar outras sequências do UniProtKB. Destas, as que apresentaram os piores índices de similaridade, segundo o modelo vetorial, com relação ao conjunto dos alvos, foram classificadas como "não-alvos". Este critério para seleção dos casos negativos foi aplicado até somar quatro vezes o tamanho do grupo positivo da matriz inicial. A partir daí, empregou-se as regressões univariada e multivariada para selecionar os atributos relevantes e definir uma fórmula que fornece a probabilidade π de uma dada protína ser drogável. Por fim, o conjunto de validação foi usado na determinação do valor de corte para π que otimiza a sensibilidade e a especificidade.

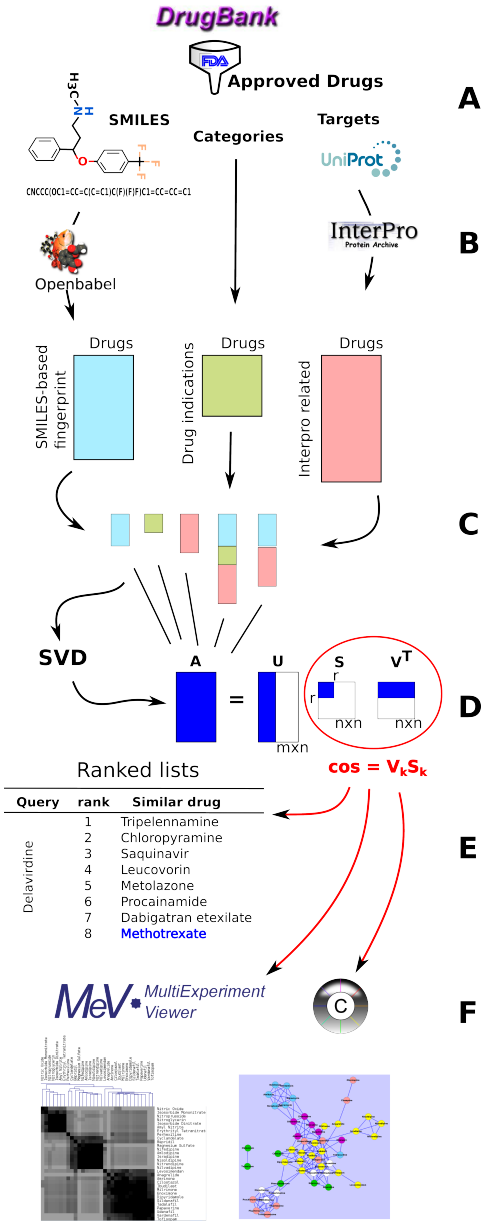


Figura 4.17. Fluxograma para outros modelos usando a integração de dados químicos, biológicos e farmacológicos. Foram extraídos do DrugBank: os SMILES canônicos; as categorias terapêuticas e os alvos proteicos associados a fármacos aprovados pela *FDA* **(A)**; depois, os SMILES foram convertidos a *fingerprints* usando o Openbabel e cada alvo foi descrito por suas respectivas anotações no InterPro **(B)**; foram construídas três matrizes representando os espaços químico, farmacológico e biológico (cada uma tendo os fármacos representados em suas colunas) e diferentes combinações dessas matrizes foram construídas justapondo-as verticalmente **(C)**; cada nova matriz foi submetida à fatoração e redução de ruído usando SVD **(D)**; para cada abordagem, foram construídas listas ordenadas para cada fármaco segundo a métrica de similaridade/dissimilaridade computada a partir do coseno dos vetores no espaço reduzido **(E)**; foram também construídos mapas de calor e redes de relacionamento a partir de algorimos de agrupamento hierárquico e baseado em força usando o MeV e o Cytoscape **(F)**.

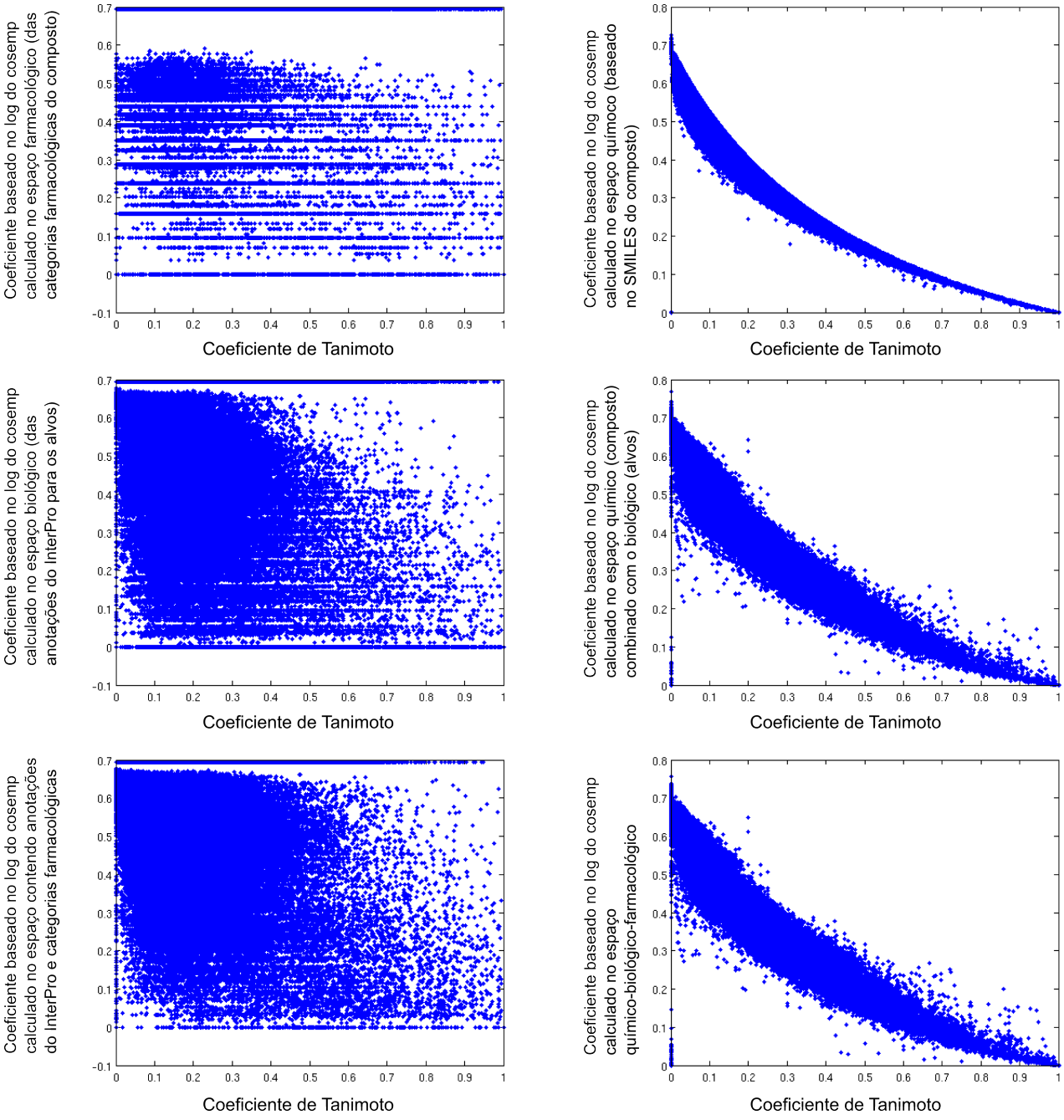


Figura 4.18. Estudos de correlação. Relação entre o coeficiente de Tanimoto e o coeficiente de dissimilaridade obtido por SVD e redução de posto usando diferentes combinações de descritores.

Capítulo 5

Resultados e discussão

Nesta seção são discutidas

- a consistência e a complementaridade do modelo em relação à métrica de similaridade entre sequências,
- a sua representatividade em relação ao subconjunto de alvos drogáveis conhecidos,
- a sua capacidade na recuperação de informação implícita e
- a sua escalabilidade em relação a alvos projetados (posteriormente) no espaço vetorial.

A consistência do modelo foi avaliada, inicialmente, por testes de correlação com o alinhamento de sequências (dado pelo BLAST) e por análise qualitativa considerando-se as propriedades biológicas conhecidas dos alvos iniciais e dos novos alvos inseridos no espaço. A distribuição dos dados foi analisada e visualizada usando diferentes algoritmos de agrupamento: hierárquico, fatoração de matriz não-negativa, método baseado em força além de algoritmo de otimização, visualização em espaço tridimensional, mapas de calor e rede de interações. Estudos de caso foram discutidos citando-se referências na literatura que corroboram alguns resultados específicos. Para avaliar se o espaço definido pelo modelo é representativo para a distinção entre alvos drogáveis e alvos não-drogáveis ou, ao menos, dos dificilmente drogáveis, foi realizado um estudo de curva ROC (*receiver operating characteristic*) relacionando alvos drogáveis e não-drogáveis.) A escalabilidade do modelo foi avaliada projetando-se outros alvos conhecidos, não presentes no espaço original, e avaliou-se a consistência do posicionamento atribuído a esses novos objetos em relação aos biologicamente similares. Um

estudo de caso envolvendo um potencial alvo terapêutico (KMO_HUMAN), corroborado por publicações recentes, é discutido para ilustrar a capacidade de recuperação de relacionamentos implícitos.

Complementariamente, proteínas de *P. falciparum* e de *T. gondii* foram projetadas e um estudo de caso discutido mostrando a possibilidade de se detectar oportunidades de utilização de fármacos, planejados para agir sobre alvos humanos, no combate aos parasitas. Além disso, foram selecionadas as anotações do InterPro mais importantes para a caracterização de alvos humanos drogáveis. Ao final, a metodologia foi empregada para descrever os medicamentos de modo a permitir a avaliação de associações implícitas que podem sugerir novas oportunidades de reposicionamento de fármacos e alguns resultados preliminares são avaliados comparativamente com dados de outras publicações.

Além da comparação quantitativa, comparamos qualitativamente os resultados fornecidos pelas duas métricas considerando as anotações biológicas funcionais conhecidas para as proteínas pesquisadas.

A Figura 5.1 apresenta o gráfico de correlação entre a métrica de dissimilaridade calculada no espaço vetorial do modelo e o alinhamento par-a-par dado pelo BLAST (*bitscore*). A curva exibe correlação negativa conforme esperado. O *bitscore* é uma medida de similaridade enquanto a métrica de "distância" dada por $d_{ij} = -ln((1+cos_{ij})/2)$ é uma medida de dissimilaridade. Além disso, a curva apresenta uma forma que indica uma relação exponencial. Esta última característica parece uma consequência da redução de posto obtida por decomposição por valores singulares. Esse procedimento evidencia os extremos acentuando as similaridades e dissimilaridades entre os pares de membros do conjunto.

Para obter os dados do alinhamento de sequências, foi construído uma base local do BLAST [Altschul et al., 1990] contendo todas as sequências dos alvos conhecidos (extraídas do *UniProt*). Esta base local foi construída usando o *formatdb* que faz parte do pacote ncbi-blast para Linux.

Após construir a base local, o *blastall* foi utilizado para cada sequência. Os valores de *bitscore* foram normalizados em cada caso dividindo-se pelo valor obtido para o alinhamento da sequência com ela mesma.

A forma da curva de correlação sugere uma melhor separação dos objetos e, talvez daí, a melhor performance do algoritmo de agrupamento hierárquico no espaço vetorial reduzido do que no espaço do alinhamento de sequências (Figura 5.2). Agrupamentos funcionais são facilmente identificados no espaço vetorial e dificilmente detectados no espaço definido pelos valores de *bitscore*. Comportamento semelhante é observado ao serem comparados os pontos no espaço tridimensional resultante da projeção pelo

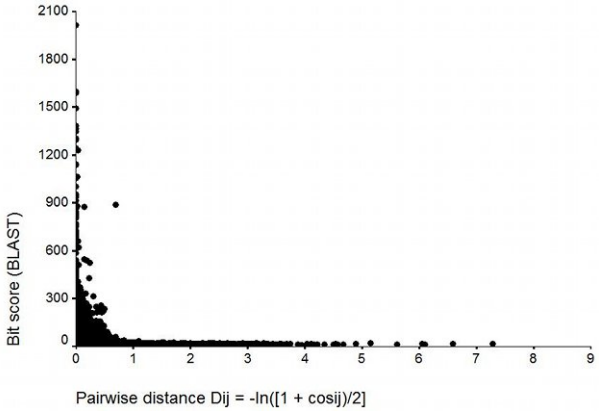


Figura 5.1. Correlação entre o *bitscore* do BLAST e o coeficiente de dissimilaridade obtido pelo modelo. A correlação negativa é explicada porque quanto maior o *bitscore*, menor a dissimilaridade medida pela "distância". O comportamento exponencial é explicado pela característica da decomposição por valores singulares evidenciar os extremos.

algoritmo de otimização (Figura 5.3).

Os mapas de calor das Figuras 5.4 e 5.5 foram gerados para um subconjunto contendo menos alvos. A representação dos alvos é a mesma quando usados todos os alvos do conjunto (ou seja, não foi realizada nova fatoração e redução de posto). Os resultados obtidos com a similaridade vetorial são bastante compatíveis com aqueles obtidos com o alinhamento de sequências. Mas nossa similaridade vetorial apresenta a vantagem de ser independente do tamanho da sequência e de produzir dados que favorecem a eficiência do algoritmo de agrupamento. A proteína PE2R3_HUMAN da família GPCR foi deixada "órfã" pelo algoritmo de agrupamento sobre a matriz de similaridade par-a-par, mas foi agrupada com o restante das GPCRs no conjunto ao ser aplicado sobre a matriz de similardade vetorial. Os 42 alvos usados foram selecionados de uma amostra usada por Keiser et al. [2007] para servir como comparação.

O MeV só calcula o coeficiente de correlação cofenética quando é usado o algoritmo de agrupamento por fatoração de matriz não-negativa (NMF). Este algoritmo foi aplicado às matrizes de similaridade vetorial (Figura 5.6) e par-a-par (Figura 5.7). O melhor valor para o coeficiente de correlação cofenética ocorreu para 12 processos (resultando 12 *clusters*) no caso vetorial e 16 no caso da similaridade par-a-par. Observa-se ainda que os 12 agrupamentos formados com o algoritmo NMF sobre a matriz de similaridade dada pelo modelo vetorial são inteiramente compatíveis com o dendrograma fornecido pelo algoritmo hierárquico.

As Figuras 5.8, 5.9 e 5.10 ilustram a capacidade do algoritmo de agrupamento hierárquico, quando aplicado sobre a matriz de similaridade vetorial, detectar estruturas

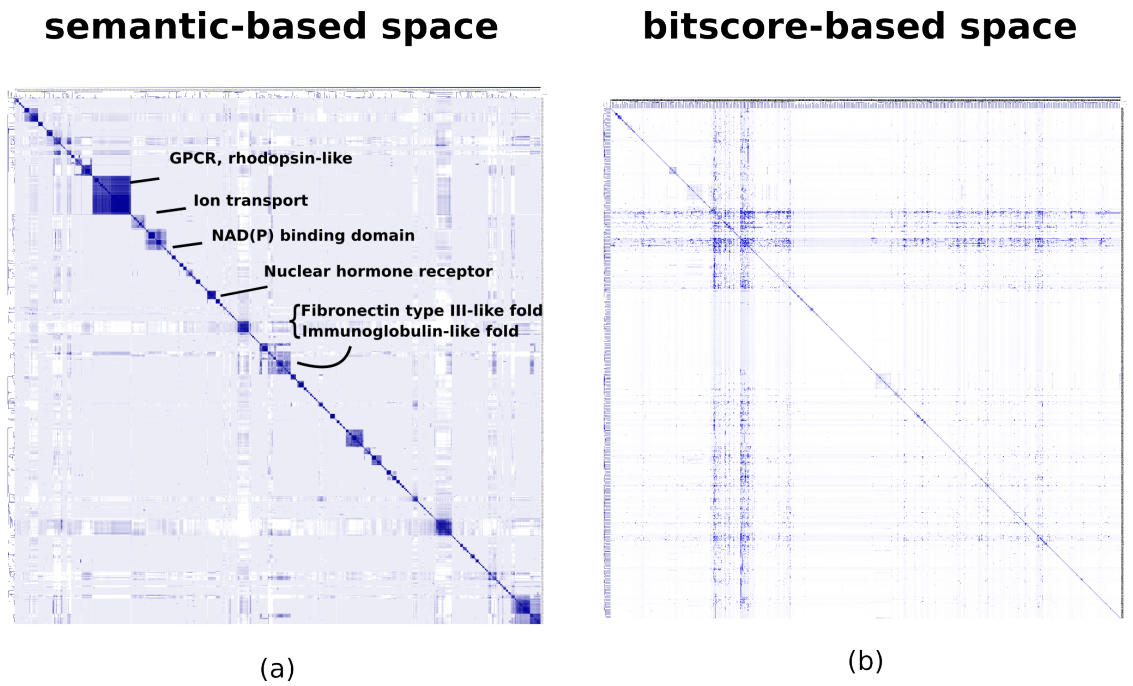
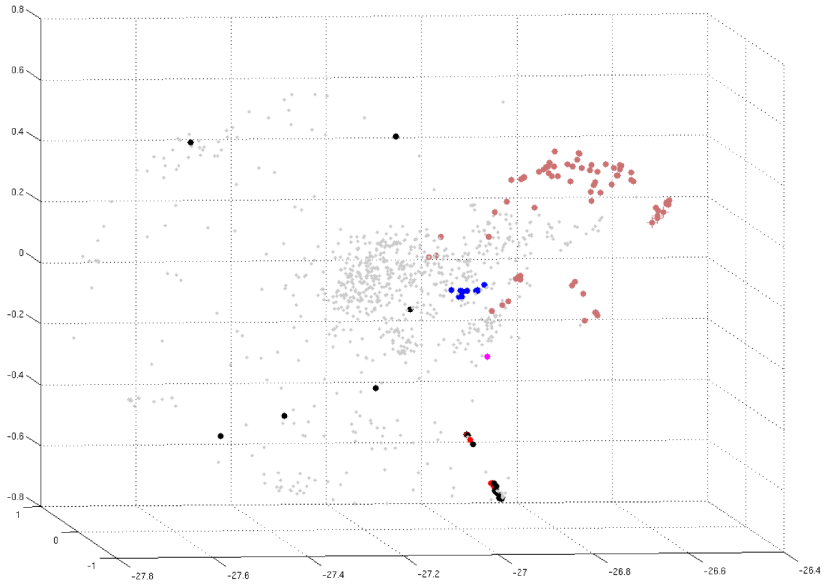


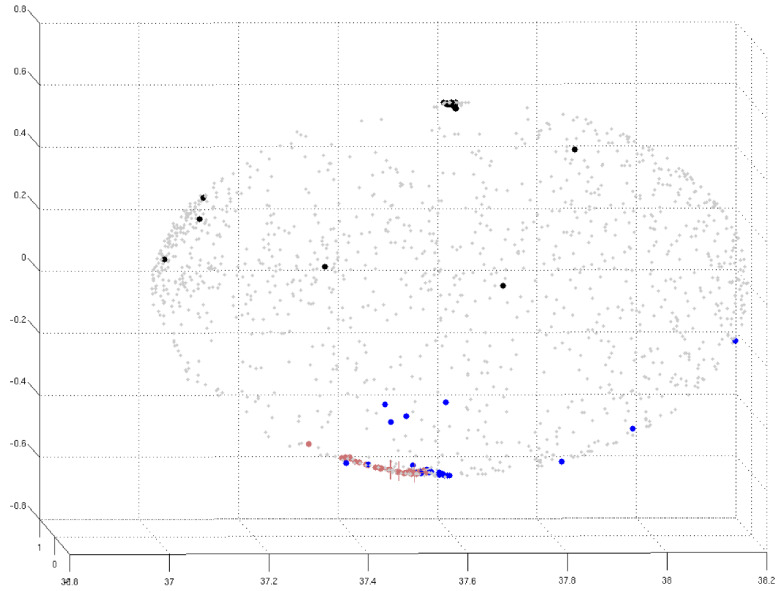
Figura 5.2. Visualização do resultado do algoritmo de agrupamento hierárquico aplicado sobre a matriz de similaridade do modelo vetorial (a) e sobre a matriz de similaridade de sequências (b). No espaço vetorial, mesmo com todos os 1541 alvos presentes, é possível identificar vários agrupamentos correspondentes às propriedades biológicas desses alvos. Indentifica-se, por exemplo: o conjunto das GPCRs (o maior grupo); as proteínas com domínios de receptores nucleares (mostrado em detalhe na Figura 5.8); transportadores de íons; proteínas com domínios de ligação NAD(P) (mostrado em detalhe na Figura 5.9); proteínas com os domínios *Fibronectin type III-like fold* e *Immunoglobulin-like fold* – domínios similares estruturalmente e usualmente co-exitentes (mostrado em detalhe na Figura 5.10). A visualização do conjunto completo também reagrupado pelo mesmo algoritmo hierárquico, mas sobre a matriz de *bitscores*, deixa claro a maior dificuldade de se identificar agrupamentos com esta métrica. Foram utilizados o algoritmo de agrupamento hierárquico e o recurso de visualização por mapas de calor implementados no *Multi environment tool* (MeV) [Howe et al., 2010]. Foi necessário converter a métrica de dissimilaridade em uma métrica de similaridade para utilizar o MeV.

hierárquicas compatíveis com a hierarquia das anotações biológicas dos alvos.

Na Figura 5.11, vê-se novamente os dados no espaço tridimensional, desta vez com algumas regiões vistas em destaque e aumentadas. Observa-se a coerência dessa metodologia em relação à estrutura hierárquica das anotações biológicas e em relação aos resultados fornecidos pelos algoritmos de agrupamento. Conclui-se que o modelo vetorial favoreceu o desempenho de diferentes técnicas de agrupamento e visualização fornecendo resultados coerentes com a classificação hierárquica das assinaturas de funções biológicas catalogadas no InterPro.



(a)



(b)

Figura 5.3. Representação dos dados em espaço tridimensional. Utilizou-se o algoritmo de otimização de Marcolino et al. [2010] aplicado sobre a matriz de dissimilaridade fornecida pelo modelo vetorial **(a)** e sobre a matriz de *bitscore* **(b)**. Os pontos pretos representam alvos da família das GPCR, os pontos azuis representam receptores nucleares e os vermelhos representam integrinas. Como no caso do algoritmo hierárquico, nota-se um agrupamento mais claro no modelo vetorial.

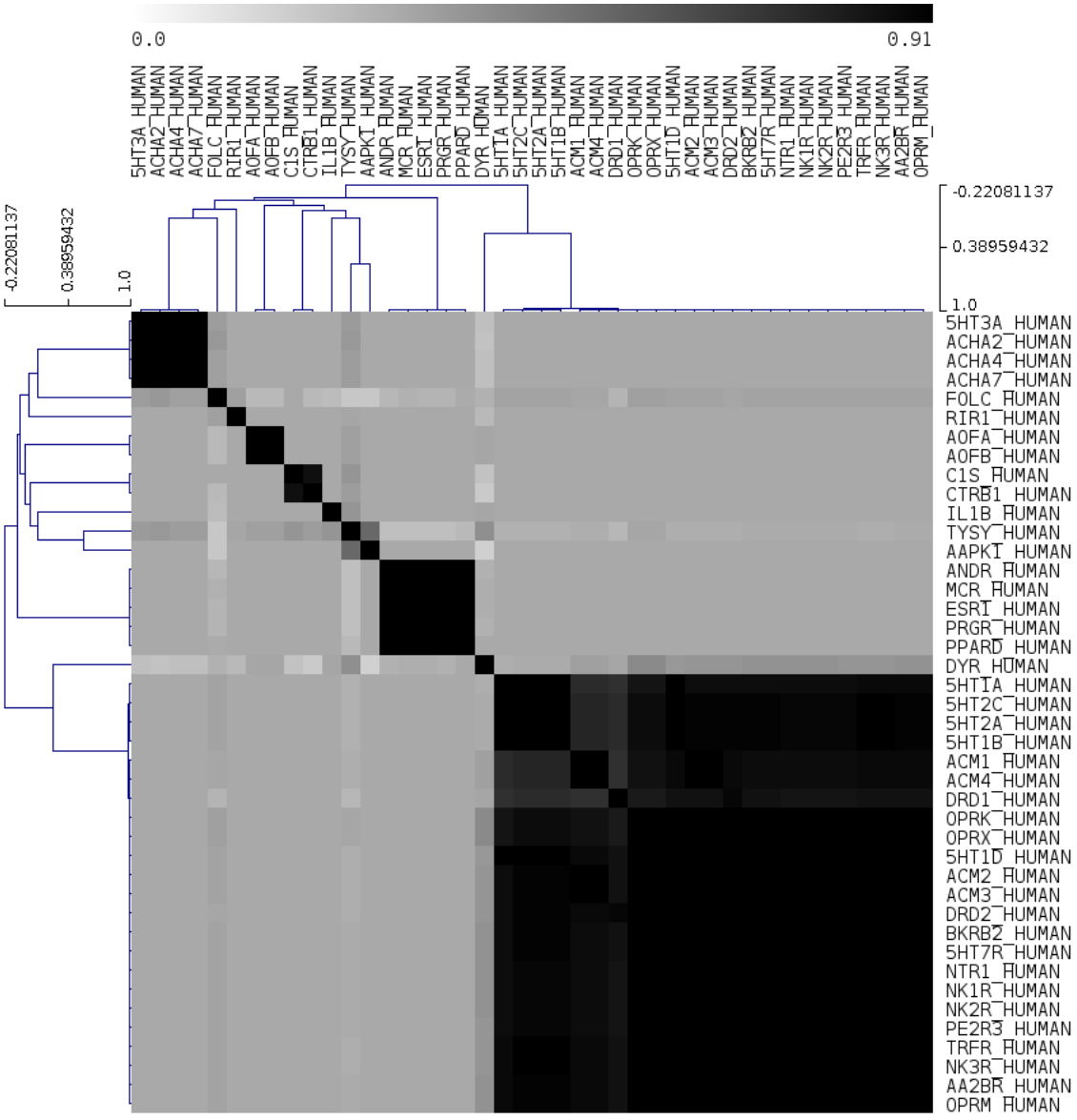


Figura 5.4. Resultado do algoritmo de agrupamento hierárquico sobre a matriz de similaridade semântica para uma amostra com 42 alvos selecionados a partir de [Keiser et al., 2009].

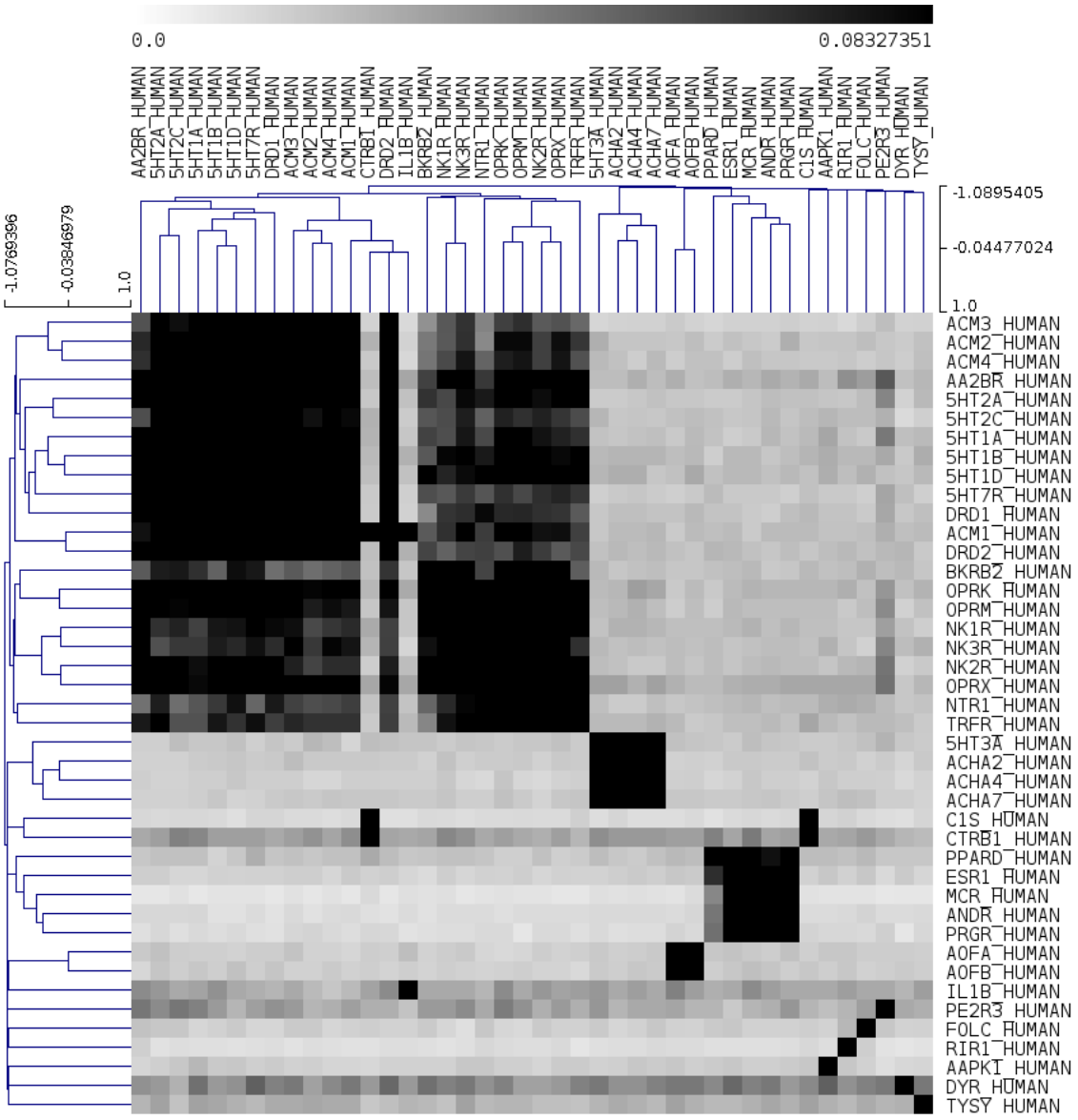


Figura 5.5. Resultado do algoritmo de agrupamento hierárquico sobre a matriz de similaridade par-a-par para a mesma amostra com 42 alvos da Figura 5.4.

```
SVD clusters - best cophenetic correlation = 0.9249313
Number of clusters: 12
Clustering algorithm: Non-negative factorization

Cluster 1:      PPARD PRGR ESR1 ANDR MCR
Cluster 2:      C1S CTRB1
Cluster 3:      AOFB AOFB
Cluster 4:      ACHA7 ACHA2 ACHA4 5HT3A
Cluster 5:      TYSY AAPK1
Cluster 6:      5HT1B 5HT2A 5HT1A 5HT2C
Cluster 7:      FOLC RIR1
Cluster 8:      IL1B
Cluster 9:      DYR
Cluster 10:     DRD1
Cluster 11:     ACM3 ACM1 ACM4 ACM2 DRD2 AA2BR NTR1 OPRM
                OPRK OPRX BKRB2 NK2R NK3R 5HT7R TRFR PE2R3
Cluster 12:     5HT1D
```

Figura 5.6. Resultado do agrupamento por fatoração de matriz não-negativa (NMF) sobre a matriz de similaridade vetorial. O melhor valor para o coeficiente de correlação cofenética ocorreu para 12 processos (resultando em 12 *clusters*). Os 12 agrupamentos indicados são consistentes com aqueles de último nível no dendrograma fornecido pelo algoritmo hierárquico (Figura 5.4).

```
SVD clusters - best cophenetic correlation = 0.9076
Number of clusters: 16
Clustering algorithm: Non-negative factorization

Cluster 1:      NK2R
Cluster 2:      IL1B
Cluster 3:      CTRB1
Cluster 4:      DRD2
Cluster 5:      ACM3 ACM1 ACM4 ACM2
Cluster 6:      DRD1 5HT7R 5HT1D 5HT1B 5HT2A 5HT1A 5HT2C
Cluster 7:      AOFB AOFB
Cluster 8:      NK3R NK1R
Cluster 9:      C1S RIR1
Cluster 10:     AA2BR OPRM OPRK OPRX
Cluster 11:     TRFR NTR1 BKRB2
Cluster 12:     AAPK1
Cluster 13:     PE2R3
Cluster 14:     FOLC ACHA7 ACHA2 ACHA4 5HT3A
Cluster 15:     DYR TYSY
Cluster 16:     ESR1 PRGR ANDR MCR PPARD
```

Figura 5.7. Resultado do agrupamento por fatoração de matriz não-negativa (NMF) sobre a matriz de similaridade sequência. O melhor valor para o coeficiente de correlação cofenética ocorreu para 16 processos (resultando em 16 *clusters*).

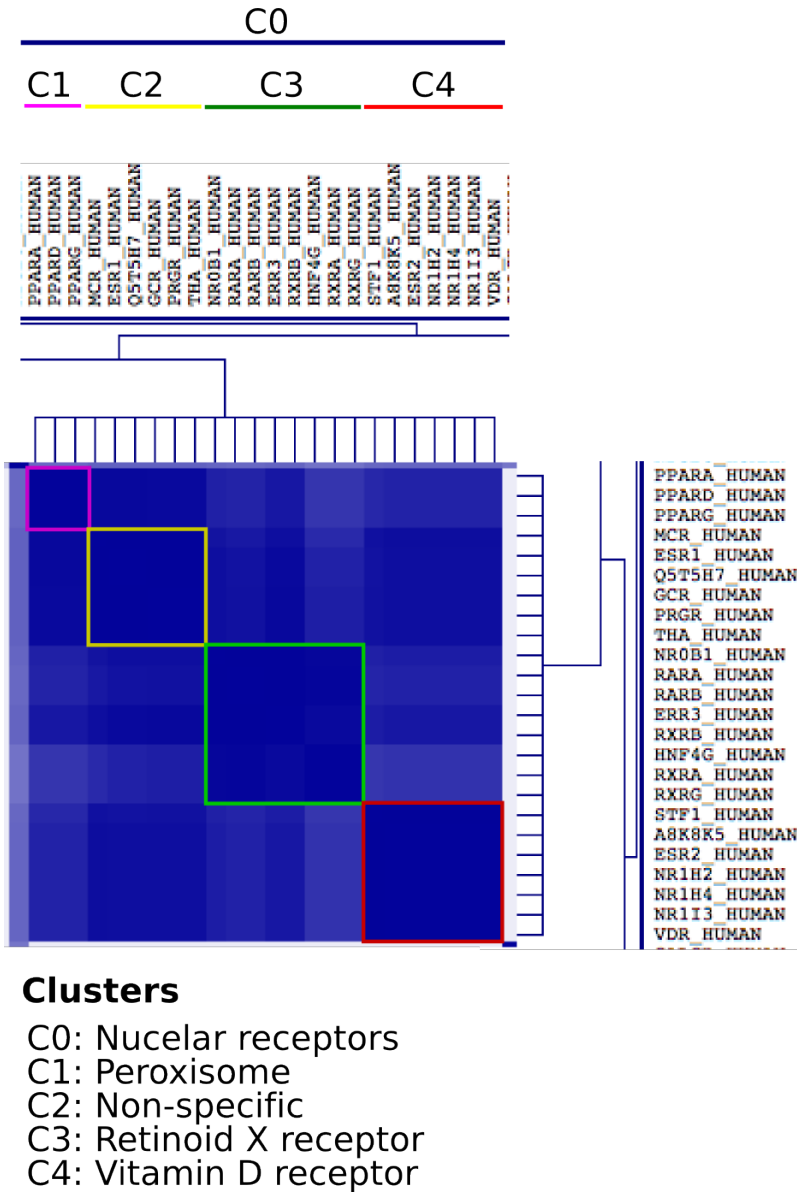
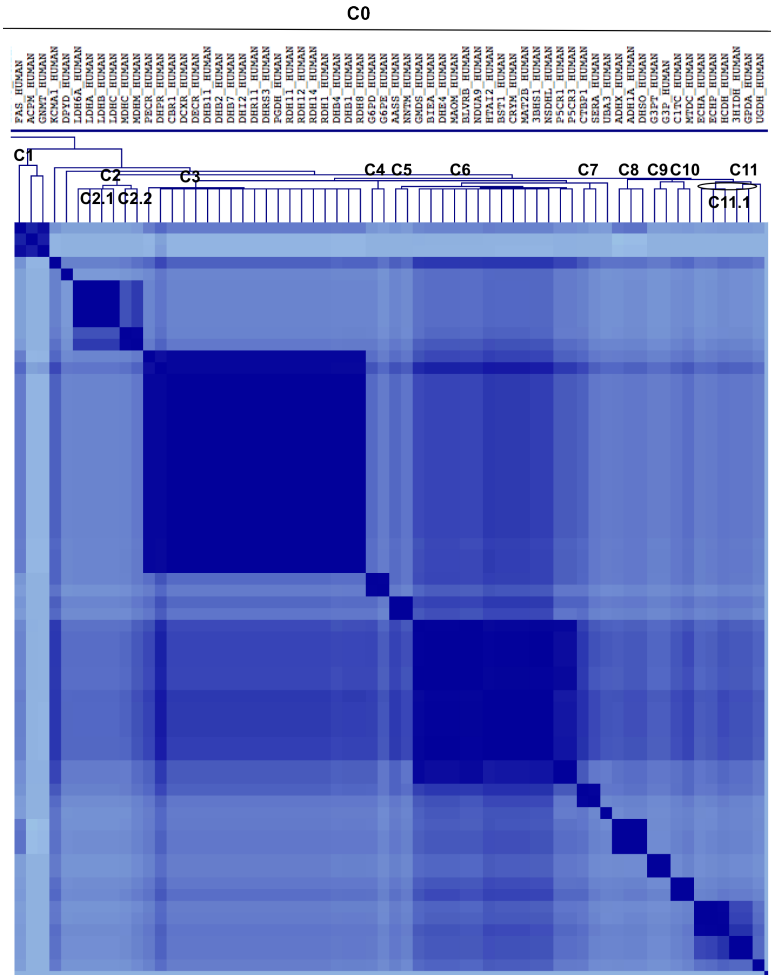


Figura 5.8. Visão expandida da Figura 5.2-a na região que contém os receptores nucleares. Observa-se a consistência com a subclassificação em categorias mais específicas.

Dadas as evidências da consistência do modelo, obtidas a partir da comparação com o alinhamento de sequências e da análise qualitativa da visualização dos dados em relação às anotações funcionais, passa-se a outra pergunta: O modelo consegue indicar alvos drogáveis? Essa pergunta é importante para a indústria farmacêutica que procura otimizar seus investimentos; oportunidades isentas de patentes podem advir das respostas às *queries* submetidas ao modelo.

A resposta a esta pergunta começa com a avaliação tradicional do modelo no que tange ao discernimento entre alvos e não-alvos.



Clusters

C0: NAD-P binding domain
- C1: Acyl carrier protein-like
- C2: Lactate/malate dehydrogenase
  - C2.1: L-lactate dehydrogenase, active site
  - C2.2: Malate dehydrogenase, active site
- C3: Short-chain dehydrogenase/reductase SDR
- C4: Glucose-6-phosphate dehydrogenase
- C5: Alanine dehydrogenase/PNT, N-terminal OR C-terminal
- C6: Various
- C7: D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain
- C8: Alcohol dehydrogenase superfamily, zinc-containing
- C9: Glyceraldehyde 3-phosphate dehydrogenase family
- C10: Tetrahydrofolate dehydrogenase/cyclohydrolase, NAD(P)-binding domain
- C11: 6-phosphogluconate dehydrogenase, C-terminal-like
  - C11.1: Dehydrogenase, multihelical

Figura 5.9. Visão expandida de outra região da Figura 5.2-a relacionada a alvos com domínios de ligação NAD-P. Novamente, fica evidente a consistência com a subclassificação em categorias mais específicas, neste caso, com um grupo com cerca do dobro de subcategorias e alvos em relação ao grupo dos receptores nucleares, além de ser um caso que exibe um terceiro nível de subcategorias.

Para a construção da curva ROC, é necessário conhecer um grupo de alvos conhecidos não presentes originalmente no conjunto, chamado de grupo positivo, e também

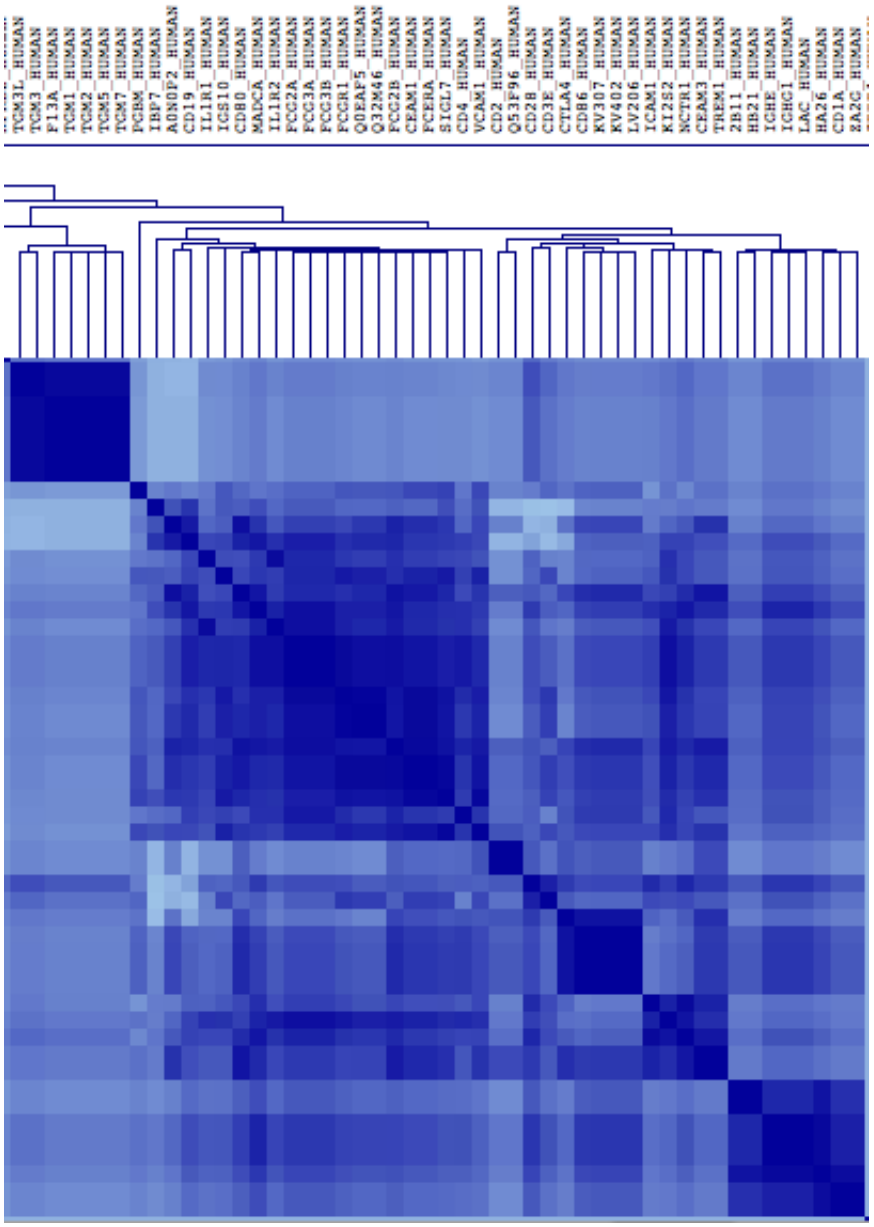


Figura 5.10. Visão expandida de outra região da Figura 5.2-a. Pode-se ver o agrupamento apropriado de alvos com domínios dos tipos *Fibronectin type III-like fold* e *Immunoglobulin-like fold*. Sabe-se que esses dois tipos de domínios ocorrem concomitantemente entre as proteínas conhecidas Leahy [1997].

um grupo de "não-alvos" formado por proteínas que sabe-se (ou ao menos estima-se) não serem drogáveis, chamado de grupo negativo. Para o grupo positivo usamos os 365 alvos reservados desde o início do processo. Quanto ao grupo negativo, não existe um conjunto bastante abrangente e confiável. Cheng et al. [2007] utilizam um grupo positivo com 23 alvos conhecidos contra um grupo negativo com apenas quatro alvos reconhecidamente difíceis de serem modulados pela ação de um fármaco. Desses qua-

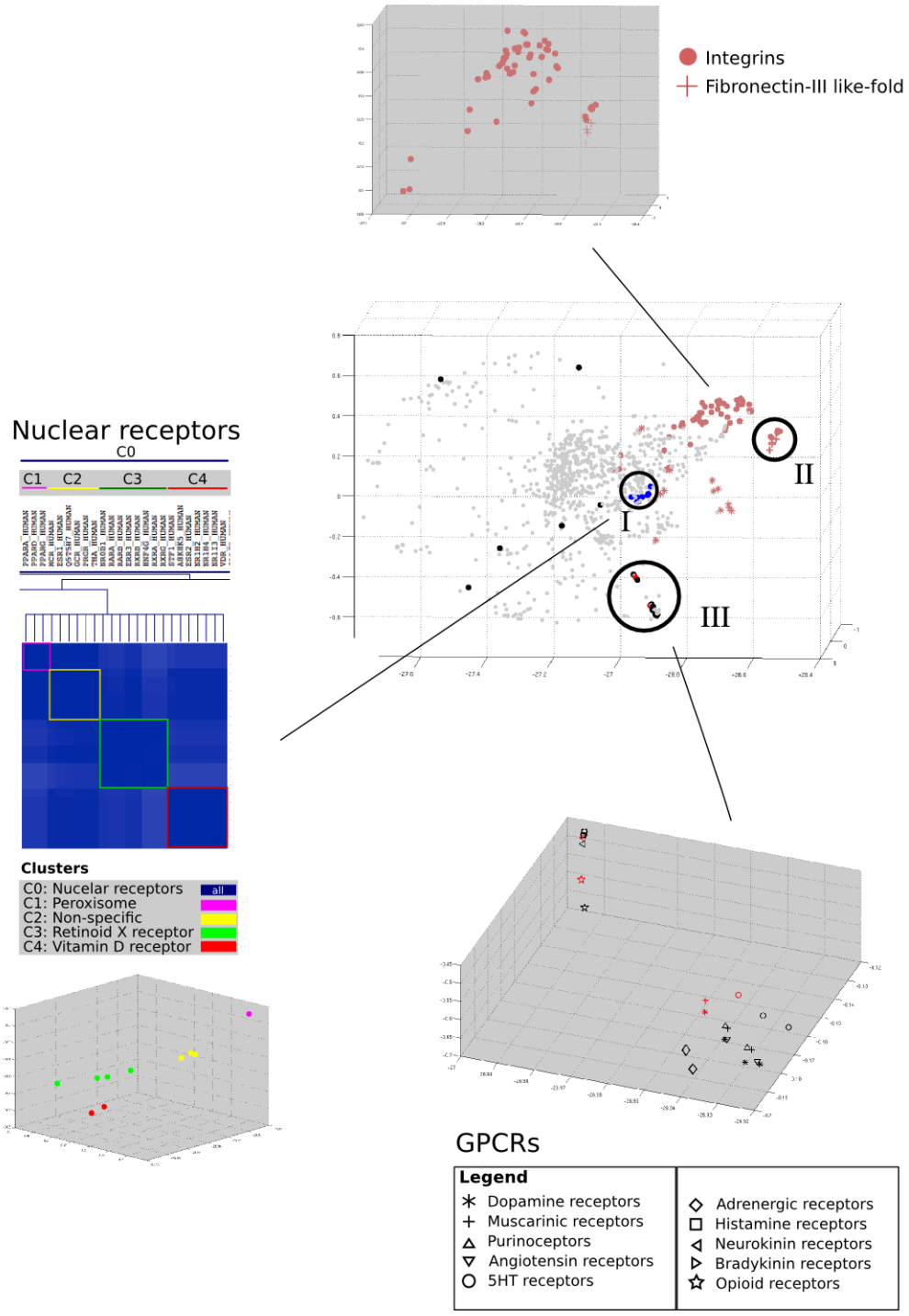


Figura 5.11. Algumas subcategorias visualizadas no espaço tridimensional obtido pelo algoritmo de otimização de Marcolino et al. [2010]. A região I contém os receptores nucleares. O mapa de calor apresentado à esquerda é uma miniatura da Figura 5.8 – foi colocado aqui para permitir a comparação das duas formas de visualização para o caso dos receptores nucleares. A região II ilustra a correlação entre *Fibronectin type III-like fold* e *Immunoglobulin-like fold*. As GPRRs aparecem agrupadas na região III. Os elementos destacados em vermelho nessa região correspondem aos alvos projetados posteriormente no espaço vetorial de posto reduzido $k = 320$.

tro alvos "não-drogáveis", três são proteínas humanas e o outro é uma enzima do HIV. A abordagem de Cheng et al. [2007] é baseada em estudos da estrutura cristalizada dos alvos. Portanto, ela se aplica a um conjunto bem menor do que o proposto neste trabalho. Mesmo para uma abordagem baseada na estrutura, um conjunto de apenas quatro casos negativos é preocupante. Para piorar a situação, todos os três alvos humanos considerados não-drogáveis, assim foram considerados por serem desconhecidas na época (novembro de 2005) interações com pequenos compostos, mas atualmente já estão associados a algum fármaco aprovado na relação do DrugBank. Como foi adotado neste trabalho o critério de considerar drogáveis os alvos relacionados no DrugBank para o conjunto de fármacos aprovados, nenhum dos três alvos humanos rotulados como não-drogáveis por Cheng et al. [2007] poderiam fazer parte de nosso grupo de não-alvos.

Diante da inexistência de um conjunto abrangente e confiável de casos negativos foi construído um grupo negativo da seguinte forma: tomamos inicialmente todas as proteínas humanas não classificadas como drogáveis em qualquer das três bases de dados consultadas (TTD, DrugBank e KEGG-DRUG) e que compartilham ao menos um termo do InterPro com algum alvo no conjunto dos 1541 presentes na matriz A. Esse procedimento inicial excluiu os casos que resultariam obviamente em similaridade zero com relação a todos os alvos mas nos deixou ainda com um grande número de candidatos (29580). Entre esses, foram selecionados aqueles que apresentaram valor mínimo de *e-value* não menor do que 10 quando comparados com os alvos inicialmente presentes no modelo. Isso resultou em um grupo de não-alvos com 242 sequências.

Assim, a amostra usada para a análise da curva ROC é constituída por 607 proteínas, sendo 365 reconhecidamente drogáveis e as outras 242 considereradas não-drogáveis por apresentarem os piores alinhamentos com relação a cada proteína no conjunto inicial de alvos conhecidos. Cada uma das 607 proteínas da amostra foi projetada no espaço vetorial. Depois disso, calculou-se o valor do coeficiente de dissimilaridade para cada uma em relação a cada alvo do conjunto inicial. Para o próximo passo, tomou-se o menor valor do coeficiente obtido para cada proteína da amostra.

A análise de *boxplot* mostra que alvos drogáveis conhecidos apresentaram valores mínimos menores do que as proteínas não-alvos (Figura 5.12). Apesar de alguns casos discrepantes (*outliers*) e extremos nos dois grupos, a comparação do *boxplot* para cada caso indica que o coeficiente de dissimilaridade desenvolvido com decomposição por valores singulares pode ser usado em uma análise discriminante para a identificação de alvos drogáveis.

A análise da curva ROC apresentou uma área sob a curva (AUC) de $0,92$. O intervalo de confiança assintótico de 95% é $[0,897; 0,942]$ mostrando uma boa qualidade

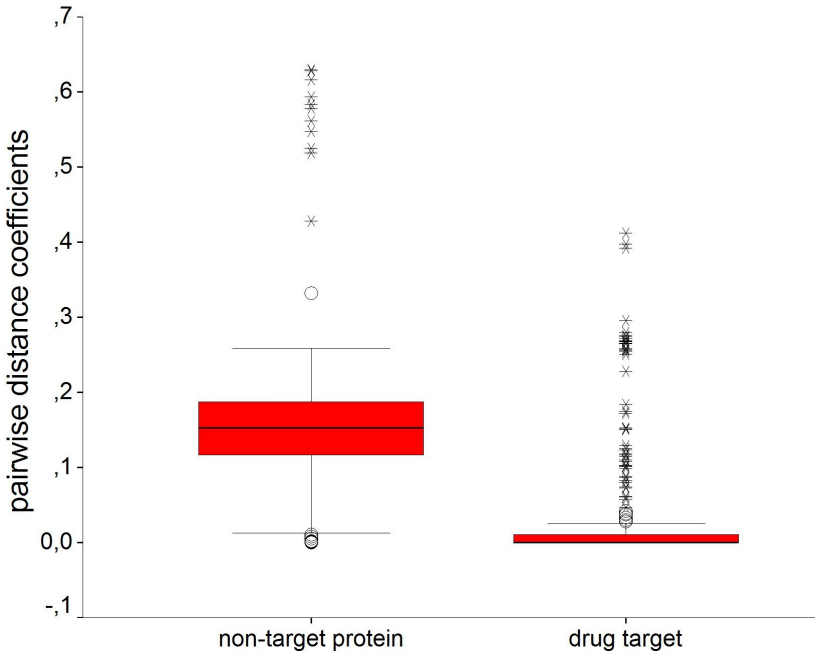


Figura 5.12. Comparação do *boxplot* para drogáveis e não-drogáveis. Valores menores do coeficiente de dissimilaridade mínima indicam uma forte evidência de que a proteína seja drogável.

Tabela 5.1. Tabela de classificação para o coeficiente de dissimilaridade na classificação de novas proteínas como drogáveis ou não-drogáveis. Classificação baseada no valor mínimo da dissimilaridade usando um corte de $0,10$.

Druggable target	(+)	(-)	Total
Yes	323	42	365
No	29	213	242
Total	352	255	607

de discriminação do coeficiente de dissimilaridade na classificação de novas proteínas como drogáveis ou não-drogáveis. A Tabela 5.1 exibe os resultados relativos à análise da curva ROC (Figura 5.13). O melhor valor de corte para o mínimo valor do coeficiente de dissimilaridade é $C = 0,10$. Neste cenário, tanto a sensibilidade como a especificidade são de 88%.

Para avaliar a consistência do processo de projeção de novas proteínas no espaço vetorial do modelo, mapeamos oito alvos conhecidos e não presentes no espaço originalmente. Desses oito alvos, sete são proteínas humanas da família GPCR (DRD2_HUMAN, ACM3_HUMAN, 5HT1D_HUMAN, 5HT2B_HUMAN, 5HT1F_HUMAN, NK3R_HUMAN, OPRM_HUMAN). Esses alvos estão no conjunto dicutido por Keiser et al. [2009].

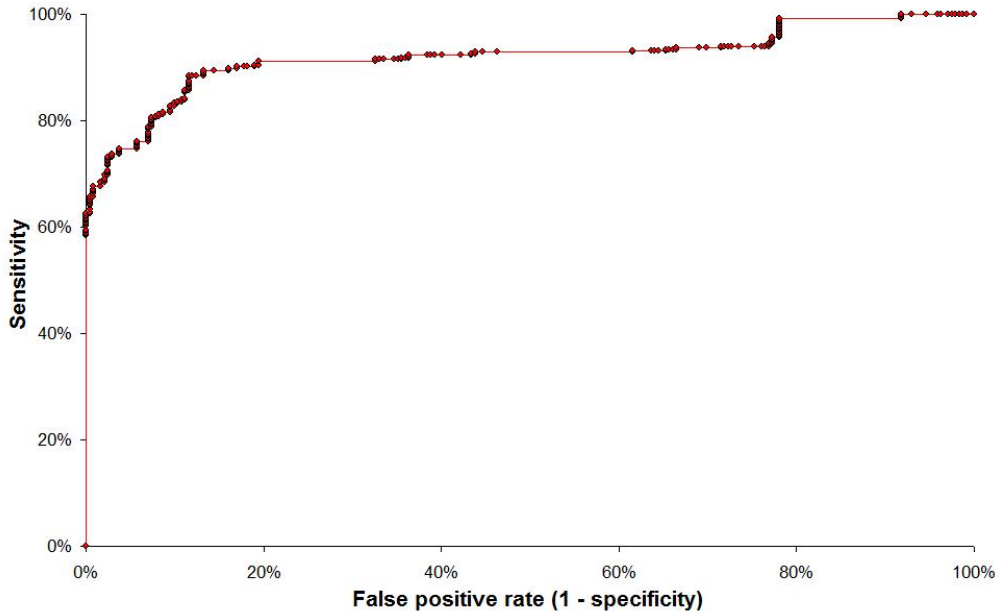


Figura 5.13. Análise de curva ROC para o modelo vetorial. O valor de ACU é maior do que 0,90 mostrando uma boa qualidade de discriminação para o coeficiente de dissimilaridade na classificação de novas proteínas como drogáveis ou não-drogáveis.

A projeção de cada novo alvo se fez conforme representado pela Equação 4.4. O posicionamento dos alvos projetados no espaço mostrou-se consistente com os agrupamentos obtidos com os alvos iniciais e com as propriedades biológicas conhecidas e descritas pelas respectivas anotações no InterPro. A Figura 5.11 ilustra o posicionamento das GPCRs humanas (região III da figura). As novas GPCRs projetadas foram todas posicionadas pelo algoritmo de otimização junto às outras GPCRs e, além disso, próximas a receptores com os quais compartilham subcategorias.

Este exemplo indica que o modelo é consistente e parece representar adequadamente as proteínas drogáveis. Permite também analisar alvos adicionais projetados no espaço. A pergunta mais interessante é se o modelo é capaz de sugerir potenciais alvos, mesmo sem haver ligação direta com os alvos conhecidos.

A seguir, tem-se um exemplo em que o modelo é capaz de descobrir associações implícitas que podem revelar interessantes oportunidades para pesquisa. Avaliamos o caso da *Kynurenine 3-monooxygenase* (KMO_HUMAN) e a *Peroxisomal sarcosine oxidase* (SOX_HUMAN). A proteína KMO_HUMAN foi projetada no espaço vetorial do modelo e foi calculada a dissimilaridade desta em relação aos alvos conhecidos. A proteína SOX_HUMAN foi o segundo alvo conhecido na lista de mais similares ao novo candidato. E o valor da dissimilaridade vetorial entre essas duas proteínas (0,00143) é bem menor do que o valor de corte definido no modelo para classificar potenciais alvos

drogáveis (0,1). A Figura 5.14 apresenta os primeiros alvos na lista ordenada. Somente os dois primeiros apresentaram dissimilaridade abaixo de 0,1. O mais interessante nesse caso é que é possível explicar como a relação indireta entre KMO_HUMAN e SOX_HUMAN pode ter surgido. Apesar de não compartilharem nenhuma assinatura entre si, cada uma compartilha outra assinatura com a ERG1_HUMAN, que é o alvo com menor índice de dissimilaridade em relação à KMO_HUMAN. O identificador IPR003042 relaciona KMO_HUMAN com ERG1_HUMAN e o IPR006076 relaciona a ERG1_HUMAN com a SOX_HUMAN. Os outros três identificadores do InterPro que aparecem na figura por estarem associados a um dos três envolvidos não estão associados a nenhum outro alvo.

A proteína KMO_HUMAN não é um alvo reconhecido como drogável em nossa base de dados inicial. Mas é um caso de interesse porque estudos têm sugerido que esta proteína está relacionada com o desenvolvimento da esquisofrenia [Aoyama et al., 2006; Holtze et al., 2011]. Por sua vez, SOX_HUMAN é um alvo conhecido da glicina, um aminoácido não-essencial, e a glicina é usada como uma alternativa no tratamento da esquisofrenia [Semba, 1998; Heresco-Levy et al., 2004]. Talvez essa relação entre a SOX e a KMO não seja suficiente para indicar a drogabilidade da KMO, mas ainda assim serve para indicar um potencial alvo terapêutico atacado sozinho ou em conjunto com a SOX em terapias adjuntivas.

Apesar de ter sido desenvolvido exclusivamente com alvos humanos, o modelo pode ser útil em pesquisas envolvendo o desenvolvimento de medicamentos antiparasitas. Se um dado domínio proteico humano é inibido com sucesso no tratamento de outra doença, ele pode ser útil para identificar novas classes de compostos que ajam de forma efetiva no mesmo domínio em patógenos [Hasan et al., 2006]. Além disso, similaridades funcionais entre um potencial alvo patológico e alvos humanos podem indicar a possibilidade de efeitos adversos.

Foi realizado um estudo de caso para averiguar o potencial do modelo na predição de alvos relacionados a doenças tropicais negligenciadas a partir do conhecimento de alvos humanos, que recebem, em geral, maior atenção e investimento pela indústria farmacêutica. Projetamos alvos do *Plasmodium falciparum* e do *Toxoplasma gondii* no espaço vetorial e analisamos sua similaridade com os alvos humanos conhecidos. Foram usadas todas as sequências das cepas designadas no UniProt por PLAFA e TOXGO presentes na versão do InterPro que utilizamos. A tabela com os pares mais relevantes está disponível como material suplementar. Nesta seção, analisamos especificamente as proteínas desses parasitas que representam uma boa oportunidade de reposicionamento de medicamentos que relacionam-se com o metabolismo de ácidos graxos (Tabela 5.2).

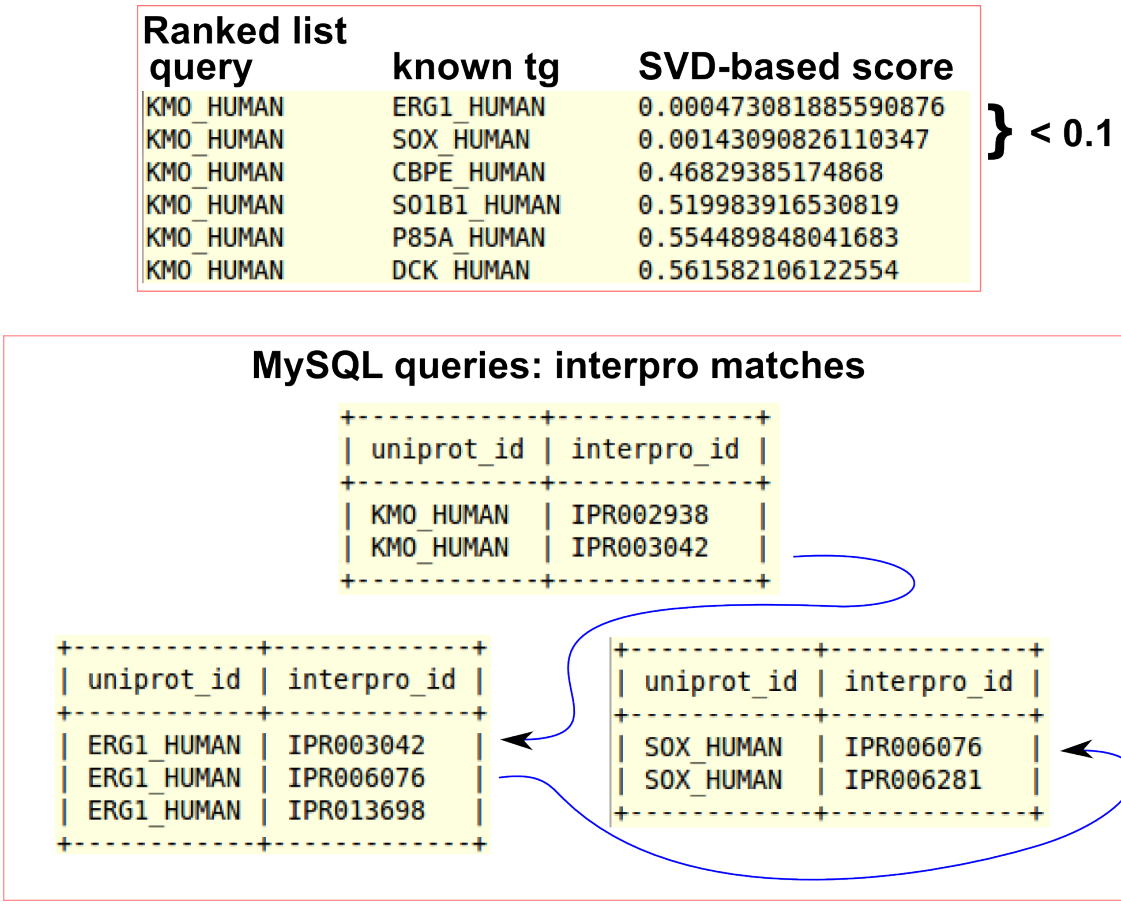


Figura 5.14. Exemplo de recuperação de informação implícita. A sequência KMO_HUMAN foi projetada no espaço do modelo vetorial e teve calculado o índice de dissimilaridade em relação a todos os alvos conhecidos. A tabela com os valores de dissimilaridades ordenados exibe dois alvos com índice abaixo do valor de corte 0.1 definido no modelo para classificar potenciais alvos drogáveis. O primeiro alvo conhecido na lista compartilha um termo do InterPro com a nova proteína, mas o segundo não. A alta similaridade apontada pelo modelo entre KMO_HUMAN e SOX_HUMAN pode ser explicada pela presença do alvo ERG1_HUMAN que compartilha um termo do InterPro com cada uma das outras proteínas. Estudos têm sugerido que a proteína KMO_HUMAN esteja relacionada com o desenvolvimento da esquisofrenia [Aoyama et al., 2006; Holtze et al., 2011]. Por sua vez, SOX_HUMAN é um alvo conhecido da glicina e a glicina é usada como uma alternativa no tratamento da esquisofrenia [Semba, 1998; Heresco-Levy et al., 2004].

Tabela 5.2. Potenciais alvos de *P. falciparum* e *T. gondii* para o agente anti-obesidade orlistat

rank	query	alvo humano conhecido	dissimilaridade vetorial	BLAST rank	BLAST bitscore	BLAST e-value	medicamento
1	Q1JTE1_TOXGO	FAS_HUMAN	0.01943580	1	263	3e-70	orlistat e cerulenin
2		GNMT_HUMAN	0.07633852	1296	20	8622	
3		ACPM_HUMAN	0.15228237	10	32	1.4	
4		ADH1A_HUMAN	0.34592475	19	30	7.4	fomepizole
5		DHSO_HUMAN	0.34592475	3	42	0.001	
1	B9PY50_TOXGO	FAS_HUMAN	0.08307553	1	74	1e-14	orlistat, cerulenin
2		GNMT_HUMAN	0.11404464	559	17	1895	
3		ACPM_HUMAN	0.21707082	1271	15	14392	
4		THIC_HUMAN	0.22979400	34	22	66	
5		PA2G6_HUMAN	0.54079587	550	17	1848	orlistat, quinacrine
1	O77078_PLAFA	THIC_HUMAN	0.04721318	102	20	216	
2		FAS_HUMAN	0.27036699	548	17	1815	orlistat, cerulenin
3		GNMT_HUMAN	0.36741667	1315	13	19247	
4		ACPM_HUMAN	0.44949792	1251	14	13901	
5		DGAT1_HUMAN	0.59385427	1145	15	9236	hesperetin

Destacamos os seguintes fármacos associados aos alvos humanos que aparecem na Tabela 5.2:

- orlistat: agente anti-obesidade projetado originalmente como um inibidor de lipase, identificado posteriormente como um inibidor da sintase de ácidos graxos (FAS) em ensaios clínicos com pacientes câncer;
- cerulenin: antibiótico inibidor da biosíntese dos ácidos graxos de bactérias e que também inibe a FAS.
- fomepizole: pirazol inibidor da álcool desidrogenase (ADH1A), usado como antídoto contra o envenenamento por metanol ou por etilenoglicol;
- quinacrine: anti-malárico aprovado nos anos 30;
- hesperetin: reduz o colesterol inibindo a atividade da acil-coenzima (ACAT1, ACAT2)

A síntese de ácidos graxos em bactérias e mamíferos processa-se pelas mesmas reações mas que são catalisadas por sistemas enzimáticos diferentes. Esses complexos contêm domínios não enzimáticos carregadores de acila. Na nossa análise, o alinhamento de sequências acusou similaridade significativa com a sintase de ácidos graxos humana (uniprot:FAS_HUMAN) apenas para duas proteínas do *T. gondii*: Q1JTE1_TOXGO (E-value=10^{-70}) e B9PY50_TOXGO (E-value=10^{-14}). Mas a similaridade vetorial sugere também a proteína do *P. falciparum* O77078_PLAFA (BLAST E-value extremamente alto: 1815) (Tabela 5.2). Destaca-se ainda a similaridade apontada pelo modelo vetorial para essas três sequências com os seguintes alvos terapêuticos conhecidos: a fosfolipase A2 (PA2G6_HUMAN), que catalisa a liberação de ácidos graxos

nas vias metabólicas da digestão e absorção de lipídios (KEGG-pathway:hsa04975) e da toxoplasmose (KEGG-pathway:hsa05145); a proteína carregadora de acil mitocondriana (ACPM_HUMAN) que articipa da ligação dos ácidos graxos (GO:0005504) e a diacilglicerol O-aciltransferase (DGAT1_HUMAN), que atua no processo metabólico dos ácidos graxos de cadeia longa.

Essas relações podem servir para sugerir oportunidades de reposicionamento de fármacos. Como a biosíntese de ácidos graxos é essencial para a sobrevivência de protozoários apicomplexos, diversos componentes desta via metabólica têm sido extensivamente pesquisados como potenciais alvos terapêuticos e a *Beta-ketoacyl-acyl carrier protein synthase III* (gene:FabH - uniprot:O77078_PLAFA) tem recebido especial atenção por ser fundamental no início do processo e por ser muito conservada em bactérias Gram-positivas e Gram-negativas [Li et al., 2011; Castillo & Pérez, 2008; Mazumdar et al., 2006; Nie et al., 2005; Khandekar et al., 2003]. O orlistat é um agente anti-obesidade, projetado inicialmente como um inibidor da lipase, mas que inibe também a FAS e, por esta característica, tem sido estudado como um potencial inibidor do crescimento de *P. falciparum*, *T. gondii* e outros parasitas apicomplexos [Miculka et al., 2011].

A FAS e a ação do orlistat e do cerulenin sobre ela, são o cerne também de diversos estudos sobre tratamento oncológicos [Flavin et al., 2010; Carvalho et al., 2008; Okawa et al., 2008; Kridel et al., 2007; Menendez et al., 2005; Kridel et al., 2004]. A FAS é geralmente muito expressa em células cancerígenas fazendo com que ela seja um atrativo alvo em terapias antitumorais. Entretanto, o orlistat apresenta várias limitações (permeabilidade celular baixa, baixa solubilidade, a falta de seletividade e biodisponibilidade e estabilidade metabólica pobres) que têm motivado o desenvolvimento de compostos análogos [Flavin et al., 2010]. Esse cenário pode resultar em novas descobertas para o tratamento das doenças causadas por parasitas muitas vezes negligenciadas. Casos de insucesso para tratamentos oncológicos prolongados podem mostrar-se interessantes para o tratamento de patologias causadas pelos microorganismos apicomplexos.

A integrase (IN) é um alvo clinicamente validado para o tratamento de infecções causadas pelo vírus da imunodeficiência humana (HIV) e a pesquisa por novas gerações de inibidores desta enzima continua sendo importante devido a resistências desenvolvidas por determinadas mutações [Métifiot et al., 2010].

Há no UniProtKB/Swiss-Prot 49 registros identificados como integrase do HIV. Apesar de apresentar diferenças na sequências de aminoácidos, todos apresentam exatamente as mesmas anotações do InterPro, que é o que importa no nosso modelo. Usamos o UniProtID POL_HV1H2 para denotar este alvo.

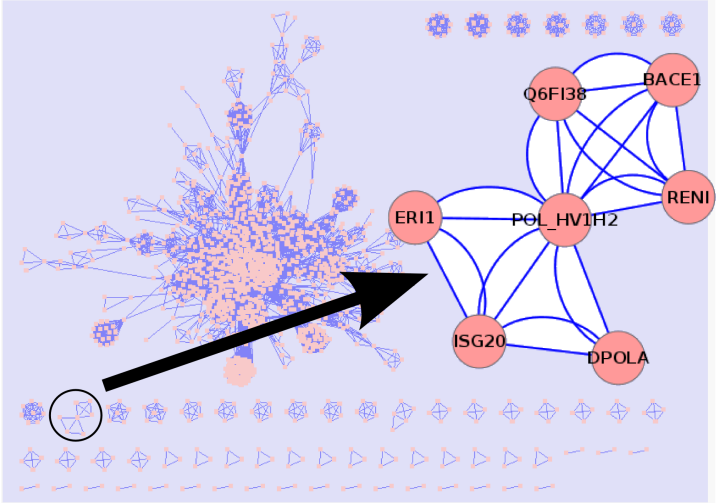


Figura 5.15. Visualização dos dados em uma rede construída com o algoritmo baseado em força implementado no Cytoscape. Em destaque, a integrase (IN) do HIV, projetada *a posteriori*, localizada bem no centro de um grupo contendo outros alvos com domínios de peptidase aspártica ou *Polynucleotidyl transferase, ribonuclease H fold* (domínios também encontrados e essenciais no HIV). Isso indica a possibilidade de desenvolvimento racional de novos compostos para agirem sobre a transcriptase do HIV.

Projetamos a IN no espaço vetorial do modelo e avaliamos sua similaridade com os demais alvos. Avaliando os alvos em uma rede, a IN é encontrada bem no centro de um grupo formado por outros alvos com domínios de peptidase aspártica ou *Polynucleotidyl transferase, ribonuclease H fold*, domínios também encontrados na IN (Figura 5.15).

Os relacionamentos apontados pelo modelo entre a IN e os alvos conhecidos podem ser usados para sugerir estratégias no desenvolvimento racional de um agente anti-HIV a partir do conhecimento das propriedades químicas dos fármacos associados aos alvos humanos.

5.1 Predição de alvos drogáveis usando regressão logística

A análise logística forneceu uma fórmula para se calcular a probabilidade p de uma dada proteína humana ser drogável:

$$p = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{66} \hat{\beta}_i x_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{66} \hat{\beta}_i x_i}}, \tag{5.1}$$

Tabela 5.3. Variáveis explicativas (InterPro) e seus respectivos coeficientes definidos pelo modelo de regressão logística para a predição da drogabilidade de uma proteína humana. O valor da constante é $\hat{\beta}_0 = -0,9$.

InterPro	$\hat{\beta}$	*p-value*	InterPro	$\hat{\beta}$	*p-value*
IPR016175	-7.3	0.067	IPR001023	-1.9	0.009
IPR012677	-4.8	0.000	IPR020685	-1.9	0.000
IPR010993	-4.5	0.004	IPR003593	-1.6	0.000
IPR004000	-3.7	0.000	IPR003596	-1.4	0.034
IPR000883	-3.4	0.001	IPR016040	-0.6	0.001
IPR008973	-3.3	0.000	IPR001452	1.3	0.032
IPR001173	-2.8	0.006	IPR020683	1.3	0.045
IPR016137	-2.6	0.011	IPR013099	1.4	0.065
IPR013783	-2.6	0.000	IPR000980	1.5	0.016
IPR013766	-2.5	0.012	IPR015421	1.6	0.000
IPR002213	-2.4	0.001	IPR011029	1.6	0.006
IPR011009	-2.4	0.000	IPR000472	1.8	0.030
IPR000873	-2.1	0.003	IPR013816	1.8	0.041
IPR000010	-2.1	0.040	IPR000889	1.8	0.031
IPR003597	-2.1	0.001	IPR011348	2.2	0.080
IPR008753	-2.0	0.052	IPR007698	2.2	0.080
IPR001353	-1.9	0.008	IPR014756	2.2	0.004
IPR011497	2.2	0.074	IPR015741	3.3	0.001
IPR005225	2.3	0.013	IPR017193	3.3	0.023
IPR001251	2.3	0.028	IPR000626	3.3	0.023
IPR002035	2.4	0.000	IPR020663	3.3	0.000
IPR001841	2.5	0.028	IPR008979	3.3	0.002
IPR011304	2.6	0.028	IPR009130	3.6	0.018
IPR000157	2.6	0.012	IPR014729	3.8	0.000
IPR013027	2.7	0.014	IPR001828	3.8	0.000
IPR002314	2.8	0.034	IPR003116	3.9	0.002
IPR008957	2.9	0.000	IPR020722	4.0	0.039
IPR015015	2.9	0.020	IPR020727	4.0	0.023
IPR011992	3.0	0.000	IPR009134	5.0	0.007
IPR005834	3.1	0.010	IPR002126	5.2	0.001
IPR009030	3.2	0.001	IPR008424	5.2	0.000
IPR005821	3.2	0.000	IPR000353	5.6	0.000
IPR000001	3.2	0.030	IPR016243	7.7	0.000

onde a constante $\hat{\beta}_0 = -0,9$ e os outros coeficientes $\hat{\beta}_i$ e seus respectivos termos do InterPro aparecem relacionados na Tabela 5.3.

Observa-se que, dos 66 termos do InterPro mantidos no modelo, 22 contribuem para reduzir a probabilidade da proteína ser alvo (coeficientes negativos) e os outros 44 contribuem para aumentar essa mesma probabilidade (coeficientes positivos).

Além de fornecer uma fórmula simples para o cálculo da probabilidade de uma proteína ser drogável, o modelo probabilístico aqui apresentado permite inferir sobre

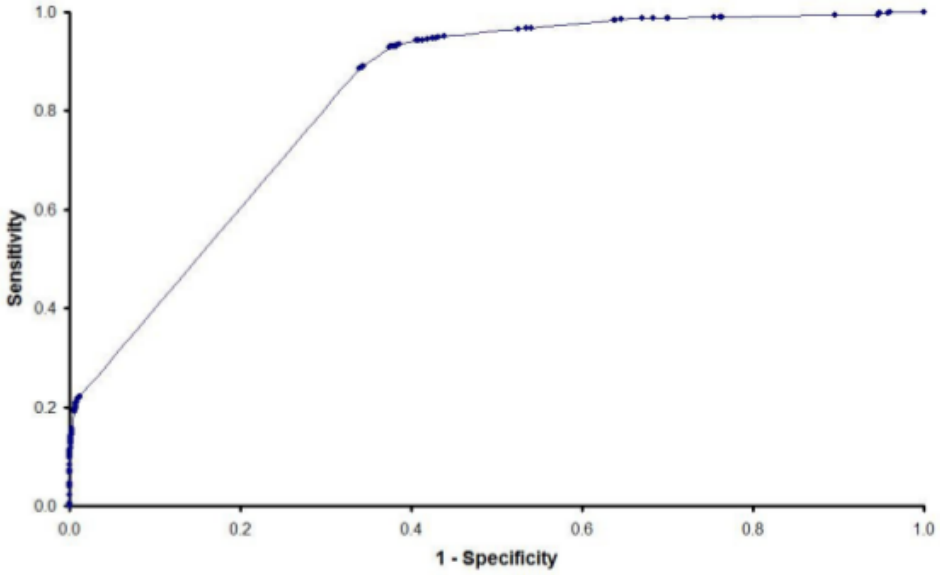


Figura 5.16. Curva ROC para a predição de um alvo drogável. O melhor valor de corte para maximizar a sensibilidade e a especificidade é uma probabilidade maior do que 0,25 (área sob a curva = 0,828) com sensibilidade de 89% e especificidade de 67%.

as propriedades biológicas que mais contribuem para garantir a drogabilidade de uma molécula. Isto pode ser feito observando-se as anotações funcionais que são mais significativas para aumentar ou diminuir a probabilidade de uma dada proteína ser drogável (Tabelas 5.4 e 5.5).

5.1.1 Classificação de alvos usando regressão logística

Para concluir a construção do modelo probabilístico, faltava apenas determinar o valor de corte para a probabilidade calculada pela fórmula desenvolvida. Esse valor de corte deve ser o que melhor discrimina as proteínas drogáveis das não-drogáveis. Usualmente, utilizaria-se o ponto de corte de $0,50$ significando que a probabilidade da proteína ser drogável é maior do que 50%. Entretanto, outros valores de corte podem ser avaliados em uma análise de curva ROC. Esse procedimento foi realizado (Figura 5.16) e sugeriu um valor de corte de $0,25$ para a probabilidade que maximiza ambas: sensibilidade e especificidade. Para validar o modelo, foi utilizado neste caso o conjunto de 384 alvos reservados juntamente com os 2009 "não-alvos" também reservados.

Tabela 5.4. Termos do InterPro mantidos no modelo probabiístico e que contribuem para aumentar a probabilidade da proteína ser drogável.

InterPro ID	InterPro Name
IPR001452	Src homology-3 domain
IPR020683	Ankyrin repeat-containing domain
IPR013099	Ion transport 2
IPR000980	SH2 motif
IPR015421	Pyridoxal phosphate-dependent transferase, major region, subdomain 1
IPR011029	DEATH-like
IPR000472	TGF-beta receptor/activin receptor, type I/II
IPR013816	ATP-grasp fold, subdomain 2
IPR000889	Glutathione peroxidase
IPR011348	17beta-dehydrogenase
IPR007698	Alanine dehydrogenase/PNT, C-terminal
IPR014756	Immunoglobulin E-set
IPR011497	Protease inhibitor, Kazal-type
IPR015741	Protein kinase, Snf1-like AMPK
IPR005225	Small GTP-binding protein
IPR017193	Anti-muellerian hormone receptor, type II
IPR001251	Cellular retinaldehyde-binding/triple function, C-terminal
IPR000626	Ubiquitin
IPR002035	von Willebrand factor, type A
IPR020663	Spindle assembly checkpoint kinase
IPR001841	Zinc finger, RING-type
IPR008979	Galactose-binding domain-like
IPR011304	L-lactate dehydrogenase
IPR009130	Tyrosine-protein kinase, JAK3
IPR000157	Toll-Interleukin receptor
IPR014729	Rossmann-like alpha/beta/alpha sandwich fold
IPR013027	FAD-dependent pyridine nucleotide-disulphide oxidoreductase
IPR001828	Extracellular ligand-binding receptor
IPR002314	Aminoacyl-tRNA synthetase, class II (G/H/P/S), conserved region
IPR003116	Raf-like Ras-binding
IPR008957	Fibronectin, type III-like fold
IPR020722	Tyrosine-protein kinase, vascular endothelial growth factor receptor 1
IPR015015	F-actin binding
IPR020727	Tyrosine-protein kinase, platelet-derived growth factor receptor beta
IPR011992	EF-hand-like domain
IPR009134	Tyrosine-protein kinase, vascular endothelial growth factor receptor
IPR005834	Haloacid dehalogenase-like hydrolase
IPR002126	Cadherin
IPR009030	Growth factor, receptor
IPR008424	Immunoglobulin C2-set
IPR005821	Ion transport
IPR000353	MHC class II, beta chain, N-terminal
IPR000001	Kringle
IPR016243	Tyrosine-protein kinase, CSF-1/PDGF receptor

Tabela 5.5. Termos do InterPro mantidos no modelo probabiístico e que contribuem para diminuir a probabilidade da proteína ser drogável.

InterPro ID	InterPro Name
IPR016175	Cytochrome b/b6
IPR001023	Heat shock protein Hsp70
IPR012677	Nucleotide-binding, alpha-beta plait
IPR020685	Tyrosine-protein kinase
IPR010993	Sterile alpha motif homology
IPR003593	ATPase, AAA+ type, core
IPR004000	Actin/actin-like
IPR003596	Immunoglobulin V-set, subgroup
IPR000883	Cytochrome c oxidase, subunit I
IPR016040	NAD(P)-binding domain
IPR008973	C2 calcium/lipid-binding domain, CaLB
IPR001173	Glycosyl transferase, family 2
IPR016137	Regulator of G protein signalling superfamily
IPR013783	Immunoglobulin-like fold
IPR013766	Thioredoxin domain
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase
IPR011009	Protein kinase-like domain
IPR000873	AMP-dependent synthetase/ligase
IPR000010	Proteinase inhibitor I25, cystatin
IPR003597	Immunoglobulin C1-set
IPR008753	Peptidase M13
IPR001353	Proteasome, subunit alpha/beta

5.2 Modelos para a representação de fármacos integrando dados químicos e biológicos

O estudo comparativo entre fármacos usando uma metodologia similar àquela usada para desenvolver o modelo de espaço vetorial apresentou resultados preliminares interessantes descritos nesta seção.

Primeiramente, foram analisados relacionamentos envolvendo fármacos e classes de ação farmacológica discutidos por Keiser et al. [2009, 2007] (método SEA). Apresenta-se os resultados obtidos com duas diferentes abordagens: a primeira usa apenas descritores relacionados com a estrutura química dos compostos (Tabela 5.6); a segunda combina os mesmos descritores químicos com os descritores biológicos que representam as anotações do InterPro associadas aos alvos relacionados (Tabela 5.7).

Em ambas as tabelas, as associações compartilhadas pelos dois compostos envolvidos aparecem denotadas em itálico. "Novas associações" previstas e comprovadas pelo método SEA aparecem em negrito. Por fim, denotamos em azul, outras associações que consideramos importantes para serem avaliadas considerando nossas pesquisas

bibliográficas.

Na Tabela 5.7, destacamos o caso da relação entre *delavirdine*, um inibidor da transcriptase reversa com atividade específica para o vírus HIV e *methotrexate* (MTX), um antimetabólito antineoplásico com propriedades imunosupressoras usado no tratamento da psoríase. Há preocupações a respeito do uso do MTX em pessoas infectadas pelo HIV porque isso pode acelerar o desenvolvimento da AIDS e aumentar a incidência de infecções oportunistas. Discussões sobre as condições de segurança no uso do MTX por pacientes infectados por HIV tem sido publicado na literatura científica há bastante tempo baseando em observações clínicas, mas falta ainda encontrar uma sólida fundamentação lógica [Civatte, 1989; Witty et al., 1992; Maurer et al., 1994; Chernyshov, 2006; Menon et al., 2010; Chen et al., 2012].

Apresentamos também uma breve análise em rede onde analisamos os relacionamentos entre algumas classes de fármacos discutidas por Luo et al. [2011]. As redes foram obtidas com o algoritmo baseado em força implementado no Cytoscape.

A rede na Figura 5.17 sugere a suposição sobre potenciais aplicações para antipsicóticos antivirais e antibióticos.

A inclusão de 17 vasodilatadores, usados para o tratamento de disfunção erétil (ER) (sildenafil, tadalafil, udenafil, vardenafil and papaverine), hipertensão, insuficiência cardíaca ou angina, mostra que esta classe de compostos relaciona-se com alguns antipsicóticos e antivirais (Figuras 5.17 e 5.18).

Algumas dessas relações são corroboradas pela literatura, como: o caso de agentes para o tratamento da ER e antivirais inibidores de protease, em particular tadalafil e ritonavir [Loulergue et al., 2011]; efeitos colaterais sexuais de antidepressantes e antipsicóticos [Baldwin & Mayers, 2003; Clayton et al., 2002]; a atividade antiviral do dipyridamole [Tonew et al., 1977; Tonew et al., 1978; Tonew et al., 1982] e a associação do dipyridamole (um inibidor não-seletivo da PDE com atividade inibitória do PDE5) com outros inibidores da PDE5 usados no tratamento da impotência [Ghofrani et al., 2006].

Tabela 5.6. Similaridade entre fármacos pela abordagem usando SVD sobre o espaço químico

Query	SVD rank	Tanimoto rank	Similar drug by SVD approach	Activity class of the similar drug	Tanimoto coefficient
Delavirdine	1	13	Tripelennamine	**Histamine H1 receptor inhibitor**	0.426
	2	16	Chloropyramine	**Histamine H1 receptor inhibitor**	0.396
	3	69	Imatinib	BCR/ABL fusion inhibitor	0.396
	4	?	Procainamide	Sodium channel inhibitor and other	< 0.3
	5	5	Leucovorin	Thymidylate synthase inhibitor	0.468
	6	57	Enoxacin	DNA topoisomerase/gyrase inhibitor	0.403
	7	6	Saquinavir	HIV-1 protease inhibitor	0.452
	8	2	Dabigatran etexilate	Prothrombin inhibitor	0.487
Domperidone	1	1	Pimozide	*Dopamine antagonist*	0.743
	2	11	Trazodone	**α-Adrenergic antagonist** Histamine H1 antagonist; Serotonin inhibitor	0.403
	3	2	Droperidol	**α-Adrenergic antagonist**; D(2)-Dopamine antagonist	0.576
	4	?	Pirenzepine	***Muscarinic antagonist***	< 0.3
	5	14	Chloroquine	Ferriprotoporphyrin IX antagonist; Toll-like antagonist	0.396
Fluoxetine	1	1	Atomoxetine	Noradrenaline/*Serotonin* inhibitor	0.798
	2	2	Reboxetine	Noradrenaline inhibitor	0.635
	3	4	Duloxetine	Noradrenaline/*Serotonin*/Dopamine inhibitor	0.573
	4	3	Bisoprolol	**β-1/2 Adrenergic receptor antagonist**	0.595
	5	5	Tramadol	Opioid receptor agonist; Noradrenaline/*Serotonin* inhibitor; NMDA antagonist	0.493
	6	8	Dyclonine	Sodium channel inhibitor	0.534
	7	7	Esmolol	**β-1 Adrenergic receptor antagonist**	0.556
Paroxetine	1	8	Dyclonine	Sodium channel inhibitor	0.497
	2	1	Nebivolol	**β-1/2 adrenergic receptor antagonist**	0.539
	3	10	Atomoxetine	*Noradrenaline/Serotonin inhibitor*	0.494
	4	5	Fluoxetine	*Serotonin transporter inhibitor*	0.506
	5	21	Esmolol	**β-1 Adrenergic receptor antagonist**	0.478
	6	4	Tramadol	**α-Adrenergic antagonist**; Histamine H1 antagonist; Serotonin inhibitor	0.509
	7	13	Reboxetine	*Noradrenaline inhibitor*	0.492
	8	27	Cycrimine	*Muscarinic receptor M1 antagonist*	0.468
Methadone	1	1	Isopropamide	**Muscarinic M3/4 antagonist**	0.651
	2	9	Milnacipran	Sodium-dependent serotonin/dopamine transporter inhibitor	0.561
	3	bla	Levomethadyl Acetate	*Opioid agonist*	0.577
	4	5	Methadyl Acetate	*μ-Opioid agonist*	0.577
	5	bla	Doxapram	Potassium channel inhibitor	0.566
	6	7	Levocabastine	Histamine H1 antagonist; Neurotensin receptor partial antagonist	0.569
	7	19	Glutethimide	GABA(A) agonist	0.520
	8	2	Fexofenadine	Histamine H1 antagonist	0.590
	9	bla	Maprotiline	**Muscarinic M1/2/3/4/5 antagonist**; α **1-A Adrenergic antagonist** ; and other	0.527
	10	13	Ketamine	*NMDA receptor blocker*; **Neurokinin NK1 receptor antagonist**	0.550
Loperamide	1	2	Isopropamide	**Muscarinic M3/4 antagonist**	0.669
	2	1	Haloperidol	Dopamine antagonist; NMDA receptor antagonist	0.681
	3	13	Halofantrine	Ferriprotoporphyrin; Potassium channel inhibitor; HERG channels blocker	0.559
	4	4	Milnacipran	Sodium-dependent serotonin/dopamine transporter inhibitor	0.636
	5	5	Chlophedianol	Histamine H1 antagonist	0.632
	6	3	Doxapram	Potassium channel inhibitor	0.656
	7	15	Fexofenadine	Histamine H1 antagonist	0.549
	8	25	Methadone	*μ-Opioid agonist*; NMDA receptor antagonist; Neuronal acetylcholine antagonist	0.511
	9	16	Diphenidol	**Muscarinic M1/2/3 Antagonist**	0.537
	10	9	Cycrimine	**Muscarinic M1 Antagonist**	0.568

[1]Em negrito: classes bioativas preditas por [Keiser et al., 2009, 2007] para a *query* dada.
[2]Em itálico: associações esperadas (por *prima faccia*).
[3]Em azul: nova associação corroborada por outras publicações.

Tabela 5.7. Similaridade entre fármacos pela abordagem usando SVD sobre o espaço químico e biológico

Query	SVD rank	Tanimoto rank	Similar drug by SVD approach	Activity class of the similar drug	Tanimoto coefficient
Delavirdine	1	13	Tripelennamine	**Histamine H1 receptor inhibitor**	0.426
	2	16	Chloropyramine	**Histamine H1 receptor inhibitor**	0.424
	3	6	Saquinavir	HIV-1 protease inhibitor	0.452
	4	5	Leucovorin	Thymidylate synthase inhibitor	0.468
	5	99	Metolazone	Thiazide-sensitive Na-Cl cotransporter inhibitor	0.380
	6	?	Procainamide	Sodium channel inhibitor and other	< 0.3
	7	2	Dabigatran etexilate	Prothrombin inhibitor	0.487
	8	22	***Methotrexate***	Dihydrofolate reductase inhibitor	0.419
Domperidone	1	1	Pimozide	*Dopamine antagonist*	0.743
	2	2	Droperidol	**α-Adrenergic antagonist**; D(2)-Dopamine antagonist	0.576
	3	?	Pirenzepine	Muscarinic antagonist	< 0.3
	4	11	Trazodone	**α-Adrenergic antagonist** Histamine H1 antagonist; Serotonin inhibitor	0.403
	5	8	Azelastine	Noradrenaline inhibitor	0.409
Fluoxetine	1	1	Atomoxetine	Noradrenaline/*Serotonin* inhibitor	0.798
	2	2	Reboxetine	Noradrenaline inhibitor	0.635
	3	4	Duloxetine	Noradrenaline/*Serotonin*/Dopamine inhibitor	0.573
	4	3	Bisoprolol	**β-1/2 Adrenergic receptor antagonist**	0.595
	5	7	Esmolol	**β-1 Adrenergic receptor antagonist**	0.556
	6	5	Tramadol	Opioid receptor agonist; Noradrenaline/*Serotonin* inhibitor; NMDA antagonist	0.493
	7	9	Fesoterodine	Muscarinic M1/2/3/4/5 antagonist	0.482
	8	8	Dyclonine	Sodium channel inhibitor	0.534
	9	?	Metoprolol	**β-1/2 Adrenergic receptor antagonist**	< 0.3
Paroxetine	1	5	Fluoxetine	*Serotonin transporter inhibitor*	0.506
	2	1	Nebivolol	**β-1/2 adrenergic receptor antagonist**	0.539
	3	10	Atomoxetine	*Noradrenaline/Serotonin inhibitor*	0.494
	4	8	Dyclonine	Sodium channel inhibitor	0.497
	5	38	Trihexyphenidyl	*Muscarinic M1/2/3/4/5 antagonist*	0.459
	6	34	Procyclidine	*Muscarinic M1/2/3/4 antagonist*	0.462
	7	77	Tolterodine	**Muscarinic M1/2/3/4 antagonist**	0.421
	8	4	Tramadol	**α-Adrenergic antagonist**; Histamine H1 antagonist; *Serotonin inhibitor*	0.509
	9	45	Tridihexethyl	*Muscarinic M1/2/3 antagonist*	0.452
	10	27	Cycrimine	*Muscarinic receptor M1 antagonist*	0.468
	11	21	Esmolol	**β-1 Adrenergic receptor antagonist**	0.478
Methadone	1	6	Levomethadyl Acetate	*Opioid agonist*	0.577
	2	1	Isopropamide	**Muscarinic M3/4 antagonist**	0.651
	3	10	Meperidine	*κ-Opioid receptor agonist; NMDA receptor antagonist*	0.558
	4	5	Methadyl Acetate	*μ-Opioid agonist*	0.577
	5	13	Ketamine	*NMDA receptor blocker*; **Neurokinin NK1 receptor antagonist**	0.550
Loperamide	1	2	Isopropamide	**Muscarinic M3/4 antagonist**	0.669
	2	1	Haloperidol	Dopamine antagonist; NMDA receptor antagonist	0.681
	3	13	Halofantrine	Ferriprotoporphyrin; Potassium channel inhibitor; HERG channels blocker	0.559
	4	4	Milnacipran	Sodium-dependent serotonin/dopamine transporter inhibitor	0.636
	5	5	Chlophedianol	Histamine H1 antagonist	0.632
	6	3	Doxapram	Potassium channel inhibitor	0.656
	7	15	Fexofenadine	Histamine H1 antagonist	0.549
	8	25	Methadone	*μ-opioid agonist*; NMDA receptor antagonist; Neuronal acetylcholine antagonist	0.511
	9	16	Diphenidol	**Muscarinic M1/2/3 Antagonist**	0.537
	10	9	Cycrimine	**Muscarinic M1 Antagonist**	0.568

[1]Em negrito: classes bioativas preditas por [Keiser et al., 2009, 2007] para a *query* dada.
[2]Em itálico: associações esperadas (por *prima faccia*).
[3]Em azul: nova associação corroborada por outras publicações.

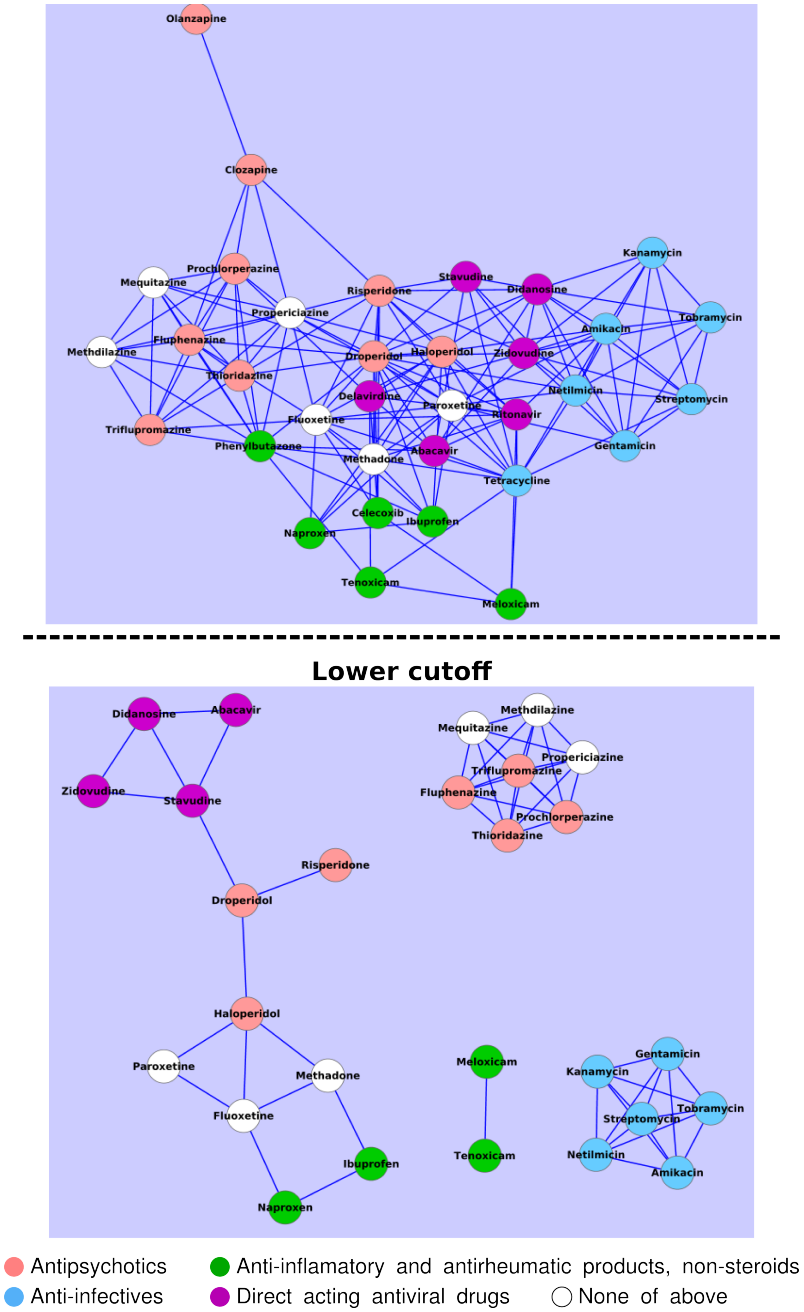


Figura 5.17. Rede de relacionamento entre fármacos. O agrupamento foi realizado pelo algoritmo baseado em força implementado no Cytoscape aplicado sobre a matriz de dissimilaridade no espaço químico/biológico. Há nove antipsicóticos, mostrados em laranja claro, quatro deles são fenotiazinas (fluphenazine, prochlorperazine, thioridazine e trifluoperazine) e cinco são não-fenotiazinas (clozapine, droperidol, haloperidol, olanzapine e risperidone). Também no conjunto: propericiazine, uma fenotiazina usada como sedativo com fracas propriedades antipsicóticas; fluoxetina e paroxetina, dois antidepressivos inibidores seletivos de recaptação serotonina (SSRIs) methadone, um analgésico opióidico mequitazine e methdilazine, duas fenotiazinas com atividade antihistamínica; seis anti-inflamatórios e antireumáticos, mostrados em verde (phenylbutazone, celecoxib, naproxen, ibuprofen, tenoxicam, meloxicam); sete antibióticos, mostrados em azul: tetracycline e seis outros que são aminoglycosides (gentamicin, streptomycin, netilmicin, amikacin, kanamycin e tobramycin; seis antivirais (stavudine, didanosine, zidovudine, bacavir, ritonavir e delavirdine).

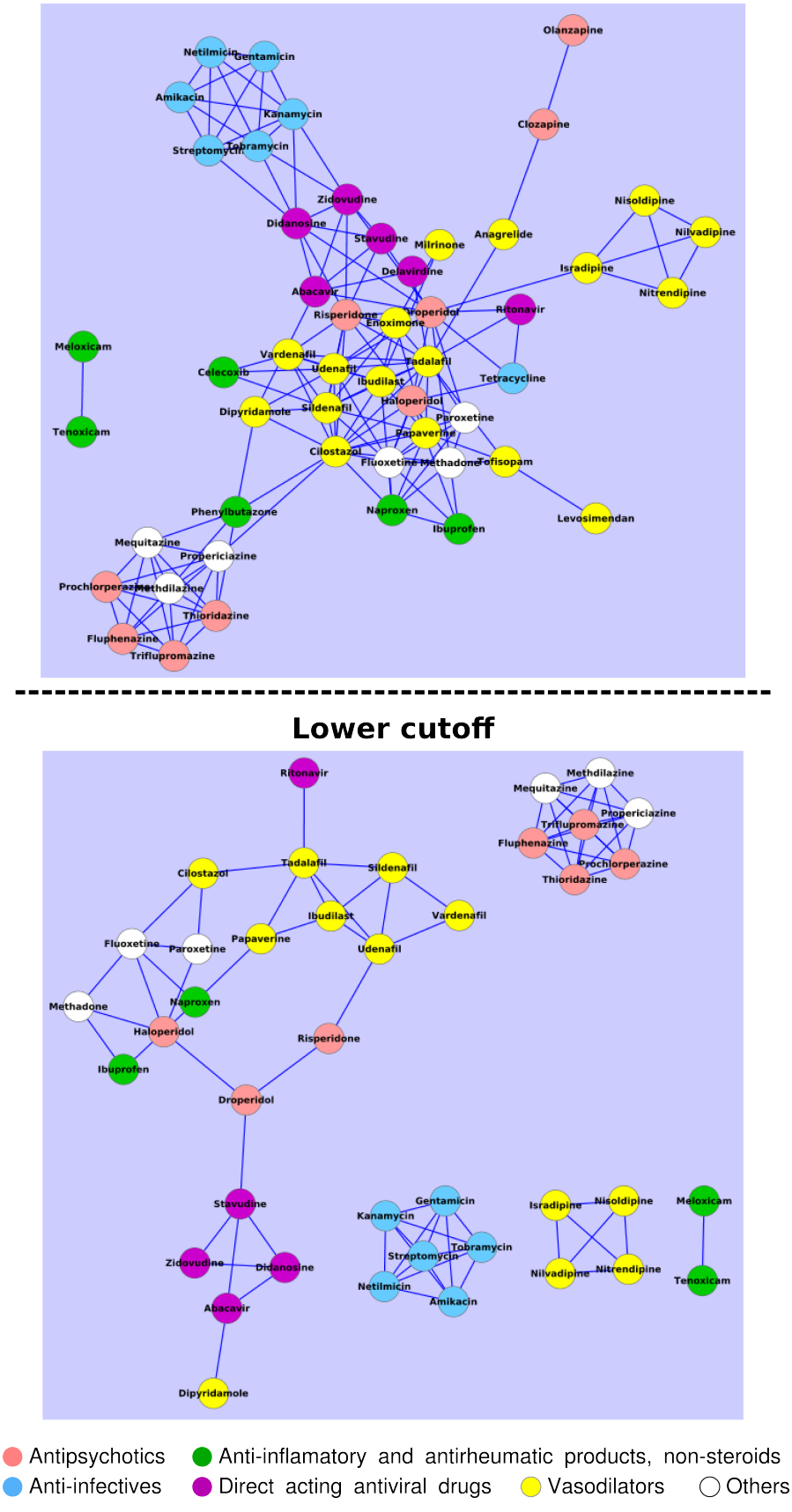


Figura 5.18. Rede de relacionamento entre fármacos. Foram acrescentados 17 vasodilatadores (em amarelo).

Capítulo 6

Conclusões

Neste trabalho foi proposto um método para identificar potenciais alvos drogáveis humanos no contexto de suas propriedades biológicas apreendidas por anotações do InterPro. A partir de uma extensiva e criteriosa coleta de informações distribuídas em diferentes repositórios públicos, cada alvo foi representado por um vetor de dimensão 2700. Todas as redundâncias e discrepâncias foram resolvidas. Ao final do processo de avaliação deste trabalho, estarão disponíveis na página do programa (ou em outro local a ser definido) todas as rotinas utilizadas, *parsers* produzidos e as *queries* SQL necessárias para extrair as informações de um banco de dados local, construído exclusivamente para dar suporte ao trabalho realizado.

A decomposição por valores singulares seguida de uma redução de posto foi o método utilizado para organizar semanticamente os alvos. O método "claramente encontra conclusões similares com aquelas apresentadas por Hopkins & Groom [2002] exibindo correlação entre famílias drogáveis e *motifs* proteicos"[1]. Além desta validação, mostrou-se que o modelo proposto "é eficiente e potencialmente efetivo para descobrir relacionamentos dificilmente detectados por similaridade entre sequências"[2]. Conclui-se que "o estudo é uma interessante aplicação da decomposição por valores singulares a uma grande matriz de alvos versus descritores do InterPro e tem mérito por encontrar resultados comparáveis com o BLAST"[3]. Isso foi verificado usando diferentes técnicas de agrupamento e visualização de conjunto de dados.

Dos 1906 alvos drogáveis humanos conhecidos, foram separados 365 e procedeu-se às análises convencionais para determinar a efetividade da métrica de dissimilaridade utilizada na busca por novos alvos drogáveis. Os candidatos são todas as 29580 sequên-

[1]Comentário de um *referee* anônimo sobre o artigo submetido à *Bioinformatics*.

[2]Comentário de um *referee* anônimo sobre o artigo selecionado para apresentação oral no *BICoB-2011*.

[3]Comentário de outro *referee* anônimo do *BICoB-2011*.

cias humanas depositadas no UniProtKB que compartilham alguma anotação do InterPro com algum alvo conhecido. Devido à inexistência de "não-alvos", foi utilizado um estratagema em que foram selecionados entre os candidatos, aqueles que apresentaram *e-value* sempre superior a 10 para o alinhamento par-a-par com cada alvo.

A métrica de dissimilaridade mostrou-se adequada para discriminar os alvos drogáveis. O valor de corte sugerido pela curva ROC permitiu classificar um alvo drogável com uma sensibilidade e especificidade de 88%.

Para avaliar os descritores (que são termos do InterPro), foi criado um modelo probabilístico baseado em regressão logística que selecionou 66 deles como relevantes. A validação do modelo foi no contexto de classificação (sensibilidade de 89% e especificidade de 67%). A análise de como uma propriedade interfere na drogabilidade de um alvo será objeto de um trabalho que ainda será desenvolvido.

Outros modelos vetoriais, que também usam vetores binários, foram construídos para representar os fármacos aprovados pela FDA e catalogados no DrugBank. Inicialmente, cada medicamento foi representado por um vetor obtido a partir de uma transformação do seu SMILES canônico. Um estudo de correlação com a métrica de Tanimoto (usual para comparar estruturas químicas de compostos) mostrou um pequeno ganho na qualidade da discriminação ao usar a métrica de dissimilaridade vetorial. Depois, foram acrescentados aos descritores de origem química, outros descritores que descrevem as classes terapêuticas dos fármacos ou as anotações biológicas relativas a seus alvos. Para cada conjunto de descritores foi avaliada novamente a correlação com a métrica de Tanimoto. Esse estudo mostrou que, da forma que foram usados, os descritores químicos são mais influentes na definição do coeficiente de dissimilaridade. Para avaliar a contribuição que pode ser extraída desta representação, alguns casos foram avaliados comparativamente com resultados discutidos em outros trabalhos [Keiser et al., 2009, 2007; Luo et al., 2011]. Estes resultados preliminares apresentam bons prognósticos que motivaram a realização de outro trabalho em andamento.

Como material suplementar, estão disponíveis duas tabelas: na primeira é exibido para cada alvo humano conhecido, o outro alvo (também entre os conhecidos) que apresentou maior similaridade pelo modelo vetorial. Além disso, para apresentar os casos menos triviais, são exibidos os quatro primeiros alvos que não compartilham anotações InterPro com o alvo da *query*. Além do *ranking* e da informação sobre o compartilhamento de anotações do InterPro, é informado nesta tabela se o par compartilha algum fármaco conforme as interações conhecidas. Na segunda tabela, é exibido os candidatos classificados como potencialmente drogáveis pelo modelo. Essas tabelas permitem que os biólogos possam avaliar o *status* definido pelo método para as proteínas de seu interesse.

Outras proteínas podem ser especuladas como alvos drogáveis. Por exemplo, foram projetados no espaço dos alvos humanos, proteínas de *P. falciparum* e de *T. gondii*. Esse estudo permitiu sugerir alvos dos patógenos com base no conhecimento sobre os alvos drogáveis humanos. Este tipo de análise pode servir para indicar possíveis casos de reação adversa. Permite também apontar oportunidades de reposicionamento de fármacos, originalmente projetados para tratar alguma enfermidade humana, para o combate de parasitas conforme exemplificado com o estudo de caso do orlistat, um agente anti-obesidade, predito pelo modelo como uma terapia alternativa contra a malária e a toxoplasmose.

6.1 Perspectivas

Este trabalho abre diversas frentes de pesquisa:

- extender o estudo sobre alvos de patógenos usando o modelo construído com alvos humanos e também aplicando a metodologia para construir novos modelos específicos para certas classes de patógenos;
- acrescentar atributos de outros tipos ao modelo vetorial de alvos, como, por exemplo, descritores relacionados a doenças (OMIM, MeSH etc);
- reconstruir o modelo vetorial usando apenas os descritores apontados como mais relevantes pela regressão logística e avaliar também o que ocorre se esses descritores forem retirados;
- particionar o modelo construindo espaços vetoriais separados para as principais categorias de alvos (enzimas, transportadores, canais iônicos, receptores e outros);
- construir e analisar redes de relacionamento em busca de padrões interessantes;
- validar o modelo vetorial representativo dos fármacos e reavaliar os resultados usando descritores químicos com significados estruturais;
- aplicar a regressão logística para a seleção de atributos que melhor caracterizam os medicamentos em relação a outros compostos químicos.

Referências Bibliográficas

Abadio, A. K.; Kioshima, E.; Teixeira, M.; Martins, N.; Maigret, B. & Felipe, M. S. (2011). Comparative genomics allowed the identification of drug targets against human fungal pathogens. *BMC Genomics*, 12(1):75.

Agarwal, P. & Searls, D. B. (2008). Literarture mining in support of drug discovery. *Brief Bioinform*, 12(4):383–389.

Altschul, S.; Gish, W.; Miller, W.; Myers, E. & Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Ampazis, N. & Perantonis, S. J. (2004). Lsisom - a latent semantic indexing approach to self-organizing maps of document collections. *Neural Process. Lett.*, 19(2):157–173.

Aoyama, N.; Takahashi, N.; Saito, S.; Maeno, N.; Ishihara, R.; Ji, X.; Miura, H.; Ikeda, M.; Suzuki, T.; Kitajima, T.; Yamanouchi, Y.; Kinoshita, Y.; Yoshida, K.; Iwata, N.; Inada, T. & Ozaki, N. (2006). Association study between kynurenine 3-monooxygenase gene and schizophrenia in the japanese population. *Genes, Brain and Behaviour*, 5(4):364–368.

Armitage, P.; Berry, G. & Matthews, J. (2002). *Statistical Methods in Medical Research*. Armitage, Statistical Methods in Medical Research. Blackwell Science.

Baldwin, D. & Mayers, A. (2003). Sexual side-effects of antidepressant and antipsychotic drugs. *Advances in Psychiatric Treatment*, 9(3):202–210.

Barrell, D.; Dimmer, E.; Huntley, R. P.; Binns, D.; O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009 - an integrated Gene Ontology Annotation resource. *Nucleic acids research*, 37(Database issue):D396–403.

Berriman, M.; Haas, B. J.; LoVerde, P. T.; Wilson, R. A.; Dillon, G. P.; Cerqueira, G. C.; Mashiyama, S. T.; Al-Lazikani, B.; Andrade, L. F.; Ashton, P. D.; Aslett, M. A.; Bartholomeu, D. C.; Blandin, G.; Caffrey, C. R.; Coghlan, A.; Coulson,

R.; Day, T. A.; Delcher, A.; DeMarco, R.; Djikeng, A.; Eyre, T.; Gamble, J. A.; Ghedin, E.; Gu, Y.; Hertz-Fowler, C.; Hirai, H.; Hirai, Y.; Houston, R.; Ivens, A.; Johnston, D. A.; Lacerda, D.; Macedo, C. D.; McVeigh, P.; Ning, Z.; Oliveira, G.; Overington, J. P.; Parkhill, J.; Pertea, M.; Pierce, R. J.; Protasio, A. V.; Quail, M. A.; Rajandream, M.-A.; Rogers, J.; Sajid, M.; Salzberg, S. L.; Stanke, M.; Tivey, A. R.; White, O.; Williams, D. L.; Wortman, J.; Wu, W.; Zamanian, M.; Zerlotini, A.; Fraser-Liggett, C. M.; Barrell, B. G. & El-Sayed, N. M. (2009). The genome of the blood fluke schistosoma mansoni. *Nature*, 460(7253):352–358.

Berry, M. W.; Dumais, S. T. & O'Brien, G. W. (1995). Using Linear Algebra for Intelligent Information Retrieval. Relatório técnico UT-CS-94-270, University of Tennessee.

Burge, S.; Kelly, E.; Lonsdale, D.; Mutowo-Muellenet, P.; McAnulla, C.; Mitchell, A.; Sangrador-Vegas, A.; Yong, S.-Y.; Mulder, N. & Hunter, S. (2012). Manual go annotation of predictive protein signatures: the interpro approach to go curation. *Database*, 2012.

Campbell, M. J.; Campbell (PhD.), M. J. & Swinscow, T. D. V. (2009). *Statistics at Square One*. John Wiley & Sons.

Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J. & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321:263–266.

Carvalho, M. A.; Zecchin, K. G.; Seguin, F.; Bastos, D. C.; Agostini, M.; Rangel, A. L. C.; Veiga, S. S.; Raposo, H. F.; Oliveira, H. C.; Loda, M.; Coletta, R. D. & Graner, E. (2008). Fatty acid synthase inhibition with orlistat promotes apoptosis and reduces cell growth and lymph node metastasis in a mouse melanoma model. *International Journal of Cancer*, 123(11):2557–2565.

Castillo, Y. P. & Pérez, M. A. C. (2008). Bacterial beta-ketoacyl-acyl carrier protein synthase iii (fabh): an attractive target for the design of new broad-spectrum antimicrobial agents. *Mini Rev Med Chem*, 8(1):36–45.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276.

Chagoyen, M.; Carmona-Saez, P.; Gil, C.; Carazo, J. M. & Pascual-Montano, A. (2006). A literature-based similarity metric for biological processes. *BMC Bioinformatics*, 7:363–375.

Chen, H.; Hayashi, G.; Lai, O. Y.; Dilthey, A.; Kuebler, P. J.; Wong, T. V.; Martin, M. P.; Fernandez Vina, M. A.; McVean, G.; Wabl, M.; Leslie, K. S.; Maurer, T.; Martin, J. N.; Deeks, S. G.; Carrington, M.; Bowcock, A. M.; Nixon, D. F. & Liao, W. (2012). Psoriasis patients are enriched for genetic variants that protect against hiv-1 disease. *PLoS Genet*, 8(2):e1002514.

Chen, M.-c.; sheng Chen, L.; chin Hsu, C. & rong Zeng, W. (2008). An information granulation based data mining approach for classifying imbalanced data. *Information Sciences*, 178:3214–3227.

Chen, X.; Ji, L. & Chen, Y. Z. (2002). Ttd: Therapeutic target database. *Nucleic Acids Research - Database issue*, 30:412–415.

Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C. & Huang, E. S. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, 25(1):71–75.

Chernyshov, P. (2006). Testing for hiv prior to methotrexate administration. is it an obligatory procedure? *International Journal of Dermatology*, 45(8):998–999.

Civatte, J. (1989). Psoriasis and hiv infection. *Bulletin de lAcademie nationale de medecine*, 173(8):1065–1070; discussion 1070–1071.

Clayton, A.; Pradko, J.; Croft, H.; Brendan Montano, C.; Leadbetter, R.; Bolden-Watson, C.; Bass, K.; Donahue, R.; Jamerson, B. & Metz, A. (2002). Prevalence of sexual dysfunction among newer antidepressants. *Journal of Clinical Psychiatry*, 63(4):357–366.

Cline, M.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M. & Gross, B. (2007). Integration of biological networks and gene expression data using cytoscape. *Nat. Protocol.*, pp. 2366–2382.

Coleman, T. F. & Li, Y. (1994). On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Mathematical Programming*, 67:189–224.

Consortium, T. U. (2010). The universal protein resource (uniprot) in 2010. *Nucleic Acids Research*, 38(suppl 1):D142–D148.

Cornish-Bowden, A. & Cardenas, M. L. (2003). Metabolic analysis in drug design. *Comptes Rendus Biologies*, 326:509–515.

Darapaneni, V.; Prabhaker, V. K. & Kukol, A. (2009). Large-scale analysis of influenza a virus sequences reveals potential drug target sites of non-structural proteins. *Journal of General Virology*.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029.

DiMasi, J. A.; Hansen, R. W. & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. Em *Journal of Health Economics*, volume 22, pp. 151–185.

Dumais, S. (1992). Enhancing performance in latent semantic indexing (LSI) retrieval. Relatório técnico, Bellcore.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327–384.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition.* Fundamentals of Algorithms 4. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Flavin, R.; Peluso, S.; Nguyen, P. L. & Loda, M. (2010). Fatty acid synthase as a potential therapeutic target in cancer. *Future oncology London England*, 6(4):551–562.

Frantz, S. (2005). Drug discovery: Playing dirty. *Nature*, 437(7061):942–943.

Gajria, B.; Bahl, A.; Brestelli, J.; Dommer, J.; Fischer, S.; Gao, X.; Heiges, M.; Iodice, J.; Kissinger, J. C.; Mackey, A. J.; Pinney, D. F.; Roos, D. S.; Stoeckert, C. J.; Wang, H. & Brunk, B. P. (2008). Toxodb: an integrated toxoplasma gondii database resource. *Nucleic Acids Research*, 36(suppl 1):D553–D556.

Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X. & Jiang, H. (2008). Pdtd: a web-accessible protein database for drug target identification. *BMC Bioinformatics*, 9(104).

Geyer, J. A.; Prigge, S. T. & Waters, N. C. (2005). Targeting malaria with specific cdk inhibitors. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1754(1,2):160–170.

Ghofrani, H. A.; Osterloh, I. H. & Grimminger, F. (2006). Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nat Rev Drug Discov*, 5(8):689–702.

Goozner, M. (2004). *The $800 Million Pill : The Truth behind the Cost of New Drugs*. University of California Press.

Gunther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.; Gewiess, A.; Jensen, L.; Schneider, R.; Skoblo, R.; Russell, R.; Bourne, P.; Bork, P. & Preissner, R. (2008). Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research - Database issue*, 36:D919–D922.

Hajduk, P. J.; Huth, J. R. & Tse, C. (2005). Predicting protein druggability. *Drug Discovery Today*, 10(23-24):1675–1682.

Harland, L. & Gaulton, A. (2009). Drug target central. *Expert Opinion on Drug Discovery*, 4(8):857–872.

Hasan, S.; Daugelat, S.; Rao, P. S. S. & Schreiber, M. (2006). Prioritizing genomic drug targets in pathogens: Application to *Mycobacterium tuberculosis*. *PLoS Comput Biol*, 2(6):e61.

Heikkinen, L. S.; Kazlauskas, A.; Melen, K. & Wagner, R. (2008). Avian and 1918 spanish influenza a virus ns1 proteins bind to crk/crkl src homology 3 domains to activate host cell signaling. *The Journal of Biological Chemistry*, 283:5719–5727.

Heresco-Levy, U.; Ermilov, M.; Lichtenberg, P.; Bar, G. & Javitt, D. C. (2004). High-dose glycine added to olanzapine and risperidone for the treatment of schizophrenia. *Biological psychiatry*, 55(2):165–171.

Hirji, K. F.; Tan, S. J. & Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Stat Med*, 10(7):1137–53.

Holtze, M.; Saetre, P.; Erhardt, S.; Schwieler, L.; Werge, T.; Hansen, T.; Nielsen, J.; Djurovic, S.; Melle, I.; Andreassen, O. A.; Hall, H.; Terenius, L.; Agartz, I.; Engberg, G.; Jansson, E. G. & Schalling, M. (2011). Kynurenine 3-monooxygenase (kmo)

polymorphisms in schizophrenia: An association study. *Schizophrenia Research*, 127(1-3):270–272.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4(11):682–690.

Hopkins, A. L. & Groom, C. R. (2002). The druggable genome. *Nature reviews. Drug discovery*, 1(9):727–730.

Hopkins, A. L.; Mason, J. S. & Overington, J. P. (2006). Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology*, 16(1):127–136.

Hosmer, D. & Lemeshow, S. (2004). *Applied Logistic Regression*. Wiley Series in Probability and Statistics: Texts and References Section. Wiley.

Howe, E.; Holton, K.; Nair, S.; Schlauch, D.; Sinha, R. & Quackenbush, J. (2010). Mev: Multiexperiment viewer. Em Ochs, M. F.; Casagrande, J. T. & Davuluri, R. V., editores, *Biomedical Informatics for Cancer Research*, pp. 267–277. Springer US.

Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T. K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; de Castro, E.; Coggill, P.; Corbett, M.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Fraser, M.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; McMenamin, C.; Mi, H.; Mutowo-Muellenet, P.; Mulder, N.; Natale, D.; Orengo, C.; Pesseat, S.; Punta, M.; Quinn, A. F.; Rivoire, C.; Sangrador-Vegas, A.; Selengut, J. D.; Sigrist, C. J. A.; Scheremetjew, M.; Tate, J.; Thimmajanarthanan, M.; Thomas, P. D.; Wu, C. H.; Yeats, C. & Yong, S.-Y. (2012). Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(D1):D306–D312.

Huthmacher, C.; Hoppe, A.; Bulik, S. & Holzhutter, H.-G. (2010). Antimalarial drug targets in plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC Systems Biology*, 4(1):120.

Janga, S. C. & Tzakos, A. (2009). Structure and organization of drug-target networks: insights from genomic approaches for drug discovery. Em *Molecuar BioSystems*. RSCPublishing.

Jeh, G. & Widom, J. (2002). Simrank: A measure of structural-context similarity. Em *In KDD*, pp. 538–543.

Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M. & Hirakawa, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38:D355–D360.

Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J. & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25:197–206.

Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K. & Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature*, 462:175–181.

Khandekar, S. S.; Daines, R. A. & Lonsdale, J. T. (2003). Bacterial beta-ketoacyl-acyl carrier protein synthases as targets for antibacterial agents. *Curr Protein Pept Sci*, 4(1):21–29.

Kramer, R. & Cohen, D. (2004). Functional genomics to new drug targets. *Nat Rev Drug Discov*, 3(11):965–972.

Kridel, S. J.; Axelrod, F.; Rozenkrantz, N. & Smith, J. W. (2004). Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity. *Cancer Research*, 64(6):2070–2075.

Kridel, S. J.; Lowther, W. T. & Pemble IV, C. W. (2007). Fatty acid synthase inhibitors: new directions for oncology. *Expert Opinion on Investigational Drugs*, 16(11):1817–1829.

Kuhn, M.; von Mering, C.; Campillos, M.; Jensen, L. J. & Bork, P. (2008). Stitch: interaction networks of chemicals and proteins. *Nucleic Acids Research - Database issue*, 36:D684–D688.

Leahy, D. J. (1997). Implications of atomic-resolution structures for cell adhesion. *Annual Review of Cell and Developmental Biology*, 13:363–393.

Lee, D. D. & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. Em *In NIPS*, pp. 556–562. MIT Press.

Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X. & Jiang, H. (2006). Tarfisdock: a web server for identifying drug targets with docking approach. *Nucleic Acids Research*, 34(suppl 2):W219–W224.

Li, Z.-L.; Li, Q.-S.; Zhang, H.-J.; Hu, Y.; Zhu, D.-D. & Zhu, H.-L. (2011). Design, synthesis and biological evaluation of urea derivatives from o-hydroxybenzylamines and phenylisocyanate as potential fabh inhibitors. *Bioorganic & Medicinal Chemistry*, 19(15):4413–4420.

Light, D. W. & Warburton, R. (2011). Demythologizing the high costs of pharmaceutical research. *BioSocieties*, 6(1):34–50.

Loulergue, P.; Gaillard, R. & Mir, O. (2011). Interaction involving tadalafil and cyp3a4 inhibition by ritonavir. *Scand J Infect Dis*, 43(3):239–240.

Luo, H.; Chen, J.; Shi, L.; Mikailov, M.; Zhu, H.; Wang, K.; He, L. & Yang, L. (2011). Drar-cpi: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. *Nucleic acids research*, 39:W492–W498.

Marcolino, L.; Couto, B. & dos Santos, M. A. (2010). Genome visualization in space. Em Rocha, M.; Riverola, F.; Shatkay, H. & Corchado, J., editores, *Advances in Bioinformatics*, volume 74 of *Advances in Intelligent and Soft Computing*, pp. 225–232. Springer Berlin Heidelberg.

MATLAB (2010). *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

Maurer, T.; Zackheim, H.; Tuffanelli, L. & Berger, T. (1994). The use of methotrexate for treatment of psoriasis in patients with hiv infection. *Journal of the American Academy of Dermatology*, 31(2):372–375.

Mazumdar, J.; H. Wilson, E.; Masek, K.; A. Hunter, C. & Striepen, B. (2006). Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in toxoplasma gondii. *Proceedings of the National Academy of Sciences*, 103(35):13192–13197.

Menendez, J. A.; Vellon, L. & Lupu, R. (August 2005). Antitumoral actions of the antiobesity drug orlistat (xenical) in breast cancer cells: blockade of cell cycle progression, promotion of apoptotic cell death and pea3-mediated transcriptional repression of her2/neu (erbb-2) oncogene. *Annals of Oncology*, 16(8):1253–1267.

Menon, K.; Voorhees, A. S. V.; Bebo, B. F.; Gladman, D. D.; Sylvia Hsu, R. E. K.; Lebwohla, M. G. & Strober, B. E. (2010). Psoriasis in patients with hiv infection: From the medical board of the national psoriasis foundation. *Journal of the American Academy of Dermatology*, 62(2):291–299.

Métifiot, M.; Marchand, C.; Maddali, K. & Pommier, Y. (2010). Resistance to integrase inhibitors. *Viruses*, 2(7):1347–1366.

Meur, N. L. & Gentleman, R. (2011). Analyzing biological data using r: Methods for graphs and networks. Em *Bacterial Molecular Networks - Methods and Protocols*, volume 804 of *Methods in Molecular Biology*. Springer.

Miculka, C.; Tran, H. Q.; Meyer, T.; Heckeroth, A. R.; Baumeister, S.; Seeber, F. & Selzer, P. M. (2011). *Orlistat: A Repositioning Opportunity as a Growth Inhibitor of Apicomplexan Parasites?*, pp. 481–492. Wiley-VCH Verlag GmbH & Co. KGaA.

Nie, Z.; Perretta, C.; Lu, J.; Su, Y.; Margosiak, S.; Gajiwala, K. S.; Cortez, J.; Nikulin, V.; Yager, K. M.; Appelt, K. & Chu, S. (2005). Structure-based design, synthesis, and study of potent inhibitors of beta-ketoacyl-acyl carrier protein synthase iii as potential antimicrobial agents. *Journal of Medicinal Chemistry*, 48(5):1596–1609.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T. & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33+.

O'Driscoll, C. (2004). A virtual space odyssey. Em *Horizon Symposia*. Nature.

Okawa, Y.; Hideshima, T.; Ikeda, H.; Raje, N.; Vallet, S.; Kiziltepe, T.; Yasui, H.; Enatsu, S.; Pozzi, S. & Breitkreutz, I. (2008). Fatty acid synthase is a novel therapeutic target in multiple myeloma. *British Journal of Haematology*, 141:659–671.

Ondetti, M. A.; Rubin, B. & Cushman, D. W. (1977). Design of specific inhibitors of angiotensin-converting enzyme: New class of orally active antihypertensive agents. *Science*, 196(4288):441–444.

Pagano, M. & Gauvreau, K. (2004). *Principles of Biostatistics*. Statistics Series. Duxbury.

Perlman, L.; Gottlieb, A.; Atias, N.; Ruppin, E. & Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *Journal of computational biology a journal of computational molecular cell biology*, 18(2):133–145.

Projan, S. J. (2003). Why is big pharma getting out of antibacterial drug discovery? *Current Opinion in Microbiology*, 6(5):427–430.

Raman, K.; Kalidas, Y. & Chandra, N. (2008). *Model-Driven Drug Discovery: Principels and Practices*. jake Chen and Amandeep S. Sidhu.

Riggs, T. L. (2004). Research and development costs for drugs. *The Lancet*, 363(9404):184.

Roberts, R. J. (2001). PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):381–382.

Ruffolo, R. (2006). Why has r&d productivity declined in the pharmaceutical industry? *Expert Opinion on Drug Discovery*, 1(2):99–102.

Santner, T. J. & Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag.

Scannell, J. W.; Blanckley, A.; Boldon, H. & Warrington, B. (2012). Diagnosing the decline in pharmaceutical r&d efficiency. *Nat Rev Drug Discov*, 11(3):191–200.

Schlesselman, J. J. (1982). *Case-Control Studies Design, Conduct, Analysis*. Oxford University Press.

Schmidtke, P. & Barril, X. (2010). Understanding and predicting druggability. a high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858–5867.

Semba, J. (1998). Glycine therapy of schizophrenia; its rationale and a review of clinical trials. *Nihon Shinkei Seishin Yakurigaku Zasshi*, 18(3):71–80.

Sokal, R. R. & Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2).

Stuart, G. W.; Moffett, K. & Leader, J. J. (2002). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol*, 19(4):554–562.

Tonew, E.; Indulen, M. K. & Dzeguze, D. R. (1982). Antiviral action of dipyridamole and its derivatives against influenza virus a. *Acta Virol*, 26(3):125–129.

Tonew, M.; Laass, W.; Tonew, E.; Franke, R.; Goldner, H. & Zschiesche, W. (1978). Antiviral activity of dipyridamole derivatives. *Acta Virologica*, 22(4):287–295.

Tonew, M.; Tonew, E. & Mentel, R. (1977). The antiviral activity of dipyridamole. *Acta virologica*, 21(2):146–150.

Trouiller, P.; Olliaro, P.; Torreele, E.; Orbinski, J.; Laing, R. & Ford, N. (2002). Drug development for neglected diseases: a deficient market and a public-health policy failure. *The Lancet*, 359(9324):2188–2194.

Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G. & III, W. A. G. (2002). Prediction of structure and function of g protein-coupled receptors. *PNAS*, 99(20):12622–12627.

Weinstein, J. N. (2008). A postgenomic visual icon. *Science*, 319(5871):1772–1773.

Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B. & Hassanali, M. (2008). Drugbank: a knowledge base for drugs, drug actions and drug targets. *Nucleic Acids Research - Database issue*, 36:D901–D906.

Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. & Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34:668–672.

Witten, I. H. & Frank, E. (2011). *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Witty, L.; Steiner, F.; Curfman, M.; Webb, D. & Wheat, L. (1992). Disseminated histoplasmosis in patients receiving low-dose methotrexate therapy for psoriasis. *Arch Dermatol*, 128(1):91–93.

Xie, D.; Tropsha, A. & Schlick, T. (2000). An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-newton minimization. *Journal of Chemical Information Computer Sciences*, 40(1):167–177.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848.

Zhao, S. & Li, S. (2010). Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS ONE*, 5(7):e11764.

Zheng, C. J.; Han, L. Y.; Yap, C. W.; Ji, Z. L.; Cao, Z. W. & Chen, Y. Z. (2006). Therapeutic targets: Progress of their exploration and investigation of their characteristics. *Pharmacological Reviews*, 58(2):259–279.

Zhu, F.; Han, B.; Kumar, P.; Liu, X.; Ma, X.; Wei, X.; Huang, L.; Guo, Y.; Han, L.; Zheng, C. & Chen, Y. (2010). Update of ttd: Therapeutic target database. *Nucleic Acids Research - Database issue*, 38:D787–D791.

Zygmunt, M. (2010). Ontologies and agents for better information flow in logistics. Em *Logistics Systems and Management*, volume 6, pp. 135–148. Inderscience Enterprises Ltd.