USO DE FINGERPRINTS DE FARMACÓFOROS POTENCIAIS PARA COMPARAÇÃO DE SÍTIOS PROTÉICOS E LIGANTES ATIVOS.

FÁBIO MENDES DOS SANTOS

USO DE FINGERPRINTS DE FARMACÓFOROS POTENCIAIS PARA COMPARAÇÃO DE SÍTIOS PROTÉICOS E LIGANTES ATIVOS.

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

Orientador: Júlio César Dias Lopes

Belo Horizonte

Agosto de 2015

Santos, Fábio Mendes dos
Uso de Fingerprints de Farmacóforos Potenciais para Comparação de Sítios Protéicos e Ligantes Ativos. / Fábio Mendes dos Santos. — Belo Horizonte, 2015
xxxiii, 156 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Júlio César Dias Lopes

1. Bioinformática Estrutural. 2. Triagem Virtual. 3. Farmacóforo. 4. Fingerprint. I. Título.

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE

Fábio Mendes dos Santos

57/2015
entrada
1°/2011
CPF:
056.329.996-70

Às quatorze horas do dia **28 de agosto de 2015**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: **"Uso de Fingerprints de Farmacóforos Potenciais para Comparação de Sítios Prótéicos e de Ligantes Ativos"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Vasco Ariston de Carvalho Azevedo**, indicado pelo orientador do aluno, onde a ausência foi justificada devido a Residência Pós-Doutoral no exterior, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Vasco Ariston de Carvalho Azevedo	UFMG	283.191.225-49	APROVADO
Dra. Raquel Cardoso de Melo Minardi	UFMG	046.454.366-51	aprovada
Dr. Tiago Antônio da Silva Brandão	UFMG	03452839-37	Aprovado
Dr. Alex Gutterres Taranto	UFSJ	771.081.916-87	APROVADO
Dra. Célia Maria Corrêa	UFOP	22701184649	Aprovado

Pelas indicações, o candidato foi considerado: APROVADO
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 28 de agosto de 2015.

Dr. Vasco Ariston de Carvalho Azevedo- Presidente _____
Dra. Raquel Cardoso de Melo Minardi _____
Dr. Tiago Antônio da Silva Brandão _____
Dr. Alex Gutterres Taranto _____
Dra. Célia Maria Corrêa _____

Dr. Vasco Ariston de C. Azevedo
Prof. Titular e Coordenador do Programa de Pós Graduação em Bioinformática
ICB/UFMG

Dedico este trabalho principalmente aos familiares e amigos aos quais tive que me sacrificar de sua companhia no período do doutorado. Dedico à minha mãe, Laurinda, e à minha tia, Agripina, por sempre terem me servido de inspiração e de base para que eu conseguisse sempre almejar algo melhor da vida. Aos meus irmãos, Leonardo, Nilza, Aline e Larisse agradeço pelo apoio, incentivo e compreensão por minha ausência nos compromissos de família.

Agradecimentos

Grandes pessoas passaram pela minha vida desde que descobri que queria enveredar pelos caminhos da pesquisa científica, e agradeço a todas elas. Em especial, agradeço à minha primeira orientadora, Célia Maria Corrêa, por me mostrar que era possível aplicar computação à resolução de questões Biológicas. Agradeço também ao meu atual orientador, Júlio, pela paciência e aos colegas que passaram pelo NEQUIM [1] (Andrelly, Ramon e Bernardo) no tempo em que estive de laboratório. Também quero agradecer aos grandes amigos que fiz na Bioinformática que me propiciaram, além de diversão, momentos muito acalourados de discussão sobre assuntos científicos.

[1] Núcleo de Estudos em Quimioinformática

"Que ninguém se engane, só se consegue a simplicidade através de muito trabalho."

(Clarice Lispector)

Resumo

Nas últimas décadas, devido aos avanços tecnológicos, os bancos de dados ligados às áreas da saúde humana (como os de sequências genômicas, estruturas protéicas tridimensionais e pequenas moléculas) tiveram um grande crescimento. Tamanha foi a quantidade de informações disponibilizadas que hoje uma das grandes preocupações na pesquisa é como analisar toda a gama de dados disponíveis.

A correta interpretação dessas informações pode contribuir para o esclarecimento de mecanismos biológicos causadores de doenças, para a classificação e identificação de proteínas com atividades biológicas similares e o desenvolvimento de novos fármacos para doenças já conhecidas e as que ainda estão por surgir.

Atualmente, há uma grande quantidade de softwares disponíveis para esta finalidade, cada um com suas vantagens e desvantagens. Basicamente, a busca por estruturas químicas em um banco de dados e sua identificação com potencialmente ativas é chamada de Virtual Screening (VS), e pode ser feita tanto pelo uso de informações dos ligantes ativos (Ligand-Based Virtual Screening - LBVS) como dos alvos biológicos (Target-Based Virtual Screening - TBVS).

Este trabalho propõe a aplicação de metodologia de fingerprints de farmacóforos potenciais para a comparação de estruturas protéicas e de ligantes. Para isto estão sendo desenvolvidas no NEQUIM (Núcleo de Estudos em Quimioiformática), sediado no Departamento de Química da UFMG, as ferramentas PharmaSite (para análise de similaridade de alvos biológicos) e 3D-Pharma (para VS de ligantes). Além disso, propomos aqui uma nova metodologia de modelagem de dados, baseada em "bootstrap"e validação cruzada, que permite a geração de modelos mais robustos para uma dada base de dados.

Realizamos diversas análises e, nesses estudos ambas as ferramentas apresentaram bons resultados tanto em aplicações de cálculo de similaridade entre sítios quanto à recuperação de moléculas potencialmente ativas em um conjunto de ligantes. Entretanto,

há possibilidades de avanço nas metodologias utilizadas, principalmente se aplicarmos técnicas de modelagem de dados.

Palavras-chave: Triagem Virtual, Fingerprints, Farmacóforos.

Abstract

Due to technological advances in the last decades databases containing information related to human health areas (such as genomic sequences, three-dimensional protein structures and small molecules) experienced a great growth. The amount of available information is so hignt that today a major concern is how to analyze such plethora of data.

The correct interpretation of this information may contribute to the understanding of biological mechanisms of diseases, for classification and identification of proteins with similar biological activities and for the development of new drugs for treatment of known diseases and those who still will appear.

Currently, there are serveral softwares for this purpose, each one with its own advantages and disadvantages. Basically, the search in libraries of small molecules in order to identify substances which are most likely to bind to a drug target is called Virtual Screening (VS). It can use information from active ligands (Ligand-Based Virtual Screening - LBVs) or from biological targets (Target-Based Virtual Screening - TBVS).

This work introduces a pharmacophore fingerprints methodology to analyse structures of proteins and small ligands. We developed at NEQUIM (Nucleo de Estudos em Quimioinformática- UFMG) two softwares called PharmaSite (to calculate similarities between biological targets) and 3DPharma (for VS of small ligands). Furthermore, we propose a new methodology to build data models using bootstrap and cross-validation approaches.

Several analysis were performed and our tools showed good results in both calculation of similarity between the active sites as in applications for recovery potentially active molecules in ligand databases.

Keywords: Virtual Screening, Pharmacophore, Fingerprint.

Lista de Figuras

Lista de Tabelas

Lista Símbolos e abreviações

- 4D FAP_{OA}: 4D Flexible Atom Pairs (Optimal Assignment);
- ADMET: Absorção, Distribuição, Metabolismo, Excreção e Toxicidade;
- ADT: Autodock Tools;
- ALR2: Aldose Redutase;
- AMP_c: Adenosina Monofosfato Cíclico;
- AMBER: Assisted Model Building with Energy Refinement;
- ANVISA: Agencia Nacional de Vigilância Sanitária;
- AR: Receptor de Androgênio;
- ATP: Adenosina Tri-Fosfato;
- AUC: Area Under Curve;
- BEDROC: Boltzmann-Enhanced Discrimination of ROC;
- BLOSUM: BLOcks of amino acid SUbstitution Matrix;
- CDK2: Ciclina Dependente de Cinase 2;
- CF: Chemical Fingerprint;
- CGO: Chemical Gaussian Overlay;
- CGT: Chemical Gaussian Tanimoto;
- CNPEM: Centro Nacional de Pesquisa em Energia e Materiais;
- COX2: Ciclooxigenase-2;

- CPASS: Comparison of Protein Active Site Structures;
- DNA: Ácido Desoxirribonucleico;
- DUD: Directory of Usefull Decoys;
- ECFP:Extended Conectivity Fingerprints;
- EF: Enrichment Factor;
- EGFR: Receptor de Fator de Crescimento Epidérmico;
- EMBL: European Molecular Biology Laboratory;
- ExCVBA: Extensive Cross-Validation e Bootstrap Aplication;
- FCFP: Functional-Class Fingerprints;
- FDA: Food and Drug Administration;
- FLAP: Fingerprints for Ligands And Proteins;
- FPR: False Positive Rate;
- GA: Algoritmo Genético;
- GB: Gigabyte;
- GHz: Gigahertz;
- GMM: Modelo de Mistura de Gaussianas;
- GMPc: Guanosina Monofostato Cíclico;
- HIV: Vírus da Imudodeficiência Humana;
- HIVRT: Transcriptase Reversa de HIV-1;
- HF: Filtro Hierarquico;
- HTS: Hight-Throughput Screening;
- HTVS: Hight-Throughput Virtual Screening;
- IC: Construção incremental;
- IUBMB: International Union of Biochemistry and Molecular Biology;

- IUPAC: International Union of Pure and Applied Chemistry;
- JPS: Joint Pharmacophore Space;
- LBVS: Ligand-Based Virtual Screening;
- LMCS: Low-mode Conformational Sampling;
- LNBio:Laboratório Nacional de Biociências;
- logP: Coeficiente de partição (Logaritmo);
- MC: Monte Carlo;
- MDDR: MDL Drug Data Report;
- MMFF: Merck Molecular Force Field;
- MOE: Molecular Operating Environmentmoe;
- NEQUIM: Núcleo de Estudos em Quimioinformática - UFMG;
- OA: Optimal Assignment;
- P38: Cinase Protéica Ativada por Mitogênio 14;
- PDB: Protein Data Bank;
- PDE5: Fosfodiesterase V;
- P&D: Pesquisa e Desenvolvimento;
- PF: Pharmacophore Fingerpirnt;
- PGH: Projeto Genoma Humano;
- PIR: Protein Information Resource;
- $PPAR_{\gamma}$: Receptor Ativado por Proliferador de Peroxissomo γ;
- PPP: Potential Pharmacophore Points;
- PRF:Protein Research Foundation;
- QSAR: Quantitative Structure-Activity Relationship;
- QSPR: Quantitative Structure Property Relationship;

- REF: Relative Enrichment Factor;
- RMSD: Root Mean Square Deviation;
- RENAMA: Rede Nacional de Métodos Alternativos;
- ROC: Receiver-Operator Characteristic;
- ROCE: Receiver Operator Characteristics Enrichment;
- SAR: Structure-Activity Relationship;
- SLR: Sum of logarithms of ranks;
- SVM: Suport Vector Machines;
- SVD: Singular Value Decomposition;
- TBVS: Target Based Virtual Screening;
- TPR: True Positive Rate;
- UFMG: Universidade Federal de Minas Gerais;
- VS: Virtual Screening;
- WOMBAT: World of Molecular BioAcTivity

.

Expressões latinas:

- *ad hoc*: para isto, para o caso específico
- *a priori*: admitido como evidente, independe da experiência
- e.g., *exempli gratia*: por exemplo
- i.e., *id est*: isto é, ou seja
- *in populo*: estudos em populações
- *in vitro*: ensaios laboratoriais
- *in vivo*: estudos em seres vivos, incluindo estudos clínicos
- *in silico*: ensaios computacionais

Sumário

Capítulo 1

Introdução

Estamos vivendo em uma época de grandes transformações onde, com frequência, nos são anunciadas novas descobertas científicas e tecnologias inovadoras. Esse cenário, além de nos possibilitar certos confortos na vida diária, também têm beneficiado praticamente todas os ramos da ciência e, como iremos abordar mais adiante, isso foi de grande importância para o desenvolvimento das ciências da vida nas últimas décadas.

Se pensarmos em como eram realizados experimentos científicos na década de 1960, apenas algumas décadas atrás, podemos perceber uma rápida evolução dos equipamentos e das técnicas aplicadas ao processo científico. Com aparelhos rudimentares (em comparação com os atuais) os cientistas eram obrigados à realização de um trabalho praticamente manual dentro dos laboratórios de pesquisa e, por isso, alguns experimentos que hoje podem ser feitos em poucas horas poderiam levar meses, ou mesmo anos, para serem finalizados naquela época. Além disso, não só a rapidez desses novos equipamentos mas, também, o desenvolvimento de novas técnicas experimentais mais eficientes vêm revolucionando a forma de se fazer ciência no século XXI.

Um fator primordial a esse processo certamente tem sido a evolução constante da tecnologia aplicada no desenvolvimento e construção dos computadores. Fazendo com que eles sejam peça essencial ao processo científico atual, tanto na análise de resultados como na composição de certos equipamentos ultramodernos. Hoje, eles são capazes processar milhares de informações por segundo, sendo que, mesmo os pequenos computadores de mesa podem se tornar poderosas ferramentas nas mãos cientista contemporâneo.

Além disso, na década de 1990 um acontecimento ajudou a promover um grande incentivo às ciências da vida. Esse acontecimento foi a implementação de um audacioso

projeto conhecido como Projeto Genoma Humano (PGH). Este tinha o objetivo de, em 15 anos, obter o sequenciamento completo do DNA humano e de algumas outras espécies consideradas modelo para estudos genômicos como coelho, camundongo, algumas bactérias e leveduras (1). Essa proposta foi abraçada por importantes cientistas de todo o mundo que acreditavam que o esforço coletivo, se não fosse capaz de alcançar a sua meta, ao menos poderia gerar uma grande quantidade de dados valiosos para a comunidade científica.

Esse projeto foi iniciado com a criação de grandes centros de pesquisa em países desenvolvidos e, mais tarde, o conhecimento e a tecnologia foram disseminados pelo globo de forma mais homogênea, a partir de acordos internacionais e interesses locais na participação em projetos de sequenciamento e/ou montagem de genomas e elucidação de estruturas proteicas. Acredita-se que tenham sido investidos cerca de 50 bilhões de dólares nesse projeto (1) que, em 2001, culminou na publicação do genoma humano (2) e, em 2003, foi oficialmente finalizado. Entretanto, não apenas apenas técnicas experimentais mais modernas para sequenciamento de genoma foram desenvolvidas nessse período como também, equipamentos para auxiliar os cientistas nos laboratórios e ferramentas para a análise e tratamento de dados. Isso mudou algumas concepções no modo de se pensar e fazer ciência, uma vez que os experimentos passaram a ser realizados em uma escala quase que industrial (1) e, por isso, à partir dos anos 1990, bancos de dados como GenBank (3), Swiss-Prot (4) e PDB (5) tiveram um aumento praticamente exponencial da quantidade de dados depositados (**Figura 1.1**).

Paralelamente a tudo isso, a disponibilidade de equipamentos e técnicas mais modernas e o acesso a maiores informacões de estruturas proteicas e de sequencias genômicas, também possibilitou às industrias químicas e farmacêuticas o desenvolvimento de novos métodos pequisa para o desenvolvimento de moléculas com potencial de atividade frente ao novos alvos biológicos que estavam sendo descobertos. Uma enorme quantidade dessas moléculas foi estudada muita informação foi gerada a respeito de suas propriedades, efeitos biológicos e interaçãos com alvos específicos. Assim, também houve um grande crescimento na quantidade dados de moléculas depositadas nas grandes bases de pequenos ligantes, algumas dessas bases possuem hoje informações de milhões de moléculas (**Tabela 1.3**).

Por tudo isso, amplas possibilidades foram abertas em vários ramos da ciência, principalmente para aqueles ligados à saúde, que vem a ser uma das grandes preocupações do homem moderno. Atualmente, pesquisadores em todo o mundo buscam novas formas de promover o tratamento racional da ampla gama de informações biológicas e

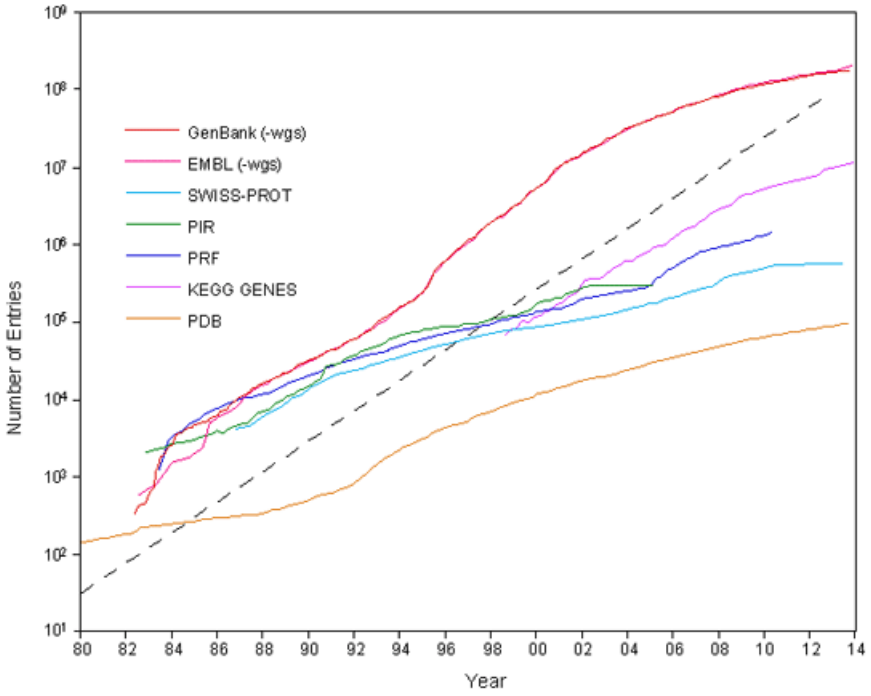


Figura 1.1: Estatísticas de crescimento de várias bases de dados de genes e proteínas. A linha preta, hachurada, representa a curva exponencial padrão da função $y = 10^x$. Fonte: `http://www.kanehisa.jp/en/db_growth.html`

químicas, de forma a viabilizar a produção de resultados rápidos e confiáveis. Visando abordar esse desafio surgiu a Bioinformática, uma ciência por si só de caráter multidisciplinar que, aliando conhecimentos de Biologia, Química, Ciências da Computação e Estatística, veio como uma alternativa para o entendimento da nova forma de se fazer ciência que vem recentemente sendo desenvolvida. Basicamente o objetivo da Bioinformática é viabilizar, através de programas de computador, a análise de dados biológicos e químicos digitalizados promovendo o entendimento de importantes mecanismos e vias metabólicas dos seres vivos.

1.1 Dados x Informação

Como vimos na seção anterior hoje há disponibilidade de uma enorme quantidade de dados de processos biológicos e químicos. Para a Bioinformática essa informação é muito valiosa pois permite a realização de extração de padrões e inferência de correlação com mecanismos biológicos reais. Entretanto, o fato de ter acesso ao dado e conseguir extrair algum conhecimento deles é algo muito diferente.

Os conceitos de extração de conhecimento à partir de dados vêm sendo delineados

desde a antiguidade mas, na década de 1930 o filósofo americano T. S. Eliot escreveu algumas frases em um poema chamado "*Choruses from The Rock*", que o tornaram um referência comum, sendo considerado um dos primeiros a abordar esse assunto. Hoje, esses conceitos, que são comumente representados por uma pirâmide e, por isso, conhecidos como "Teoria da pirâmide do conhecimento"(**Figura 1.2**), resumem o tratamento de dados em um processo de quatro estágios:

- gerar, possuir e armazenar dados;
- extrair informações relevantes;
- correlacionar as informações com fatos reais, gerando conhecimento;
- fazer com que esse conhecimento seja inquestionável (sabedoria).

Figura 1.2: Pirâmide do conhecimento.

Não se vê a menção ao último estágio (sabedoria) quando se fala em mineração ou tratamento de dados em Bioinformática. Talvez esse fato se deva à dificuldade de obtenção ou à própria geração dos dados, nem sempre confiável. Hoje em dia os grandes problemas em bases de dados biológicos são os erros presentes em sequências ou estruturas, redundâncias ou divergências de informações quando se realiza a análise de bases de dados diferentes. Além disso, existem outros fatores complicadores como as dificuldades de análise ou para a realização de cálculos decorrentes do fato de algumas bases serem caras ou necessitarem de ferramentas pagas e sofisticadas ou, ainda, por estarem vinculadas a plataformas nas quais poucos têm o domínio para realizar a exploração.

Essas dificuldades são consideradas normais uma vez que a própria Bioinformática é uma ciência recente e ainda em desenvolvimento, apresentando poucos paradigmas já estabelecidos. Se realizarmos uma análise rápida do perfil dos profissionais que trabalham nessa área é fácil identificar que a maioria é oriunda das Ciências da Computação ou da Biologia (principalmente genética e bioquímica). No entanto, diferentes áreas do conhecimento vêm percebendo a necessidade de entender esse tipo de dados e absorver os conhecimentos desenvolvidos na Bioinformática. Por isso, atualmente vários profissionais como químicos, farmacêuticos, médicos, físicos e matemáticos entre outros vem se aprofundando nos estudos de Bioinformática. Isso vem demonstrar que atualmente já é de grande importância o domínio desse conhecimento e que muitas esperanças estão sendo depositadas na Bioinformática para a resolução de problemas ligados à saúde do homem.

Contudo, a diversidade de profissionais nessa área reflete a complexidade envolvida no tratamento de dados biológicos, uma vez que cada um deles tem conhecimentos *a priori* que são de extrema relevância para o entendimento de informações extraídas dessas bases de dados. Além da quantidade de dados já ser um enorme problema, a sua natureza e diversidade torna-os extremamente difíceis de serem tratados. Por isso, uma das principais dificuldades para o bioinformata é saber o que coletar e como isso deve ser analisado.

Além disso, o fato de ser multidisciplinar apresenta outra dificuldade implícita, que é a forma de exposição dos resultados. Isso ocorre pelo fato de um resultado poder ser apresentado de diferentes maneiras e, pode acontecer de um mesmo resultado apresentar informações igualmente relevantes, mas diferentes para cada tipo de profissional, dependendo de sua origem. Como exemplo, na apresentação de um programa para análise de proteínas o informata pode dar mais atenção ao algoritmo enquanto o biólogo pode estar mais interessado nas proteínas que apresentaram atividade em determinado evento biológico e o estatístico se ponha e estudar a relevância desses dados. Por tudo isso, podemos dizer que o perfil do profissional do futuro seja o mais diversificado possível, pois o entendimento dessas divergências o faz entender a necessidade de acumular conhecimento necessário para perceber todas as nuances apresentadas nessa área.

Ainda, um fato que deve ser levado em consideração é o tempo despendido para a aplicação de uma abordagem de tratamentos de dados, uma vez que um dos principais objetivos dessas análises de dados é possibilitar um direcionamento para agilizar os experimentos de bancada. Assim, análises que tenham uma precisão elevada, mas que demandem muito tempo computacional, podem ser preteridas em relação a outras

abordagens um pouco menos precisas, mas que permitam análises de uma quantidade de dados maior.

Por fim, e esse talvez seja o ponto mais claro, mas que representa maior dificuldade, é que os resultados encontrados devem apresentar significância biológica. Isto é necessário para a identificação correta de entidades biológicas e/ou químicas e para a previsão de correlação com outras entidades às quais serão aprofundadas em estudos posteriores. È importante ressaltar a importância dessa correlação, pois não será útil um trabalho, mesmo que gere bons resultados com gráficos e tabelas pomposas, se o que for apresentado não tiver correlação com eventos biológicos ou não puder ser aplicado de forma a elucidar mecanismos de uma via metabólica ou a modulação de um evento patogênico, por exemplo.

1.2 "Virtual Screening": Triagem Virtual

Para a indústria farmacêutica a grande quantidade de dados disponíveis foi estimulante, mas também levou as empresas a investirem mais em pesquisa e desenvolvimento na busca de uma melhor utilização da grande quantidade de informações, seja de sequências genômicas e proteicas ou de informações estruturais sobre alvos biológicos e complexos supramoleculares. Com isso novas técnicas de previsão de atividade biológica foram desenvolvidas, otimizadas ou reinventadas levando o processo de seleção de novas moléculas para o desenvolvimento de fármacos a alcançar um novo estágio onde, devido à quantidade e complexidade dos cálculos necessários, passou a ser evidente a necessidade de auxílio de ferramentas computacionais.

Hoje, a principal aplicação das grandes bases de dados para o desenvolvimento racional de fármacos é a Triagem Virtual (*"Virtual Screening"*, VS), que pode ser definida como a busca computacional de pequenas moléculas, com atividades biológicas e especificidades desejadas, em grandes bases de dados. O VS passou a ser muito utilizado na década de 1990 com o objetivo de auxiliar o Hight-Througput Virtual Screening, HTVS, que deriva de *Hight-Throughput Screening* (HTS), que é uma metodologia experimental utilizada para a identificação de pequenas moléculas com certa afinidade por um determinado alvo biológico em uma biblioteca de compostos (chamada de "*library*", que é uma grande coleção de compostos que foram adquiridos ou adicionados com o tempo por terem sido obtidos por protocolos particulares de síntese ou que compartilham propriedades fisioquímicas interessantes). As grandes empresas farmacêuticas chegam a ter bibliotecas de compostos com milhões de moléculas

O HTS é um procedimento reconhecidamente rápido e relativamente eficiente, entretanto alguns fatores inerentes á sua metodologia prejudicam a sua execução. Apesar de ser totalmente automatizado (os testes em escala reduzida são feitos por robôs através de microtitulações simultâneas em bandejas que podem conter milhares de micropoços) pode ser necessária a realização milhões de testes em um só HTS demandando vários meses para terminar. Além disso, ela é uma técnica muito cara por depender de equipamentos ultrassofisticados e, ainda, o armazenamento dos compostos em condições satisfatórias para a realização do experimento é extremamente difícil, além de ser muito dispendioso. Ainda, algumas condições dos ligantes como a baixa estabilidade estrutural ou insolubilidade podem comprometer os resultados do experimento (6), gerando muitos falsos negativos e positivos.

Já o VS tende a ser mais econômico e os resultados podem ser produzidos em alguns dias, devido ao uso de computadores superpotentes. Além disso, a biblioteca de compostos virtual construída para o VS pode conter além das moléculas reais, presentes na biblioteca de compostos da empresa, outras substâncias que teoricamente podem ter algum efeito sobre um alvo e que a empresa ainda não tenha sintetizado. Apesar disso, o HTS tende a identificar moléculas ativas com uma diversidade estrutural muito maior uma vez que o VS é direcionado pelo conhecimento anterior (a priori), o que não impede que os resultados de HTS sejam somados ao VS promovendo uma otimização do processo racional de descobrimento de fármacos.

A grosso modo, o VS pode ser dividido em duas categorias principais, cujas abordagens dependem do conhecimento da estrutura tridimensional (3D) do alvo biológico (TBVS, *"Target Based Virtual Screening"* ou, Triagem Virtual Baseada em Estruturas Proteicas), ou das estruturas (2D ou 3D) de ligantes ativos conhecidos(LBVS, *"Ligand Based Virtual Screening"* ou Triagem Virtual Baseada em Ligantes). A **Tabela 1.1** mostra, de acordo com Klebe 2006 (7), uma sugestão das metodologias a serem escolhidas de acordo com o tipo e quantidade de informação disponível para a realização do VS.

A **Tabela 1.1** mostra que as metodologias a serem aplicadas são propostas de forma lógica. Assim, se há informações de ligantes e dos alvos é sugerido uma TBVS através de programas de ancoragem, entretanto mesmo nesses casos pode-se usar o LBVS se houverem muitos ligantes. Se há somente informação dos alvos conhecida, a única opção disponível é TBVS, podendo ser realizados estudos de novo para prever através de projeção na cavidade do sítio ativo as características estéricas e químicas necessárias a uma molécula para realizar a formação do complexo proteína-ligante e,

posteriormente, a realização estudos de ancoragem molecular. Caso só se tenham informações dos ligantes, tal como no exemplo anterior, o lógico é aplicação de estudos de LBVS nas suas mais amplas variações de metodologia, que será escolhida de acordo com a quantidade de informação contida nas bases. Caso sejam poucos ligantes estudos mais caros computacionalmente podem ser aplicados (como os de sobreposição estrutural ou dinâmica molecular). Caso sejam muitas moléculas será necessário escolher uma metodologia que permita a simplificação da informação molecular para a agilização dos cálculos. Com essa finalidade também podem ser aplicados filtros para o descarte precoce de moléculas potencialmente inativas. Esses filtros podem ser construídos com base na fórmula molecular (de acordo com a presença de átomos específicos ou quantidade destes), com subestruturas ou farmacóforos (toxicóforos, por exemplo). Por último, caso não se tenha nem a estrutura do alvo, nem as dos ligantes, é certo que não será possível aplicação de VS e o HTS é a única saída para a identificação de novas substâncias bioativas.

Tabela 1.1: Metodologia de triagem virtual a ser utilizada dependendo do tipo e da quantidade de informação disponível (7)

	Ligante(s) Conhecido(s)	**Ligante(s) desconhecido(s)**
A estrutura do alvo biológico é conhecida ou se dispõe da estrutura de um homólogo próximo	Triagem Virtual Baseada nos Alvos Biológicos: Ancoragem proteína-ligante	Triagem Virtual Baseada nos Alvos Biológicos utilizando metodologia de novo: Ancoragem proteína-ligante
Estrutura do alvo é desconhecida	Triagem Virtual Baseada nos ligantes: Poucos ligantes: - análise de similaridade Muitos ligantes: - análise de farmacóforos	VS não pode ser aplicado

1.3 Triagem Virtual Baseada em Alvos Biológicos

A técnica de triagem virtual onde se utiliza do conhecimento estrutural do alvo biolológico é chamada de Triagem Virtual Baseada em Alvos Biológicos (do inglês, "Target based virtual screening- TBVS). Dentre as técnicas de TBVS destaca-se o *"Docking Molecular"* ou Ancoragem Molecular (8) (9) (10) que, apesar de sua origem remontar à década de 80, passou a ser muito utilizado partir da década de 1990 devido ao aperfeiçoamento dos métodos, à crescente disponibilização de dados de estruturas de alvos e ao gradativo desenvolvimento computacional experimentado.

A Ancoragem Molecular ("Molecular Docking") consiste em procedimentos computacionais que tentam predizer o modo de ligação e a afinidade de um ligante com um alvo receptor. Cerca de 20 anos atrás, os programas de ancoragem molecular estavam na linha de frente das ferramentas computacionais para auxiliar a descoberta de fármacos. Poucos anos depois, se tornou claro que as funções de ranqueamento (funções de "scoring") usadas não eram capazes de predizer, de forma acurada, a energia livre de ligação ou mesmo de ordenar corretamente uma lista de compostos em função da afinidade prevista. Apesar desse desapontamento ferramentas de ancoragem molecular e abordagens computacionais continuam sendo utilizados no desenvolvimento e visualização de hipóteses que guiam os experimentos por novos ligantes. Entretanto, uma recente mudança de paradigmas no desenvolvimento racional de fármacos - promovendo uma mudança para respostas do tipo sim/não ao invés de predições quantitativas - tem conduzido algoritmos com abordagens baseadas em estrutura de volta aos holofotes científicos (11).

Dentre os principais programas disponíveis para ancoragem molecular estão hoje o Autodock, Glide, GolD, Surflex. Seus pacotes completos são geralmente pagos para as indústrias, mas possuem licensas acadêmicas gratuitas ou mais baratas fornecidas aos pesquisadores e cientistas para serem usados em seus experimentos de VS. A **Tabela 1.2** relaciona estes programas, bem como suas vantagens e desvantagens e o tipo de algoritmo utilizado, essa tabela foi apresentada por BIELSKA, 2011 (6) compilando os comentários e resultados publicados por vários outros pesquisadores.

Tabela 1.2: Lista de programas mais utilizados para a ancoragem proteína ligante com suas principais forças, fraquezas e os casos recentes onde foram aplicados com sucesso. GA. Algoritmo Genético, HF filtro hierárquico, IC. Construção incremental, MA algoritmo de similaridade, MC. Montet Carlo.

Programa e algoritmo	**Vantagens**	**Desvantagens**	**Exemplo de recentes de sucesso com utilização de VS**
AutoDock e AutoDock Vina (Morris et al., 1998 (12); Osterberg et al., 2002 (13); Trott and Olson, 2010 (14)) - GA	• Pequenos ligantes • Sítios de ligação muito extensos • Grau de liberdade disponibilizado	• Manter níveis bons de acurácia para ligantes altamente flexíveis. • Baixa velocidade.	Glutamate Transporter 1 (GLT1) inhibitors (Luethi et al., 2010) (15) Cdc25 phosphatase inhibitors (Park et al., 2008) (16) D-Ala:D-Ala ligase inhibitors (Kovac et al., 2008) (17) Cyclodextrin-based receptors (Steffen et al., 2007) (18)

Continua na próxima página...

Tabela 1.2 – ... ***Continuação***

Programa e algoritmo	Vantagens	Desvantagens	Exemplo de recentes de sucesso com utilização de VS
DOCK (Ewing et al., 2001 (19); Kuntz et al., 1982 (20); Lang et al., 2009 (21); Moustakas et al., 2006 (22); Oshiro et al., 1995 (23)) - IC	• Sítios de ligação pequenos • Cavidades abertas ao meio externo • Ligantes hidrofóbicos • Grau de liberdade disponibilizado	• Manter bons resultados de acurácia para ligantes muito flexíveis ou muito polares	Hepatitis C virus helicase inhibitors (Chen et al., 2009a (24)) SARS-CoV 3C-like proteinase inhibitors (Liu et al., 2005 (25)) Cyclooxygenase (COX-2) inhibitors (Mozziconacci et al., 2005 (26))
FlexX (Rarey et al., 1996 (27)) - IC	• Sítios de ligação pequenos • Ligantes hidrofóbicos pequenos	• Ligantes muito flexíveis	Bacterial NAD synthetase inhibitors (Moro et al., 2009 (28)) Lymphoid phosphatase inhibitors (Wu et al., 2009 (29)) RNA polymerase inhibitors (Kim et al., 2008 (30)) ATP phosphoribosyltransferase (HisG) inhibitors (Cho et al., 2008 (31)) Human histamine H4 receptor ligands (Kiss et al., 2008 (32))
Glide (Friesner et al., 2004 (33)) - HF+MC	• Ligantes flexíveis • ligantes hidrofóbicos pequenos	• ranqueamento de ligantes muito polares • baixa velocidade	Liver X receptor modulators (Cheng et al., 2008 (34)) HIV-1 reverse transcriptase inhibitors (Barreiro et al., 2007 (35))
GOLD (Verdonk et al., 2003 (36); Verdonk et al., 2005 (37)) - GA	• sítios de ligação pequenos • ligantes hidrofóbiocos pequenos • sítios de ligação internos	ranqueamento de ligantes polares ou quando a ancoragem é realizada em sítios muito grandes	HIV-1: CD4-gp120 binding inhibitors (Lalonde et al., 2011 (38)) Serotonin 5-HT(7)R antagonists (Kurczab et al., 2010 (39)) Non-peptide β-secretase inhibitors (Xu et al., 2010 (40)) Sarco/endoplasmic reticulum calcium ATPase inhibitors (Deye et al., 2009 (41)) Trypanosoma cruzi transsialidase inhibitors (Neres et al., 2009 (42))
Surflex (Jain, 2003 (43); Jain, 2007 (44)) - IC+MA	• Cavidades abertas • sítios de ligação muito grandes • ligantes muito flexíveis	• baixa velocidade para ligantes muito grandes	Triple helical DNA intercalators (Holt et al., 2009 (45)) ErmC methyltransferase inhibitors (Feder et al., 2008 (46)) Hepatitis C virus NS5B polymerase inhibitors (Musmuca et al., 2010 (47))

Por não ser alvo de estudo neste projeto apresentamos uma visão mais detalhada a respeito dos algoritmos aplicados nos métodos de TBVS no Apêndice (**Item A**).

1.4 Triagem Virtual Baseada em Ligantes

A utilização das informações dos ligantes ativos para a identificação de novas substâncias potencialmente ativas é chamada de Triagem Virtual Baseada em Ligantes ("Ligand Based Virtual Screening- LBVS). Neste ramo, o número de estruturas disponíveis é muito grande, o que aumenta a dificuldade de análise e o desenvolvimento de modelos válidos para representar classes de moléculas ou moléculas capazes de promover certo tipo de efeito biológico.

A quantidade de informações sobre ligantes se tornou maior à partir do início da década de 1990, quando novas técnicas de HTS (Hight-Throughput Screening) foram desenvolvidas, levando à uma geração massiça de dados. Hoje, essas informações sobre ligantes também estão armazenadas em grandes bases de dados e repositórios como, por exemplo, PubChem (48), ChEMBL (49), ChemSpider (50), MDDR(MDL Drug Data Report) (51), WOMBAT (World of Molecular BioAcTivity) (52), sendo, as duas últimas são bases de dados comerciais. Contudo, tal fato certamente corroborou para o desenvolvimento das técnicas LBVS, que usam somente dados de ligantes ativos, principalmente pequenas moléculas, em suas análises.

As bases de dados de pequenas moléculas apresentam um número bem maior de entidades se comparadas às bases de dados de alvos biológicos. As diferenças de tamanho são tão destoantes que nem o fato de as tecnologias para elucidação de estruturas proteicas ter evoluído nos últimos 20 anos foi suficiente para, sequer equiparar à quantidade de entidades presentes nas bases de pequenas moléculas (**Tabela 1.3**). No dia 31 de julho de 2015 o PDB possuía 110.790 estruturas depositadas, enquanto o PubChem possuía cerca de 68,4 milhões de compostos únicos e, destes, cerca 2 milhões possuem dados de ensaios biológicos, sendo que 1,1 milhões de compostos apresentam algum tipo de atividade relevante. O ChEMBL tem aproximadamente 1,5 milhões de compostos únicos e 13,5 milhões de medições de bioatividade. O ChemSpider tem aproximadamente 34 milhões de compostos.

Apesar de ser fácil a identificação de que a quantidade de informações disponíveis sobre ligantes é muito maior que a de alvos, isso não reflete a quantidade de métodos e programas para tratar as pequenas moléculas, pois, segundo Ripphausen, Nisius e Bajorath,2011 (54), existem aproximadamente três vezes mais técnicas baseadas em

Tabela 1.3: Principais bases de dados de pequenos ligantes. Adaptada de Kalyaanamoorthy e Chen, 2011 (53). Sim* = informação vinda de outras bases de dados ligadas

Base de dados	Nº de entidades	Busca por subestruturas	Formatos de arquivos	Busca de similaridade	Descritores Moleculares	Dados Experimetais
Pubchem	> 68 Million(C) > 196 Million(S)	Sim	Smiles, SDF, 2D, 3D	Sim	Sim	Sim
Drug Bank	Approx. 7800	Sim	Smiles, SDF, 2D, 3D, Mol, PDB	Sim	Sim	Sim
KEGG Ligand	17,402	Sim	Mol, KCF	Sim	Sim	Sim
Chem DB	>600,000	Sim	Mol, SDF	Sim	Sim	Não
ChemSpider	> 34 Million	Sim	Mol, Smiles, SDF, SKC, CDX	Sim	Sim*	Sim*
BindingDB	270,000	Sim	SDF, PDB, Mol, Smiles	Sim	Sim	Sim
Zinc	>6 Million	Sim	Smiles, Mol2, SDF	Sim	Sim*	Sim*
ChEMBL	Aprox 1.5 million	Sim	Smiles, sdf	Sim	Não	Sim

TBVS que em LBVS disponíveis. Por outro lado, em estudos prospectivos, em média, as técnicas baseadas em ligantes ativos identificam novas moléculas bioativas mais potentes que as identificadas pelas abordagens baseadas em alvos biológicos.

Ainda não é possível eleger qual abordagem é mais eficiente entre TBVS e LBVS. Na realidade, cada uma delas oferece vantagens e desvantagens a serem consideradas. No LBVS, por exemplo, a utilização de representações simplificadas dos ligantes o torna, geralmente, mais rápido que o TBVS e passível de ser aplicado a grandes bases de dados. Entretanto, os resultados tendem a ser enviesados, favorecendo a identificação de estruturas semelhantes aos modelos de compostos ativos utilizados, principalmente quando se usam métodos baseados nas estruturas 2D dos ligantes. Já no TBVS, há um número maior de softwares disponíveis, com metodologias mais complexas sendo empregadas, e os compostos potencialmente ativos recuperados têm maior diversidade estrutural. Entretanto, o custo computacional é geralmente é muito alto, demandando alto tempo de processamento e, nem sempre, com a eficiência requerida (55). Por estes motivos, alguns cientistas buscam formas de aliar os aspectos positivos das duas

metodologias, combinando informações de receptores e ligantes (por exemplo, o FLAP Fingerprints for Ligands and Proteins (55) (56) (57)), e os resultados parecem apresentar aumento de desempenho (58) em relação a estudos isolados de TBVS e LBVS.

1.4.1 Classificação das metodologias para LBVS

Baseado na quantidade de dados utilizada e no grau de complexidade desejado abordado para a resolução do problema de LBVS, vários aspectos podem ser utilizados para classificar essas metodologias de forma mais genérica. Uma das classificações mais comuns diz respeito à quantidade de dimensões (1D, 2D ou 3D) (6) (BIELSKA, 2011) consideradas no momento da coleta ou na forma de abstração dos dados das estruturas da base de ligantes (**Figura 1.3**). Os descritores 1D abrangem as propriedades que podem ser extraídas da fórmula molecular sem a necessidade de consultas às estruturas 2D ou 3D das moléculas. São exemplos desses descritores o peso molecular, número e tipo de átomos, refratividade molar e logP.

Os descritores que necessitam da estrutura molecular em duas dimensões para serem calculados são chamados de 2D. Basicamente, há dois tipos desses descritores: os índices topológicos (caracterizado por um valor único que representa uma característica molecular, como forma ou volume molecular por exemplo) e os descritores estruturais (caracterizam a substrurura química através dos fragmentos estruturais presentes em uma molécula, sendo mais frequentemente representados por grafos 2D ou vetores binários) (59). Esse último descritor costuma ser mais facilmente encotrado na literatura aplicado a estudos de VS.

Os descritores que possuem a maior dificuldade de implementação são os discritores 3D, isso acontece porque eles são contruídos levando em consideração características moleculares mais complexas como, por exemplo, a flexibilidade molecular e a possibilidade de formação de diferentes confórmeros. Dentre as propriedades utilizadas na construção desse tipo de descritor podemos citar superfícies de moléculas, distâncias inter-atômicas, campos eletrostáticos e também farmacóforos 3D, dentre outros (59).

Entretanto, apesar de existirem diferentes tipos de descritores que representam diferentes propriedades, com suas respectivas complexidades, estudos indicam que não há um descritor que apresente resultados superiores em todas as possíveis aplicações de estudos triagem virtual (60). Assim, é comum aos cientistas, visando alcançarem seus objetivos, empregarem conjuntos desses descritores na representação de moléculas (59) .

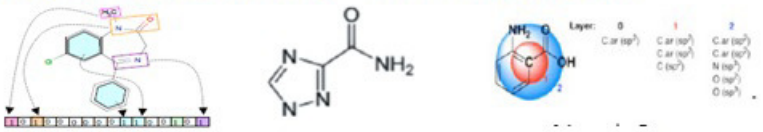
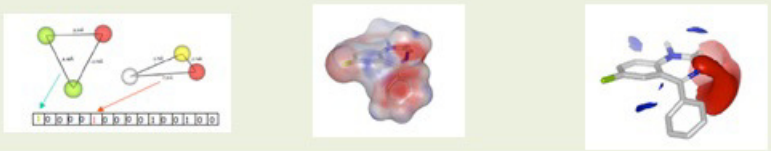
Descritor	Exemplos
1D	Fórmula Molecular, Peso Molecular, Número de Átomos, Número de anéis, Área da Superfície Polar, LogP
2D	Distancia entre pares de átomos, distância pares de farmacóforo 2D fingerprint, substruturas, 2D circular fingerprint 
3D	Tuplets de Farmacóforos, Forma, Campos Eletrostáticos 

Figura 1.3: Descritores utilizados em VS. Adaptada de (61).

Outra forma de classificar as técnicas de VS em ligantes diz respeito ao tipo de algoritmo utilizado para a predição de possíveis ativos em uma base de dados. Dentre eles estão os que realizam cálculos de similaridade, os algoritmos quantitativos e os baseados em aprendizagem de máquina.

1.4.2 Algoritmos baseados em buscas por similaridade

O princípio básico deste algoritmo vem da química medicinal que diz que moléculas com estruturas semelhantes devem compartilhar propriedades estéricas e físico-químicas que os permitem ter a probabilidade maior que a randômica de desempenhar funções biológicas semelhantes (62) (o que nem sempre é verdade).

Devido a isto, toda abordagem de LBVS pretende desenvolver ou identificar novos compostos ativos se baseando na similaridade molecular e cada uma delas faz isso utilizando diferentes pontos de vista. Por exemplo, métodos para analisar farmacóforos ou baseados na relação quantitativa entre a estrutura e a atividade ("Quantitative Structure-Activity Relationships", QSAR) focam na busca de similaridades locais pelos fatores determinantes da atividade biológica tais como grupos funcionais e seus arranjos espaciais específicos e/ou propriedades química resultantes.

Entretanto, os químicos medicinais também sabem que pequenas alterações químicas em uma molécula ativa podem tanto potencializar sua atividade como diminuir ou mesmo tornar a molécula completamente inativa, o que é chamado de "fendas de

atividade"("activity cliffs") (63), (64). Seguindo este conceito a relação entre a estrutura e atividade biológica pode ser classificada como contínua ou descontínua. Quando contínua, as alterações na estrutura molecular de um composto ativo formam um "raio de atividade biológica"em torno deste composto que irá ser povoado por diferentes moléculas em um espectro de atividade decrescente. Quando descontinuo, as alterações nas moléculas causam efeitos drásticos (65).

Os algoritmos de busca de similaridade, geralmente, necessitam de uma molécula ou mais moléculas que servem de referência. Embora seja a técnica mais utilizada, talvez por ser a mais lógica, seja com a utilização de descritores 2D (principalmente por substruturas, avaliado frequentemente por coeficientes de similaridade como o de Tanimoto que atribuem uma valor bonificador a cada entidade presente em duas estruturas), podem ser facilmente encontrados programas que utilizam superposição estrutural ou de volume e modelos de farmacóforos (estes podem ser avaliados por coeficientes de similaridade ou pelo desvio médio quadrático - RMSD, Root Main Square Deviation) para a realização dos seus cálculos de similaridade.

1.4.3 Algoritmos quantitativos

São os algoritmos que aplicam métodos matemáticos para tentar correlacionar estrutura de uma ou mais moléculas com estrutura conhecida a uma atividade biológica, os softwares que aplicam essa metodologia são os conhecidos programas de SAR (*"Structure-Activity Relationship"*, Relação Estrutura Atividade) ou QSAR (*"Quantitative Strutcture-Activity Relationship"*, Relação Estrutura Atividade Quantitativa). De forma geral, estes programas têm por premissa tentar resolver equações empíricas da forma $Y_i = F(X_1, X_2, ..., X_n)$, onde a variável Y_i representa a atividade biológica das moléculas, e as variáveis $X_1, X_2, ..., X_n$ representam propriedades estruturais (descritores moleculares como peso molar, logP, seus fragmentos, quantidade de átomos e/ou ligações) experimentais ou calculadas de compostos (66).

Cada um dos compostos é descrito por uma ou mais propriedades independente ou descritores, de tal forma que podemos dizer que cada uma dessas representa uma dimensão. Ao final cada molécula é representada num espaço n-dimensional de vai de X_1 a X_n. Os resultados dessas metodologias são dados pela medida de contribuição de cada uma das variáveis para a atividade biológica, estabelecendo um padrão ou modelo que pode ser aplicado à triagem virtual e ao desenvolvimento racional de fármacos.

1.4.4 Algoritmos baseados em técnicas de aprendizado de máquina

Os primeiros tipos de algoritmos desenvolvidos para a busca de similaridade de moléculas, precisavam de pouca informação para a geração de resultados. Em geral, esses programas usam apenas uma molécula, com atividade conhecida, como modelo para a busca de moléculas similares em uma base de dados. Entretanto, em metodologias mais recentes, modelos preditivos passaram a ser gerados à partir de uma grande qantidade de dados moleculares, utilizando inclusive compostos ativos e inativos na geração de tais modelos. Devido à analogia de os algoritmos modernos aprenderem a diferenciar entre as diferentes classes de entidades presentes nas bases de dados, esses médotos são chamados de Métodos de Aprendizado de Máquina (do inglês, ?Machine Learning?). Sendo que, dentre os algoritmos que mais têm sido aplicados ao VS, podem ser destacados o Support Vector Machine (SVM), os métodos bayesianos e as árvores de decisão.

As técnicas de SVM, por definição, tentam construir hiperplanos em uma espaço n-dimensional (onde n representa o número de características analisadas de uma entidade) de forma a separar da melhor forma possível as diferentes classes presentes na base de dados. Essa metodologia é baseada na Teoria do Aprendizado Estatístico e tende a representar bem os dados contidos no grupo de treinamento e de apresentar capacidade moderada de classificar correntamente dados externos.

Já as técnicas Bayesianas são assim denominadas por se basearem no Teorema de Bayes, base da Estatística Bayesiana, que descreve a probabilidade de um evento acontecer levando em consideração duas ou mais causas (67). O Teorema de Bayes descreve a probabilidade de um evento A acontecer dado que o evento B aconteceu (P(A/B)) é dado pela **Fórmula 1.1** onde, P(B/A) é a probabilidade de B acontecer dado que o evento A aconteceu e P(A) é a probabilidade do evento A acontecer e P(B) é a probabilidade do evento B acontecer. Tal técnica pode ser usada também para ordenar entidades das bases de dados de acordo com as probabilidades calculadas. Os métodos mais usados em VS incluem classificadores bayesianos naive e discriminação binária de kernels, os quais podem ser aplicados para vetores binários ou de frequência.

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \tag{1.1}$$

Por último, as Árvores de Decisão representam modelos preditivos que expõem as características de um objeto, direcionando a conclusões quantitativas ou qualitativas

sobre o mesmo. Neste último caso as árvores de decisão podem também ser chamadas de Árvores de Classificação (onde o valor alvo representa uma classe, como o tipo de efeito de uma molécula no organismo ou mesmo o seu nível de atividade biológica, por exemplo). À medida que se percorre uma arvore de decisão escolhas de caminhos são oferecidas, esses são os chamados ramos, e quando se finaliza essas escolhas obtemos uma classificação ou valor, que são as chamadas de folhas. Essa técnica é muito usada em quimioinformática, onde a sua aplicação abrange análises de QSAR, predição e solubilidade de água de compostos e em Triagem Virtual.

1.4.5 Principais programas para Triagem Virtual baseada em ligantes

O processo de desenvolvimento racional de fármacos baseado em métodos computacionais existe há mais de 50 anos, entretanto o desenvolvimento de novos algoritmos e métodos ainda é uma área muito ativa na pesquisa científica.

Atualmente, existem vários programas para a realização de LBVS, utilizando diversos tipos de metodologia. Entre os mais conhecidos há os que utilizam a busca de similaridades para a apresentação de resultados como o FLAP, 4DFAP$_{OA}$, FieldScreen, LigMatch e o PhamaGist. Também há os que utilizam sobreposição de volumes, como o SHAFTS, ROCS e o Phase Shape e, ainda, alguns que se utilizam da aprendizagem de máquinas, como o JPS (Joint Farmacophore Space). Todos os métodos citados acima são metodologias que utilizam descritores 3D ou 4D (como o 4DFAPOA, que considera a amostragem conformacional extensiva com a quarta dimensão) mas, os métodos que com descritores 2D também são muito utilizados e reconhecidamente rápidos e eficientes. Entre as ferramentas que utilizam os descritores 2D há o ScreenMD, que compila uma série de descritores 2D para serem utilizados.

Os programas que utilizam descritores 2D embora possam utilizar variados tipos de informação molecular, geralmente usa representações vetoriais para os cálculos de similaridade. Entre os que utilizam descritores 3D existe uma diversidade maior de algoritmos. Também é comum a utilização da representação vetorial em programas que usam descritores 3D, mas é possível encontrar programas que realizam a superposição de formas, volumes ou estruturas moleculares e métodos baseados em aprendizado de máquina.

Abaixo descrevemos, de forma simplificada, algumas das principais ferramentas disponíveis atualmente para uso.

ScreenMD (68)

Programa da ChemAxon para a realização de estudos de VS com descritores 2D. Este programa é capaz de trabalhar com vários tipos de descritores presentes nos pacotes ChemAxon, dentre eles estão o BCUT, CP, PF, ECFP e FCFP. Esses descritores representam propriedades moleculares 2D ligadas a grupos químicos ou farmacóforos ou, ainda, à topologia molecular. Os descritores CF ("Chemical Fingerprint") e PF ("Pharmcophore Fingerprint") são topológicos, baseados em caminhos moleculares utilizando propriedades químicas ou farmacofóricas, respectivamente. Esses fingerprints com caminhos moleculares são gerados pela análise das ligações realizadas entre todos os pares de átomo, podendo-se escolher o número máximo de ligações sucessivas para a construção desses caminhos.

O ECFP("Extended Conectivity Fingerprints") e o FCFP ("Functional-Class Fingerprints") são fingerprints circulares baseados em átomos ou em farmacóforos, respectivamente. Eles geralmente pegam cada átomo interativamente como um ponto de referencia e usam a vizinhança ao redor desse ponto. Os fragmentos de caminhos ou os de vizinhança foram escritos em uma matriz com um número definido de bits. Os descritores BCUT("Burden eigenvalue descriptor") são autovalores de uma matriz de conectividade modificada conhecida como matriz de Burden.

Screen3D (69)

Essa ferramenta também é disponibilzada pela ChemAxon e consiste em uma ferramenta em linha de comando para a realização rápida de um VS baseado em ligantes 3D.

As moléculas podem ser submetidas a esse programa em formatos 2D ou 3D. Caso seja submetida em formato 2D o programa gera automaticamente conformações 3D para os ligantes através de um módulo interno. Em casos onde são passadas as conformações 3D das moléculas o usuário pode optar por trabalhar com a conformação original (método rígido) ou solicitar a geração de conformações (método flexível). Essa geração de múltiplas conformações, tanto com entradas 2D quanto 3D, é feita através da livre rotação das ligações simples das moléculas.

Ainda há, nesse programa, duas formas de se calcular a similaridade. A primeira é baseada na forma molecular, onde as moléculas são superpostas com base em sua

forma 3D. Essas formas moleculares são representadas como funções matemáticas e essa forma é "colorida"(atribuição de características) por atributos dos tipos atômicos, que pode ser feita por um pacote da ChemAxom (ChemAxon's Dreiding force field) ou baseada em características farmacofóricas. O segundo métodos fornece seus resultados de alinhamento molecular de acordo com a quantidade de átomos alinhados.

FLAP (55)

O FLAP ("Fingerprints for Ligands And Proteins") foi desenvolvido para explorar informações relevantes em estruturas cristalográficas, de pequenas moléculas ou de complexos moleculares. O método utilizado compreende a quantificação de fingerprints (macro)moleculares que seguem uma aplicação de estratégias racionais para gerar estruturas de novo e para comparar e clusterizar famílias de proteínas sem um viés de conhecimento prévio. Esse programa aplica um algoritmo rápido para descrever pequenas moléculas e estruturas protéicas usando fingerprints de 4 pontos e a forma da cavidade molecular. O procedimento inicia pelo uso um campo de forças com o programa GRID para calcular os campos de interação molecular, que são então usados para identificar as peculiaridades das interações ligante-receptor. Esses pontos calculados são então usados pelo FLAP para construir todos os possíveis conjuntos de quatro pontos apresentados em um dado sítio ativo. Uma abordagem similar pode ser aplicada para pequenas moléculas, usando diretamente os tipos atômicos do GRID para identificar as características farmacofóricas, e essa descrição das complementaridades de uma alvo e seu ligante conduzem a vários novas aplicações.

Assim, o FLAP pode ser usado para análises de similaridade a fim de comparar macromoléculas estruturalmente diferentes. Famílias de proteínas podem ser comparadas e clusterizadas dentro de classes, sem viés de conhecimento prévio e sem requerer a superposição das proteínas, alinhamentos, ou comparações baseadas em conhecimento prévio das estruturas ativas. O FLAP também pode ser usado eficientemente para LBVS ou TBVS e, finalmente pode calcular descritores para análises quimiométricas ou servir de base para a realização de procedimentos de ancoragem.

FieldScreen (70)

O FieldScreen usa um alinhamento baseado em campos moleculares e um protocolo de ranqueamento baseado em VS. Para cada molécula submetida ao FieldScreen, são

calculadas as conformações mais estáveis através de um algoritmo baseado em Monte Carlo que randomiza todas as ligações rotacionáveis. A seguir, cada molécula é descrita por um grupo de quatro campos moleculares. Os três primeiros campos são definidos como a energia de interação das moléculas com sondas positivas, negativas e neutras (com propriedades estéricas de átomos de carbono hibridizado sp^3). A quarta sonda é definida por uma função de densidade empírica que considera uma função ponderada baseada nas sondas hidrofóbicas, nas distancias dos pontos e no tipo atômico.

Para a realização de buscas em bases de moléculas são realizados alinhamentos dos campos moleculares e os resultados são dados por uma lista ordenada pelos valores de similaridade.

PharmaGist (71)

Os arquivos são submetidos a esse programa como estruturas moleculares em uma representação 3D e o programa fornece como saída uma lista dos modelos de farmacóforos que representam todas ou a maior parte das moléculas do conjunto de entrada. Para isso, o programa gera várias conformações para as moléculas de entrada. O programa detecta ligações rotacionáveis dos ligantes e os divide de acordo com os grupos rígidos, ao reorganizar esses grupos com a molécula de referência são realizadas rotações para gerar novas conformações. Essa análise conformacional também é feita de maneira explícita e determinística durante o processo de alinhamento. Em sequida, o programa atribui características farmacofóricas às moléculas (doador ou aceptor de hidrogênio, anion ou cátion, anel aromático, grupo hidrofóbico ou, opcionalmente outras características definidas pelo usuário) e realiza alinhamentos par a par iterativamente para todas da base. As moléculas usadas como referencia para os alinhamentos (chamadas de "pivô") são consideradas rígidas e o restante como flexível e a sua escolha pode se dar de forma padrão (a cada rodada uma delas é usada como molécula pivô), ou passadas pelo usuário como a que tem alta afinidade com o receptor ou aquela que tem o menor número de ligações rotacionáveis. Ao realizar os alinhamentos par a par o programa considera as melhores superposições para a realização de um alinhamento multiplo. No último estagio, candidatos a farmacóforos derivados de diferentes iterações são clusterizados e os valores não redundantes mais altos são fornecidos ao usuário em uma lista ordenada. O PharmaGist está disponível através de um servidor (`http://bioinfo3d.cs.tau.ac.il/pharma/index.html`).

SHAFTS (72)
O programa SHAFTS pode ser descrito como uma abordagem híbrida para a busca de similaridade entre moléculas pois, nesse programa, os resultados são dados pelo cálculo das somas das similaridades normalizadas de fingerprints de farmacóforos e por superposição de volumes.

Para a realização dos cálculos de similaridade, primeiramente, o algorítmo realiza o mapeamento de todos os PPPs presentes nas moléculas. Isso pode ser feitos através de uma estrutura cristálográfica da molécula ou com a utilização do software Cindy (73), desenvolvido pelo mesmo grupo, que utiliza um algoritmo evolucionário para realizar a amostragem conformacional de uma dada molécula. Após isso, todas as moléculas da base de dados são submetidas a uma amostragem conformacional e, opcionalmente, a um alinhamento em relação à molécula de referência. Por fim, são realizadas buscas de similaridade utilizando triplets de PPPs , sendo que, caso seja encontrado o mesmo triplet na molécula avaliada e na de referência, é realizado o cáculo da similaridade que leva em consideração o alinhamento de PPPs e a superponsição de volumes. Ao final, apenas a conformação com o maior valor de similaridade é armazenada.

ROCS (74)

Este software tem se tornado o padrão ouro para estudos de LBVS que utilizam comparações de formas moleculares. Esse programa foi originalmente desenvolvido para usar somente o método de formas e posteriormente foi adicionado um termo baseado em complementaridades eletrostáticas (chamado de ROCS-color). O ROCS usa uma função Gaussiana de suavização para representar o volume molecular que fornece vantagens em termos de velocidade e uso de operações matemáticas simplificadas.

Para a realização dos alinhamentos estruturais, além dos volumes , o ROCS pode levar em consideração as propriedades químicas de cada átomo, conhecido como "colors"ou farmacóforos, que foi implementado justamente para facilitar a identificação dos compostos que são similares tanto em formas quanto em características químicas. Os ranqueamentos podem envolver as formas, os "colors"ou uma combinação deles. As moléculas são alinhadas visando a maximização da superposição entre elas, e para cada molécula na base de dados, é selecionado o conformero com a melhor superposição com os ligantes utilizados como referência na busca (75). Assim, o ROCS apresenta várias formas de execução de um experimento de VS e dentre as métricas possíveis de serem utilizadas estão similaridade ShapeTanimoto, ScaledColor, ColorTanimoto,

TanimotoCombo and ComboScore.

Phase Shape (76)

No programa Phase-Shape cada conformero de uma dada molécula é alinhada com a molécula de referência para a realização dos cálculos de similaridade baseados em forma. Para esses cálculos o programa pode tratar todos os átomos de forma equivalente ou favorecer alinhamentos que superpõe átomos do mesmo tipo, dependendo da escolha do usuário. Há também três tipos possíveis para tipificar um átomo na busca de similaridade baseada na forma: usando tipos do macromodel, tipos de átomos, ou tipos QSAR (onde são usadas características doador de hidorgênio, hidrofóbico/não polar, íons negativos e positivos entre outras) .

4DFAP$_{oa}$ (77)

O 4DFAP ("4D flexible atom-pairs similarity measure"), é um método sofisticado que compara moléculas utilizando o seu espaço conformacional. A abordagem pode ser dividida em uma etapa de processamento e uma etapa de cálculo de similaridade.

Na etapa de processamento é gerado um conjunto de conformações de uma molécula que são codificados através de modelos extraídos dos perfis de distância entre os pares de átomos dessas conformações. Essas distribuições de distâncias podem ser representadas graficamente como funções Gaussianas e os modelos são gerados com a utilização dessas funções e gera modelos que são chamados de GMMs (Gaussian Mixture Models). Esses GMMs são modelos que descrevem a forma probabilística de distribuição das distâncias das várias conformações de uma molécula (**Figura 1.4**)

Os cálculos de similaridade entre um par de moléculas são feitos através de uma função que leva em conta a estrutura molecular e a sobreposição dos GMMs. A similaridade final pode ser calculada através da soma normalizada dos valores dentro da matriz de distâncias entre os GMMs, como feito pelo 4D FAP original, aplicado em estudos QSAR/QSPR (78), (79). Também, para os estudos em triagem virtual, a similaridade pode ser calculada através de um algoritmo que implementa a resolução ótima do Problema de Atribuição (OA - Optimal Assignment), que maximiza a valor da similaridade entre pares de átomos de duas moléculas, resultando na variante 4D FAPOA (77).

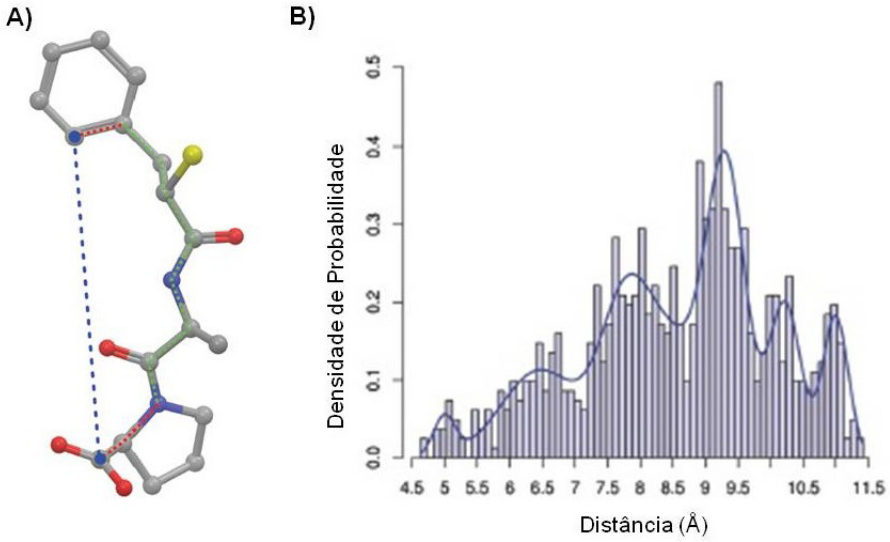


Figura 1.4: Representação de moléculas no 4DFAP$_{oa}$. **A)** Menor caminho entre átomos flexíveis. O menor caminho topológico é mostrado por linhas hachuradas vermelhas e verdes. A linha vermelha representa uma ligação rígida enquanto a linha verde marca as ligações rotacionáveis. A ultima ligação do caminho (liga um heterociclo ao carbono de um grupo carboxil) é tratada como uma ligação rígida porque a rotação dessa ligação não influencia a distancia geométrica entre os pares de átomos. **B)** Visualização do histograma baseado na distribuição de distâncias dos pares de átomos mostrados em A. A linha representa a representa a respectiva GMM que modela o comportamento das distâncias entre os átomos no espaço conformacional. Adaptado de (77).

1.5 Avaliação de Desempenho de uma Triagem Virtual

Os algoritmos de VS fornecem, geralmente, ao final de seus cálculos uma lista ranqueada qualificando ou quantificando os prováveis alvos ou ligantes de interesse (aqueles que são os mais prováveis de desempenhar a atividade biológica procurada). Para a construção da lista são utilizadas medidas de similaridade ou dissimilaridade que correlacionam as estruturas da base de dados com uma ou mais estruturas de referência. Essas medidas podem ser dadas pela presença ou ausência de certas características físico-químicas, por um coeficiente de similaridade ou pelo alinhamento de estruturas ou modelos no espaço tridimensional.

No caso de contagem da presença de certas características a medida é dada bonificando as semelhanças em relação a uma entidade de referencia, podendo ser acrescida também uma forma de penalização das diferenças encontradas. Essas medidas podem ser comparadas a um coeficiente de similaridade, com a diferença que o usuário tem maior liberdade para construir o seu modelo de medida de similaridade. Os coeficien-

tes de similaridade geralmente são associados à medida de similaridade de vetores ou fingerprints. Entre os coeficientes de similaridade mais conhecidos estão o coeficiente de tanimoto e a distância euclidiana, mas vários trabalhos listam uma série de outros coeficientes, cada um deles com determinadas particularidades (**Tabela 1.4**), contudo, nenhum deles penaliza as diferenças encontradas.

A construção dessa lista também pode ser dada pelo alinhamento estrutural e, nesse caso, a medida mais difundida é o Desvio médio quadrático (Root Mean Square Deviation, RMSD). Esta é uma medida de dissimilaridade que varia no intervalo (0, ∞), sendo 0 o valor encontrado quando as duas estruturas comparadas são exatamente iguais. A formula do RMSD é apresentada na fórmula 1.2, onde "A"e "B"representam duas estruturas moleculares comparadas (ligantes, proteínas ou modelos estruturais simplificados) e "n"é o número total de átomos ou pontos. X, Y e Z se referem às coordenadas no espaço tridimensional dos átomos das estruturas "A"e "B".

$$RMSD_{(A,B)} = \sqrt{\frac{\sum_{(i=1)}^{n} (A_{xi} - B_{xi})^2 + (A_{yi} - B_{yi})^2 + (A_{zi} - B_{zi})^2}{n}} \tag{1.2}$$

Geralmente é adotado um ponto de corte nessa medida (RMSD) para a seleção das estruturas mais semelhantes à referência. Em geral essa medida não deve ser maior que dois Angstrons, sendo que a medida ideal é zero, quando as estruturas são absolutamente iguais.

Com a quantificação dessas similaridades são construídas matrizes de similaridade (correlação todos contra todos das entidades analisada, dispostas na forma de matriz n x n, sendo "n"o numero de entidades da base de dados) ou listas organizadas de forma crescente (medidas de dissimilaridade) ou decrescente (para medidas de similaridade). Isso permite a realização de avaliações da eficiência do VS em agrupar entidades semelhantes ou selecionar as que são mais similares a um padrão.

A métrica mais popular de avaliação de eficiência de VS é o fator de enriquecimento ("Enrichment Factor", EF), talvez por ser uma medida de métodos mais direta e intuitiva. Entretanto há diversos problemas associados ao EF. Um deles é a dependência da inserção de um ponto de corte para a triagem em uma base de dados, geralmente os cálculos são automatizados calculando-se cortes sucessivos de 1% em 1% na faixa entre 1% a 5% da base de dados. Outro problema mais grave é a dependência do número de compostos ativos e inativos, pois dependendo da prevalência o valor da métrica muda.

Alguns autores defendem fortemente o uso de medidas padronizadas, incluindo as

Tabela 1.4: Vários coeficientes de similaridade usados para comparar os fingerprints em LBVS. Adaptada de (80). Todos os coeficientes calculam a similaridade entre dois fingerprints A e B, onde "a"é o número de bits presentes exclusivamente em A, "b"é o número de bits exclusivamente em B, "c"é o número de bits presentes em A e B e d é o número de bits que não estão nem em A nem em B.

Número	Medida	Fórmula
1	Cosseno	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
2	Dice	$\frac{2*c}{\sqrt{(a+c)*(b+c)}}$
3	Euclidiana	$\frac{c+d}{\sqrt{(a+b+c+d)}}$
4	Forbes	$\frac{c*(a+b+c+d)}{\sqrt{(a+c)*(b+c)}}$
5	Hamman	$\frac{(c+d)-(a+b)}{\sqrt{(a+b+c+d)}}$
6	Jaccard	$\frac{c}{\sqrt{(a+b+c)}}$
7	Kulczynski	$0.5 * [\frac{c}{(a+c} + \frac{c}{b+c}]$
8	Yule	$\frac{(c*d)-(a*b)}{(c*d)+(a*b)}$
9	Manhattan	$\frac{(a+b)}{(a+b+c+d)}$
10	Matching	$\frac{(c+b)}{(a+b+c+d)}$
11	Pearson	$\frac{(c*d)-(a*b)}{(a+c)*(b+c)*(a+d)*(b+d)}$
12	Rogers-Tanimoto	$\frac{(c+d)}{(a+b)+(a+b+c+d)}$
13	Russell-Rao	$\frac{c}{(a+b+c+d)}$
14	Simpson	$\frac{c}{min((a+c),(b+c))}$
15	Tanimoto	$\frac{c}{(a+b+c)}$

curvas ROC ("Receiver-Operator Characteristic") e a área sob as curvas ROC ("Area under the curve ROC", AUC) que é comumente aplicada em outros campos como análises estatísticas, mineração de dados e técnicas de aprendizado de máquina (81). Entre-

tanto, o AUC não leva em consideração o chamado "early recognition"(reconhecimento precoce), que é a capacidade de um programa em recuperar os compostos ativos no inicio de seu ranqueamento. Assim, (Truchon & Bayly, 2007) (82) desenvolveram a métrica Boltzmann-enhanced discrimination of ROC (BedROC), que usa uma ponderação exponencial para atribuir pesos maiores para as abordagens que conseguem identificar ativos nos topos de suas listas. No entanto, Nicholls (83) apresentou evidencias de que há uma forte correlação entre AUC e BEDROC, sugerindo que o AUC seria uma medida suficiente para a avaliação de performance em triagem virtual.

1.6 Similaridade de Sítios Ativos

A estrutura tridimensional (3D) de macromoléculas biológicas é de extrema importância para decifrar a maquinaria das células vivas, sendo utilizadas em "drug dicovery"para o desenvolvimento de pequenas moléculas com o potencial de inibir, ativar ou modular proteínas importantes para aplicações terapêuticas. Até meados da década 1990 este número de estruturas 3D era limitado, mas hoje abordagens que utilizam essas informações têm sido extremamente encorajadas.

Atualmente, devido à quantidade de reações adversas causadas por medicamentos, podendo causar toxicidade e efeitos colaterais, causados principalmente pela permissividade de ligação a vários alvos indesejados (*"off-targets"*). Atualmente a previsão de similaridades entre sítios ativos de proteínas tem se consolidado como uma estratégia importante ao desenvolvimento racional de fármacos. Estas comparações podem ser realizadas de várias formas como a realização de alinhamento de sequencias primárias de aminoácidos ou elementos de estrutura secundária (α-hélice, β-folhas e voltas), mas esse tipo de análise implica em considerarmos que o arranjo espacial dos sítios ativos são provocados por padrões idênticos presentes nas sequencia de proteínas, o que nem sempre é correto e, pode levar a resultados incorretos e/ou ambíguos (84) (85). Ainda, na década passada, começaram a surgir métodos que baseiam a comparação de sítios ativos na simplificação da estrutura tridimensional da proteína, isso graças aos avanços obtidos nos algoritmos de ancoragem e de triagem virtual de pequenas moléculas (86). Esse tipo de abordagem visa à redução do custo computacional para permitir a análise de grandes quantidades de informação, conforme necessitam as aplicações realísticas de VS.

1.6.1 Alinhamento de sequencias de proteínas

Aplicado às proteínas o alinhamento de sequencias é geralmente empregado para inferir homologia, um conceito biológico de que proteínas de uma mesma família ou de famílias próximas tem sua estrutura tridimensional semelhante e essa estrutura é refletida na sequência protéica. Por esta metodologia também é possível identificar resíduos de aminoácidos conservados utilizando apenas uma sequencia primária. Partindo do pressuposto de que a atividade funcional de uma proteína é dada pelo seu sítio ativo, os resíduos identificados como sendo conservados muitas vezes fazem alusão aos aminoácidos dessas cavidades. Entretanto isso nem sempre é verdade, podendo existir casos em que mesmo proteínas com similaridade de sequencia muito baixa têm suas estruturas tridimensionais ou os seus sítios de atividade conservados e, por isso têm a capacidade de desempenhar um mesmo papel biológico. Exemplo disso é a hemoglobina (PDB: 1jebA) e a mioglobina (PDB: 2mm1) humanas que apesar de terem uma similaridade de sequências de apenas 27% a sua similaridade estrutural é de 90%.

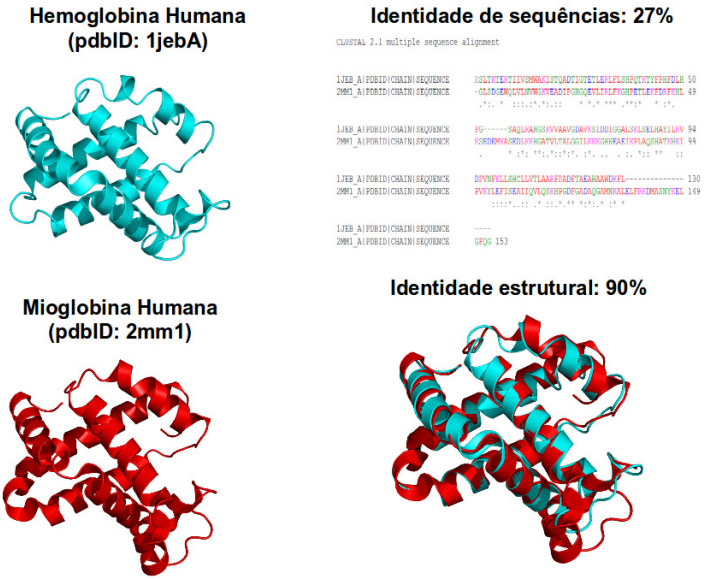


Figura 1.5: Relação similaridade sequencias e estrutural entre as proteinas 1jebA e 2mm1

Dentre as aplicações dos alinhamentos de sequências protéicas estão a identificação de domínios conservados e sítios ativos de enzimas e o descobrimento de assinaturas de famílias gênicas.

São varias os programas disponíveis para a realização de alinhamento de sequências, os principais deles estão descritos na **Tabela 1.5** . A classificação dos algoritmos utilizados nesses programas é dado pelo tipo de alinhamento realizado (ótimo ou heurístico,

local ou global) e pela quantidade de sequências que o algoritmo consegue alinhar por vez (simples ou múltiplo).

Tabela 1.5: Principais programas para a realização de alinhamentos de senquências. Fonte: Prosdocimi, 2002 (87).

Programa	**Tipo de alinhamento**	**Precisão do alinhamento**	**Número de sequências a serem alinhadas**
BLAST2Sequences	Local	Heurístico	2
SWAT (Smith-Waterman)	Local	Ótimo	2
ClustalW	Global	Heurístico	N
Multalin	Global	Heurístico	N
Needleman-Wunsch	Global	Ótimo	2

Uma vez alinhadas as sequencias, os programas necessitam de uma medida para quantificar a qualidade do alinhamento. Essas medidas geralmente são feitas geralmente por tabelas de substituição, sendo a BLOSUM (BLOcks of amino acid SUbstitution Matrix) uma das mais conhecidas. Essas tabelas referenciam a bonificação que é dada por alinhamentos corretos e penalizam os incorretos, podendo ter incrementos nessa penalização dependendo dos tipos de aminoácidos que foram alinhados incorretamente (**Figura 1.6**). Por exemplo, se um aminoácido hidrofóbico for alinhado com um outro aminoácido hidrofóbico a penalização deve ser menor que se tivesse sido alinhado com um aminoácido polar. Além disso, algumas delas também penalizam as falhas encontradas nos alinhamentos, vindas da necessidade de abrir as sequencias proteicas para a obtenção de um alinhamento melhor, chamamos isso de "gaps". Todos esses alinhamentos incorretos têm explicação biológica baseada em princípios de mutação gênica e das possíveis inserções e deleções de bases no genoma que podem acontecer com tempo.

As principais vantagens desses métodos são a rapidez dos algoritmos, a disponibilidade de ferramentas gratuitas on-line, além da maior quantidade de dados disponíveis (**Figura 1.1**). Entretanto o alinhamento de sequências tende a ser muito inacurado para utilização como parâmetro de avaliação de similaridade entre sítios ativos, uma vez que o arranjo tridimensional é quem garante a função biológica. Além disso, proteí-

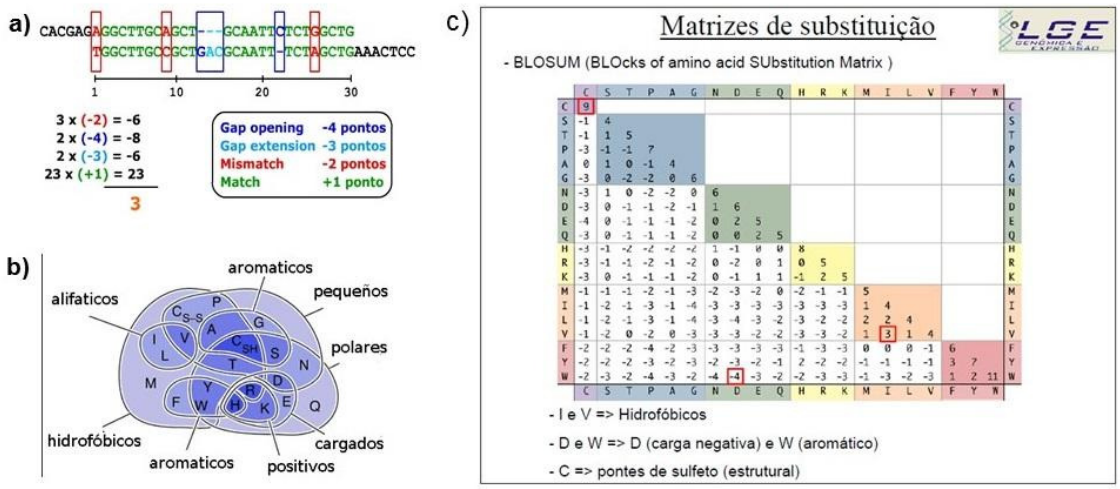


Figura 1.6: Exemplificação de como podem ser realizados os cálculos dos valores de similaridade entre sequências proteicas **(a)**. Em **(c)** mostramos a matriz de substituição BLOSUM, que leva em conta as propriedades dos aminoácidos **(b)**, para determinar um valor de penalização para os casos de "*mismatch*"encontrados.

nas de famílias diferentes tendem a apresentar baixas similaridades de sequência, o que dificulta os estudos de predição de off-targets e efeitos colaterais por essa metodologia.

1.6.2 Métodos 3D de identificação de similaridade

Os métodos de comparação e cálculo de similaridade entre sítios de proteínas geralmente passa por uma etapa de pré-tratamento dos dados que visa facilitar a realização dos cálculos. Isso pode ser feito através do uso de uma representação simplificada dos resíduos da cavidade, pela divisão da superfície do sítio em pequenos pedaços ou arranjos que então passam a ser tratados como padrões geométricos ou pelo armazenamento de fingerprints numéricos representativos da estrutura química e eletrostática da estrutura protéica.

1.6.2.1 Métodos baseados na comparação de padrões geométricos

Esse tipo de metodologia usa três passos para realizar os cálculos de similaridade. Primeiro as estruturas de duas proteínas são colocadas no mesmo sistema de coordenadas 3D a fim de diminuir a dificuldade da sobreposição. Nessa fase são detectados resíduos chave para servir de base para essa comparação, além disso, somente um determinado número de pontos é utilizado e esta seleção geralmente é feita de acordo com propriedades farmacofóricas, geométricas e/ou químicas na vizinhança de cada

uma deles. Segundo, os padrões resultantes de cada entidade são estruturalmente alinhados usando rotações e translações se necessário, produzindo um alinhamento onde o número máximo de pontos correspondentes é obtido. Por último, a função de scoring quantifica a similaridade.

1.6.2.2 Representação simplificada das cavidades das proteínas

Apesar de estarmos comparando o sítio ativo, sabemos que só alguns resíduos participam do reconhecimento proteína ligante. Assim, uma idéia básica é considerar em um primeiro momento só os resíduos que estão na superfície do sítio, pois estes aminoácidos tem maior probabilidade de estarem envolvidos no reconhecimento molecular, de forma direta ou indireta (ligação mediada por uma molécula de água, por exemplo). Assim, dependendo do objetivo da comparação, podem ser considerados todos os resíduos da superfície do sítio (88) ou somente os resíduos mais próximos de um ponto de referência da cavidade. Para efeito de representação, a cavidade dos sítios podem ser analisadas basendo-se na forma de sua superfície (ProSurfer (89), SiteEngine (88), CavBase (90)) ou de uma distancia calculada (com ponto de corte geralmente entre 5 e 7 angstrons) entre os resíduos e o ligante co-cristalizado (eF-site (91), SuMo (92), CPASS (93), SitesBase (94)).

Os resíduos de aminoácido selecionados podem ser transformados em um conjunto de pontos que podem ser regulares ou irregulares (**Figura 1.7**). Quando são irregulares, cada resíduo pode ser descrito por todos os seus átomos ou por apenas um ponto (pseudo-centros) ou, ainda, por pontos que representam um grupo de átomos (pseudo-atomos). A cada ponto considerado podem ser atribuídas propriedades fisioquímicas ou farmacofóricas dos átomos correspondentes. A informação das moléculas do sítio ativo também podem ser codificadas por uma representação cartesiana regularmente espaçada de sua superfície e os pontos podem ser gerados por uma triangulação utilizando a metodologia de Connolly, sendo os descritores atribuídos a estes de acordo com a proximidade dos átomos, pseudo-átomos ou propriedades de superfície (91) tais como potencial eletrostático, hidrofobicidade e curvatura.

1.6.2.3 Alinhamento estrutural de cavidades proteicas

A similaridade entre duas cavidades de proteína é sempre inferida a partir do melhor alinhamento estrutural dos padrões correspondentes. A busca por esse melhor alinhamento consiste em procurar transformações rígidas que otimizem o número de elementos correspondentes. A metodologia mais simples é a busca interativa pela melhor rotação/translação de uma estrutura, mantendo sempre o alvo a ser comparado

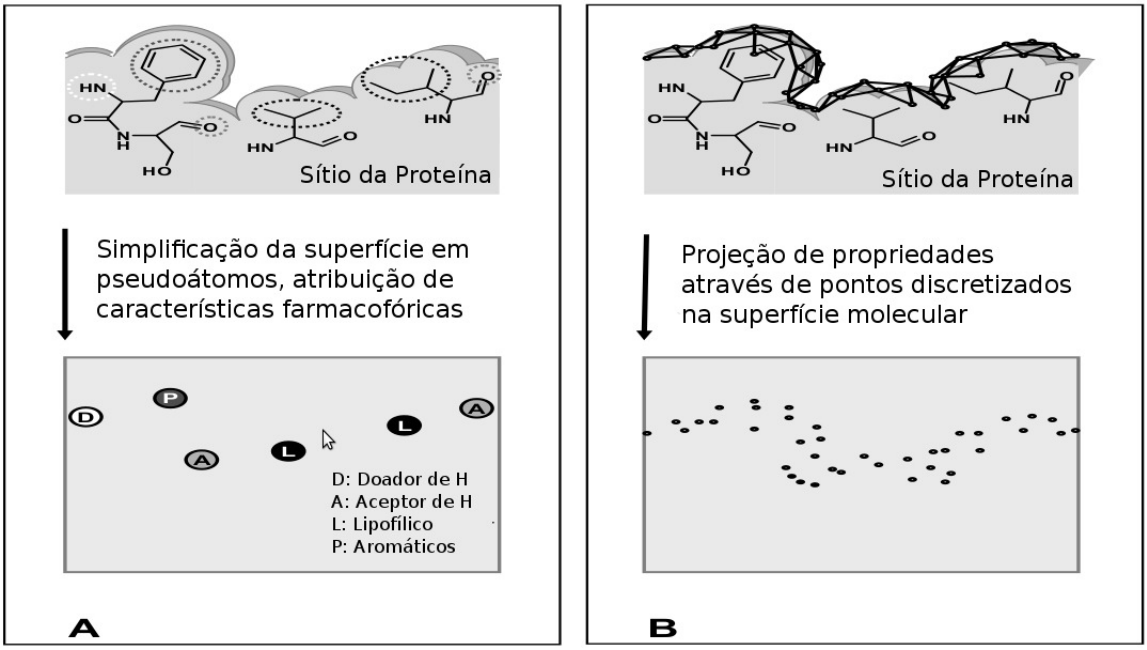


Figura 1.7: Descrição dos sítio ligante de proteínas através da padrões de pontos. (A) Exemplo de padrão de pontos irregular. O sítio da proteína é representado por pseudoátomos, também chamado de pseudo-centros, nesse exemplo elas foram atribuídas características farmacofóricas pertencentes ao respectivo grupo funcional exposto à superfície. As coordenadas desses pseudoátomos geralmente estão associadas ao centro de massa do grupo funcional correspondente. (B) Exemplo de padrão de pontos irregular. A superfície, usualmente é aplicada uma triangulação, decompondo a superfície em uma malha de triângulos. A cada vértice é atribuída uma característica de acordo com propriedades locais. Fonte (86)

fixo, este método identifica as regiões mais similares que podem ser então mais precisamente exploradas. Entretanto este método tem a desvantagem de ser relativamente lento especialmente quando entidades grandes são comparadas.

A complexidade do alinhamento estrutural pode ser reduzida pelo agrupamento dos pontos em tripletos de acordo com a sua proximidade, sendo então realizada a comparação de todos os tripletos formados (ProSurfer, SuMo, SitesBase). Há duas formas gerais de realizar essas comparações: superposição geométrica e hashing geométrico. Na superposição geométrica os triângulos formados são rotacionados visando encontrar as posições com melhor alinhamento e, no hashing geométrico, a cada triangulo formado costuma ser atribuído um valor único, que serve de base para a comparação com outro tripleto, sem levar em consideração a sua orientação (86).

Outra abordagem frequentemente utilizada é a detecção de cliques (95) através de isomorfismos de subgrafos. Aqui, as estruturas são representadas por grafos: os pontos (átomos, pseudo-centros, pesudo-átomos) são os nós e estão conectados por arestas às

quais geralmente são associadas as distancias entre um nó e outro (<12 Å no CavBase). Para calcular a similaridade o algoritmo parte do princípio da multiplicação de grafos, que é obtida pelo pareamento de nós com nomes idênticos (propriedades ou descritores fisicoquímicos - CavBase -, ou potenciais eletrostáticos e curvaturas - eF-site). Os nós serão conectados por arestas se estas possuírem os mesmos valores nas duas estruturas comparadas, entretanto este valor pode ser acompanhado por uma faixa de tolerância, que geralmente está entre 1 e 1,5 angstroms. Ao final é obtido o clique máximo comum, que é o maior subconjunto de nós completamente conectado por arestas em um grafo.

1.6.2.4 Scoring de similaridade

O valor que quantifica a qualidade do alinhamento entre dois sítios ativos, seja qual for a metodologia utilizada para a comparação, tenta refletir a quantidade de pontos alinhados corretamente, sejam átomos ou descritores. Esses cálculos tentam prever e descartar as soluções erradas para o problema proposto.

As funções de scoring mais simples utilizam a contagem desses pontos ou porcentagens derivadas desse processo. Assim, essas porcentagens dependem do alinhamento de um padrão de referencia com a estrutura a qual estamos avaliando a similaridade. Entre os índices mais empregados estão o coeficiente de Tanimoto (calcula o numero de pontos alinhados sobre o total de pontos de ambas as entidades), a contagem ponderada das coerências encontradas na geometria 3D (RMSD) ou dependente das propriedades estudadas (pode ponderar o resultado dando maior valor a determinados descritores) (86). Algumas metodologias como CavBase e o SiteEngine usam metodologias híbridas utilizando pontos de correspondência e análise da superfície molecular das estruturas em análise.

1.6.2.5 Comparação de cavidade por fingerprints

A utilização de representações numéricas das proteínas possibilita realizar a comparação entre duas cavidades de proteínas sem que seja necessário o emprego do dispendioso processo de alinhamento geométrico. Como exemplo disso, há o programa SiteAlign (96) que transforma sítios definidos pelo usuário ou sítios provenientes do scPDB em um número inteiro de tamanho fixo. Neste programa uma esfera de 1angstron de raio, cuja superfície é discretizada em 80 triângulos é colocada no centro da cavidade proteica. Em seguida são projetados vetores dos carbonos alfa dos aminoácidos presentes no sítio em direção ao centro da esfera e são atribuídas características a cada face da esfera cortada por esse vetor. Tais características dependem do tipo de aminoácido presentes sendo atribuídas 5 índices farmacofóricos e 3 topológicos . Como

resultado a esfera é convertida em um fingerprint de 640 inteiros (80 triângulos com 8 descritores cada). As similaridades calculadas neste caso dependem de um alinhamento das esferas e é dado pela média normalizada das diferenças de cada descritor de cada um dos triângulos das esferas comparadas.

Também, Nathanaël and Rognan, 2010 (97) utilizaram fingerprints de fármacóforo para a realização de análises de similaridade entre sítios protéicos. Nesse trabalho foi apresentada a ferramenta FuzCav, essa utiliza como farmacofóricas as propriedades fisicoquímicas dos aminoácidos expostos ao sítios ativos realizando após uma triangulação. Cada um dos triângulos obtidos é associado a uma das posições de um fingerprint de tamanho 4833 e, através de cálculos vetoriais, são calculados os valores de similaridade entre os sítios ativos.

Tabela 1.6: Alguns dos programas disponíveis para a busca de similaridade de ligantes e busca de alvos. Adaptada de (98)

Programa	**Método**
eF-seek (99)	Avaliar a similaridade de Sítios pela detecção clique sobre os vértices triangulados da superfície acessível do solvente.
FINDSITE (100)	Um método para a previsão do sítio-ligante e anotação funcional com base na similaridade de ligação entre os grupos locais de fraco modelo homólogo por segmentação de estruturas identificadas.
SuMo (101)	É um sistema de bioinformática para a busca de similaridades em estruturas e subestruturas 3D de proteínas o que permite investigar o PDB para procurar sítios de ligação correspondentes a estruturas de uma proteína ou, inversamente, permite encontrar estruturas de proteínas correspondentes a um sítio ativo.
Q-Sitefinder (102)	Método (energético) de previsão do sítio ligante que trabalha com sondas de ligação hidrofóbica (CH3) para a proteína, e encontrar clusters de sondas com a ligação de energia mais favorável.
THINK (103)	Sistema modular projetado para ajudar na descoberta ou otimização moléculas ativas através de comparações por similaridade. Atualmente em uso para a geração de perfis farmacofóricos, mas possui uma série de outras ferramentas para virtual Screening baseado na estrutura proteica, Geração de moléculas pelo método de síntese de Novo e análise dos dados (quando utilizado concomitantemente com o software de fluxo de trabalho KNIME).

1.7 Tratamento de estruturas moleculares

A maioria das metodologias de LBVS e TBVS abordam as estruturas moleculares como rígidas, isso implica em considerar de forma errônea que o acoplamento entre receptor e ligante é rígido, quando na verdade ele é moldado dinâmicamente e marcado pela acomodação de ambos, alvo e ligante, para atingirem o estado de menor energia. Embora a maioria dos autores descreva a relevância de utilização de vários conformeros de um mesmo ligante, como tentativa de avaliar a influencia da flexibilidade nas interações receptor-ligante, poucas são as metodologias que aplicam a exploração do espaço químico das moléculas em análises de Virtual Screening. Isto pode prejudicar a análise para o VS, uma vez que pequenas diferenças nessas moléculas podem levar à identificação errônea de estruturas e induzir falsas taxas de enriquecimentos e erros em outras medidas de desempenho do VS, além de possibilidade de introdução de viés ao experimento.

Quando se trabalha com pequenas moléculas é importante que o tratamento delas seja feito de forma adequada. Um tratamento de moléculas deve incluir a padronização dos arquivos e das estruturas, cálculos de isômeros e de conformações. A padronização dos arquivos consiste no registro de todas as estruturas das moléculas da base de dados para um mesmo padrão de arquivo, assim garante-se que todas as informações necessárias ao estudo estarão presentes e que os tratamentos subsequentes partirão de um mesmo ponto. Já os cálculos de isômeros, incluindo protomeros e tautomeros, são importantes porque nem todos os isômeros de uma molécula ativa são necessariamente ativos e a mesma explicação pode ser dada para a necessidade de cálculos de vários conformeros.

Ao iniciar um projeto de LBVS, visando à redução da possibilidade de falhas no experimento, a escolha do formato dos arquivos de entrada deve ser bastante criteriosa, pois cada um dos diferentes formatos de arquivos existentes (entre eles pdb, sdf, mol, mol2, smi) apresenta peculiaridades que podem dificultar o reconhecimento dos detalhes das estruturas por alguns programas disponíveis. Ainda, em caso de necessidade de conversão de formatos de arquivos deve-se assegurar de que todas as informações relevantes foram mantidas.

Quando se fala em correção das estruturas moleculares, os pesquisadores geralmente se preocupam muito com a presença de estruturas incorretas na base de dados. Como incorretas pode-se entender a ausência ou excesso de átomos e/ou ligações químicas inconsistentes, erros mais fáceis de serem identificados e tratados. Entretanto devem ser tomados cuidados especiais com questões menos visíveis como estereoquímica, protome-

ria e tautomeria. Arquivos de diferentes formatos apresentam características próprias que podem não ser corretamente interpretadas ou não são um formato de entrada padrão a todos os programas, o que pode ocasionar erros de leitura. Dependendo das definições do software, as informações estereoquímicas podem ser atribuídas a partir de dados da geometria 3D ou de marcadores de quiralidade, podendo ser ignorados dependendo do algoritmo. Ainda, em alguns formatos, como por exemplo os arquivos PDB provenientes de cristalografia por raio-x, os átomos de hidrogênio não são representados o que pode gerar erros se houver a necessidade de cálculo de cargas, formas, volumes ou fragmentos. Além disso, a representação e manuseio de partes aromáticas diferem em alguns programas, ocasionando problemas na análise decorrentes do fato de o arquivo usado não possuir as anotações aromáticas apropriadas (104).

Dificuldades assim podem gerar falhas no processo de VS, pois a atividade de determinada conformação ou protômero de uma molécula ativa pode estar erroneamente sendo considerada como molécula de referência e produzindo descritores incorretos para comparação (seja de substruturas, farmacóforos ou potenciais físico-químicos) ou erros de alinhamento estrutural. Na verdade, até os melhores programas podem ser negativamente afetados quando existem erros nos arquivos de entrada (104). Além disso, a escolha dos melhores formatos, tanto para a entrada quanto para a saída de dados, evita a necessidade de reanálises e facilita recuperação de informações e a demonstração dos resultados.

A segunda padronização a ser realizada deve ser a do tratamento das estruturas para a geração de isômeros e conformeros, devendo-se garantir que todas as moléculas sejam submetidas a todas as etapas do tratamento. Existem vários programas para isso, sendo que os principais estão disponíveis em pacotes de programas comerciais utilizados na descoberta racional de fármacos. As principais empresas a comercializar estes programas são a Accelrys, OpenEye , Schrodinger e Tripos. Em geral, esses programas usam dois tipos de algoritmos: os determinísticos e os estocásticos. Determinísticos são os que calculam as conformações a partir dos ângulos torcionais de maneira sistemática, sendo esse método mais recomendado para estudos com moléculas com flexibilidade conformacional limitada, apresentando problemas em tratar moléculas com muitas ligações livres devido ao grande custo computacional necessário. Os algoritmos estocásticos usam elementos aleatórios para gerar conformações no espaço tridimensional (105), podendo usar cálculos de estabilidade, de sobreposição e cálculo de cargas como referencia para a aceitação das estruturas geradas.

1.7.1 Programas para cálculos de amostragem conformacional de moléculas

Accellrys

A Accelrys (http://accellrys.com) possui três algoritmos para a amostragem conformacional de moléculas: dois deles estão implementados no CatConf (ou ConFirm) e o mais um no CAEZAR. O CatConf é parte dos protocolos e ferramentas para a modelagem de farmacóforo disponibilizado juntamente com a plataforma "Catalyst", e seus modos de busca conformacional em ligantes são o "fast"e o "best". O modo "fast"aplica uma busca exaustiva modificada (semi-exaustiva, na verdade) nas porções mais flexíveis das moléculas enquanto usa uma base de conformações de anéis predefinidos. A seguir aplica um cálculo de geometria para evitar estruturas duplicadas e selecionar aquelas que apresentarem menor energia, a seguir essas estruturas são submetidas a um método heurístico visando alcançar uma maior diversidade estrutural. O método "best"usa uma abordagem baseada em distâncias geométricas, que necessita de um tempo computacional maior que o modo "fast". Ele usa uma matriz de coordenadas internas das moléculas com valores inferiores e superiores para cada átomo para gerar todas as conformações das moléculas (106). O refinamento é realizado também com um campo de força CHARMm, além disso para alcançar a diversidade conformacional é aplicado um método chamado de "polling" (107) que usa barreiras energéticas artificiais colocadas entre os pontos de mínimo energético local.

O CAEZAR (Conformer Algorithm based on Energy Screening and Recursive Buildup) é um programa de geração conformacional baseado em busca sistemática, disponibilizado juntamente com o "Discovery Studio". Este programa representa as moléculas como um arvore onde os nós representam menores fragmentos moleculares possíveis e os vértices representam as ligações rotacionáveis que ligam esses fragmentos. Esse sistema pode ser usado tanto para a geração de conformeros quanto para a obtenção de estereoisômeros. As conformações são então recursivamente geradas de acordo com esse grafo formado, onde os fragmentos são ligados de acordo com um conjunto de ângulos de torção (entre 6 e 12 valores) e as conformações são conferidas através de cálculos de energia ou de acordo com um número máximo pré-definido de conformações intermediárias a serem geradas. O CEASAR tente a ser 20 vezes mais rápido que o modo "fast"do ConFirm e é capaz de realizar a amostragem conformacional das moléculas com maior eficiência (106) .

Chemical Computing Group

MOE, da Chemical Computing Group (http://www.chemcomp.com) contêm módulos para buscas sistemáticas e estocásticas (108), além de um módulo próprio chamado de "Conformational Import". A busca sistemática aplica uma rotação de cada ligação rotacionável utilizando um ângulo de incremento (15º para cíclicos, 60º ou 120º para ligações acíclicas). As estruturas são checadas buscando-se átomos com impedimentos estéricos e, caso existam, as estruturas são descartadas. Ao final as estruturas com menores energias são selecionadas e aquelas iguais (cálculo baseado em RMSD) são descartadas. O método de busca estocástico aplicado no MOE gera conformações por mudanças estruturais repetitivas e aleatórias nos ângulos de torção. A busca para quando um determinado número de ciclos ou de conformações é alcançado. Já o "Conformational Import"combina um sistema baseado em regras e uma base de dados pré-calculada com uma busca conformacional estocástica, onde as moléculas são primeiro opcionalmente submetidas a alguns filtros (de acordo com estado de protonação, presença de determinados grupos ativos ou propriedades fisicoquímicas, por exemplo) e, a seguir, cada molécula é fragmentada. As conformações são calculadas de acordo com certos padrões geométricos determinados por uma biblioteca pré-calculada. Se os fragmentos não estiverem presentes nesta biblioteca então é aplicado um método estocástico como o descrito acima. Essa metodologia é recomendada para a aplicação em bases de dados muito grandes (106).

OpenEye Scientific Software

O OMEGA (109), da OpenEye Scientific Software (http://www.eyesopen.com) realiza uma busca sistemática rápida, baseada em conhecimentos *a priori*. Ele é baseado em regras pré-definidas que descrevem as características torsionais de fragmentos. O algoritmo utilizado divide a molécula em fragmentos conectados, os quais podem conter entre uma e cinco ligações rotacionáveis e estes fragmentos são submentidos a uma minização de energia com o campo de força MMFF94 (Merck Molecular Force Field). Conformações para cada fragmento são geradas através de uma biblioteca de ângulos torsionais pré-definidos, os quais são recolocados na estrutura inicial através de regras químicas e geométricas, visando construir várias conformações para cada molécula. Conformações duplicadas ou com sérias restrições estéricas são removidas e as restantes são usadas como base para uma amostragem torcional mais refinada, gerando um

número fixo de conformações (definidas pelo usuário) dentro de um intervalo de energia.

Schrödinger

O MacroModel, da Schrödinger (http://www.schrodinger.com) é um pacote de modelagem baseado em campos de força moleculares que oferece uma variedade de métodos para a geração de conformeros (110). Entre eles estão métodos sistemáticos e estocáticos e os métodos conhecidos com o LMCS (low-mode conformational search) e LLMOD (large scale low-mode) que usam métodos baseados em modelos matemáticos como autovetores para a realização de amostragens conformacionais. A otimização geométrica no MacroModel pode ser feita utilizando diferentes pacotes de campos de força tais como MM2, MM3, MMFFs, AMBER, AMBER94 e muitos outros. Além disso, esse programa pode usar modelos de solvatação de moléculas na tentativa de obter conformações bioativas de forma mais acurada. A desvantagem desse método para o VS é o alto tempo necessário à realização cálculos o que o torna inviável para a utilização em grandes bases de dados.

1.8 Modelagem de dados

Um modelo pode ser entendido como o conjunto de características mínimas que necessárias para a qualificação ou quantificação de um objeto. No tratamento de estruturas de pequenas moléculas e proteínas um modelo é o conjunto de relações matemáticas que são calculadas para tentar predizer resultados experimentais ou atividade de uma estrutura com base em suas propriedades estéricas fisicoquimicas. Quando referido a pequenas moléculas o método usado para definir essa modelagem é o QSAR (Quantitative Structure-Activity Relationship).

Qualquer método de QSAR pode ser definido como uma aplicação de métodos matemáticos e estatísticos para encontrar relações empíricas do tipo $P_i = k'(D_1, D_2, ..., D_n)$ onde P_i é a atividade biológica (ou outra propriedade de interesse) da molécula, $D_1, D_2, ..., D_n$ são propriedades estruturais calculadas (ou medidos experimentalmente) dos compostos, e k' é um fator de correção que pode ser aplicado (111). Em geral, a preparação de uma base de dados para a geração de um modelo passa por duas fases:

1. Preparação dos dados: geralmente inclui a coleta dos dados das estruturas, des-

carte de informações desnecessárias, cálculo de descritores moleculares, fusão dos descritores (se necessário) para tornar os dados manuseáveis.

2. Geração do modelo: utilização de métodos estatísticos para a identificação de propriedades de representem o conjunto de dados. As técnicas mais utilizadas são a regressão linear, regressão múltipla, regressão logística, SVD (Singular Value Decomposition) e Machine Learning.

Geralmente as bases de dados são divididas em grupo(s) treino e grupo(s) teste, sendo que os grupos teste são usados para avaliar a acurácia do modelo gerado com os dados do grupo treino. Em modelos atuais usa-se dividir a base de dados em múltiplos grupos mutuamente exclusivos para realizar a validação dos modelos, isso é chamado de k-particionado ("k-fold"), onde k é o número de grupos no qual a base foi particionada. Em geral, 1 dos grupos é usado para teste e k-1 grupos são usados para treino (validação cruzada). Não há um valor padrão de k para ser utilizado, apesar de o valor k = 10 ser muito encontrado. Entretanto, quanto maior o valor de k maior será o número de cálculos a ser realizado, por isso é conveniente que uma otimização entre o valor de k e o tempo computacional necessário seja realizada. Pode também acontecer de os k grupos nos quais a base foi dividida sejam reagrupados (inclusive testando todas as combinações possíveis) para a geração múltiplos grupos treino e teste.

Após a geração e a triagem dos modelos mais bem sucedidos, é necessária a realização de uma validação externa. Isso é feito com a aplicação do modelo a um conjunto de dados independente, que não tenha sido utilizado na geração do modelo ou nos grupos de teste. De acordo com **Tropsha (2010)** (111), um modelo não é confiável sem a utilização de uma validação externa. Caso não haja um grupo externo para ser utilizado, existem técnicas matemáticas, como o bootstrap, para isso. A técnica de bootstrap consiste na retirada aletória de amostras de uma base que passa a serem utilizadas como grupo externo, e o restante dos dados passa a ser utilizado para treino e teste. A cada passo os dados anteriormente retirados são repostos e uma nova amostra é aleatóriamente retirada.

1.9 Farmacóforos

O conceito de farmacóforo é baseado na suposição de que o reconhecimento molecular entre um alvo biológico e uma família de compostos pode ser descrito por um conjunto de características comuns a estes compostos, geralmente associadas à interação de um ou mais fragmentos moleculares com o sítio de interação do alvo (112).

De acordo com recente definição da IUPAC, um modelo farmacofórico é "um conjunto de características estéricas e eletrônicas necessárias para garantir a ótima interação supramolecular com um alvo biológico para provocar (ou bloquear) a resposta biológica"(113).

Ainda sobre o conceito de farmacóforo é necessário lembrar que o farmacóforo não é a representação real de uma molécula ou uma associação real com grupos funcionais, mas apenas uma abstração que descreve complementaridades entre um composto bioativo e um alvo de interesse, sejam elas estéricas ou eletrostáticas, (114). De acordo com **Drie (2010)** (115) pode-se descrever o farmacóforo de uma forma mais simples, como "o arranjo espacial de características químicas essenciais para a atividade biológica", isto é, um padrão apresentado por um conjunto de moléculas bioativas.

Primeiramente atribuído a Paul Erlich, hoje se reconhece que a primeira pessoa a definir o termo farmacóforo foi Monty Kier, em uma série de artigos publicados entre 1968-1971 (116) (117) (118) (119). Além de deduzir e calcular os primeiros farmacóforos de forma manual (**Fig. 1.8**), auxiliado apenas por ferramentas simples interativas de visualização molecular gráfica (modelos moleculares), ele também mapeou o processo hoje conhecido como 'ligand-based design', inspirando vários pesquisadores da época, inclusive Garland Marshall, co-fundador da Tripos (115). De acordo com **Leach et al. (2010)** (60) isso foi possível somente porque Kier estudava moléculas com flexibilidade limitada para a geração desses farmacóforos.

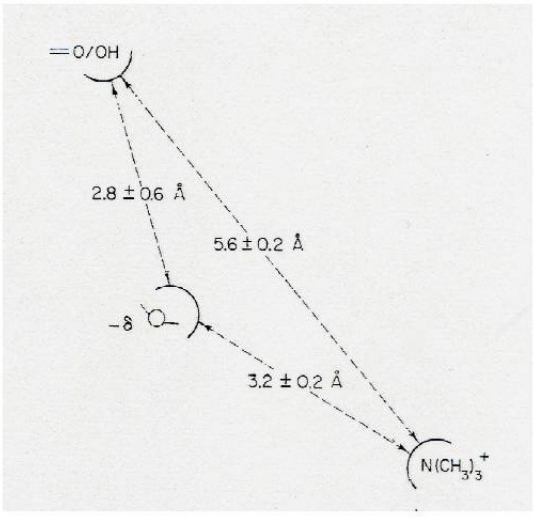


Figura 1.8: O primeiro modelo de farmacóforo para agonistas muscarínicos publicado por Kier em 1967 (117)

Hoje, o conceito básico de farmacóforo como uma representação geométrica e físico-

química das chaves para a interação molecular permanece inalterado (60), mas o seu significado e área de aplicação tem se expandido consideravelmente. Nas últimas décadas surgiram sofisticados algoritmos para a elucidação, manipulação e uso dos modelos de farmacoforos que vem ganhando espaço nas pesquisas de desenvolvimento racional de fármacos, talvez devido às decepções com as triagens virtuais feitas com docking e outras metodologias baseadas em funções de ranqueamento ("scoring functions") (112).

Os métodos de modelagem de farmacóforos tem despertado bastante interesse nos últimos anos devido principalmente ao crescimento constante das estruturas tridimensionais de alvos depositadas. Apesar de necessitar do mesmo nível de informação da ancoragem molecular eles demandam menor tempo de execução com muito mais eficiência (112).

Um modelo farmacofórico clássico pode ser obtido pela resolução de um alinhamento estrutural de moléculas que apresentam uma determinada função biológica, o resultado dessa superposição fornece similaridades que determinam quais características farão parte do modelo (**Fig. 1.9**) . Ainda, o modelo de farmacóforo pode ser inferido da estrutura tridimensional do sítio ativo de uma proteína através da projeção na sua cavidade das características ideais a um ligante para estabelecer interações com o receptor. È importante salientar que essa abordagem tende a dificultar o trabalho de análise por gerar pontos potenciais em excesso no modelo de farmacóforo. A escolha de qual abordagem será utilizada dependerá do tipo de informação que o pesquisador tiver em mãos.

Os programas para a elucidação de farmacóforos hoje são amplamente usados. Eles costumam ser disponibilizados junto com como pacotes de softwares comerciais para o descobrimento racional de fármacos. Entre os principais estão o CATALYST, GALAHAD, GASP e módulos para cálculos de farmacóforos no MOE, PHASE e CHEMAXON (60). Na ultima década surgiram muitos programas para este tipo de cálculo mostrando que a elucidação desse problema ainda não foi considerada resolvida.

1.10 *Fingerprints* - Assinatura digital

Esta técnica é bastante utilizada na computação para caracterização e descrição de um objeto. No contexto da quimio-bioinformática, as "*fingerprints*"são vetores únicos, binários ou não, que codificam de forma simplificada informações variadas e que permitem produzir resultados rápidos. Em VS, o tipo de fingerprint mais simples é o vetor binário, onde cada posição é associada univocamente a uma ocorrência de subestrutura,

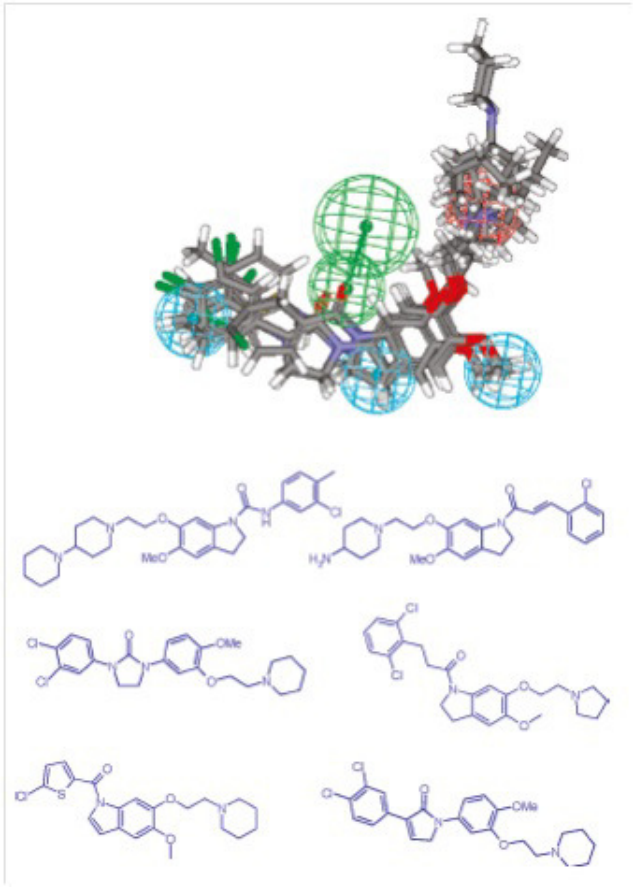


Figura 1.9: Uma ilustração do conceito básico de farmacóforo. Na figura, os antagonistas de recpetor 5HT2C foram alinhados gerando o modelo farmacofórico 3D; Em verde estão representados os grupos aceptores hidrogênio, vermelho os ionizados positivamente carregados e em ciano os grupos hidrofóbicos (Andrew R. Leach, 2010).

fragmento, arranjo de átomos ou farmacóforos, e o estado do bit (aceso ou não) designa a presença ou ausência daquela característica na molécula representada. Caso se deseje agregar informações à fingerprint que extrapolem a sua natureza binária, o vetor pode armazenar o número de ocorrências de cada representação, tornando-se uma fingerprint de frequência, que ainda pode ser normalizada, transformando os contadores em pesos que variam entre zero e um, ou qualquer outro tipo de normalização.

Além de representar uma única molécula, fingerprints podem ser usados para representar um conjunto de moléculas diferentes. Ao incorporar a informação de múltiplas estruturas, sistemas baseados em fingerprints podem condensar os dados em uma única representação através de fingerprints modais, onde um bit é acesso se sua frequência nas moléculas está acima de um determinado corte, ou ponderadas, onde cada bit recebe um peso de acordo com sua frequência (120).

As fingerprints modais podem ser consideradas um exemplo de fingerprint fuzzy, uma vez que, assim como na Lógica Fuzzy, uma premissa varia em grau de verdade de 0 a 1. Ou seja, se as informações totalmente falsas recebem valor 0 e, informações totalmente verdadeiras recebem valor 1, as características passam a poder serem classificadas como parcialmente verdadeiras ou parcialmente falsas, recebendo valores de acordo com o grau de pertinência a um conjunto (grupo)(121) (122). Este tipo de fingerprint é útil ao tratamento de informações do mundo real, pois elas podem ser vagas, ambíguas e qualitativamente incompletas e imprecisas. Assim, sua aplicação aos conjuntos de dados de farmacóforos pode revelar características marcantes (comuns) entre eles. As grandes vantagens em se aliar essas técnicas são as formas de representação dos dados que, apesar de simples, possuem notável capacidade de armazenamento da informação estrutural de moléculas, o que permite, caso seja necessário, mapear todas as interações em uma estrutura orgânica complexa como a das proteínas por exemplo.

1.11 Processo de desenvolvimento de um novo Fármaco

O processo de desenvolvimento de novos fármacos (termo em inglês, "Drug Discovery") é demorado, complicado e dispendioso. Em média o tempo para que uma molécula um candidata a fármaco chegue ao mercado é de 10 a 17 anos, incluindo desde o descobrimento do alvo até o registro do medicamento e gasta-se, em média, de 1-2 bilhões de dólares. Por isso os estudos de VS são tão essenciais por diminuírem o tempo final pelo direcionamento dos estudos e aumento das probabilidades de encontrar novas moléculas ativas.

Inicialmente é escolhida uma enfermidade ou atividade biológica a ser tratada ou modulada, baseado em estudos de mercado ou casos de necessidade de desenvolver tratamentos para alguma epidemia, enfermidade ou anomalia. O direcionamento do projeto também pode ser dado por conhecimento empírico de efeitos biológicos provocados por substâncias ou plantas. Como exemplo deste último caso podemos citar o caso recente do Acheflan ® Cordia verbenacea DC, analgésico e anti-inflamatório tópico cujos efeitos da planta do qual é fabricado (Cordia verbenacea DC, popularmente conhecida como "Maria-milagreira") há muito tempo eram conhecidos em regiões do litoral brasileiro do qual é nativa.

Iniciado o projeto, o desenvolvimento de um fármaco pode ser divido em duas fases principai: a pré-clínica e a clínica. Na fase pré-clinica são realizados estudos para a

identificação dos alvos biológicos e moléculas potencialmente ativas através de testes *in silico*, *in vitro* ou utilizando modelos biológicos animais. Nesta etapa, mais de 90% das substâncias estudadas são descartadas por não apresentarem atividade terapêutica suficiente ou por apresentarem toxicidade elevada.

Definidas as estruturas com maior atividade iniciam-se as etapas de otimização das moléculas buscando a maior relação efeito/toxicidade. O tempo necessário para a realização de todos esses experimento varia de 3,5 a 7 anos e, no fim, sobram poucas moléculas promissoras e estas são submetidas aos testes clínicos.

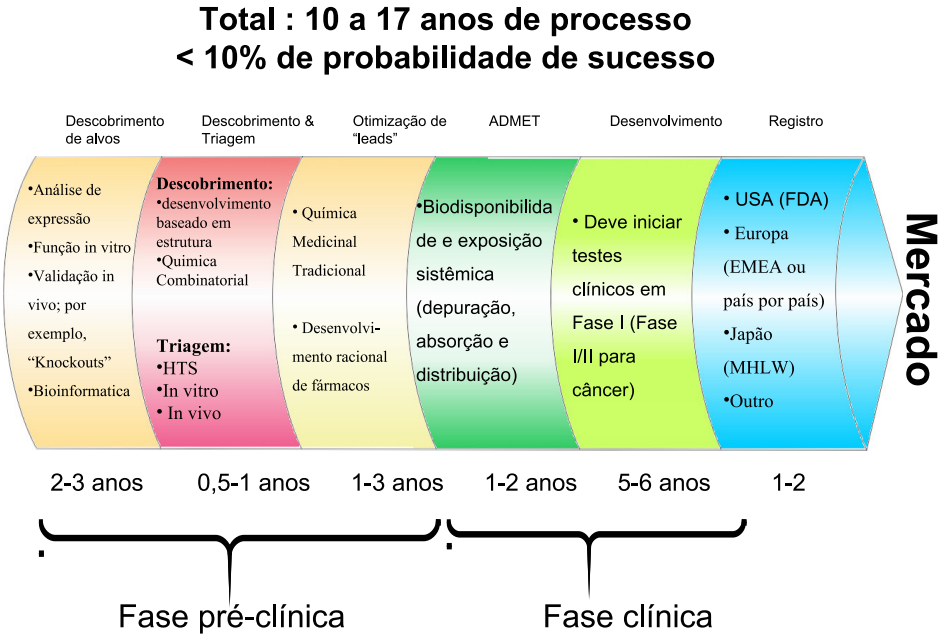


Figura 1.10: Processo de desenvolvimento de Fármacos. Adaptada de (123).

A fase clínica do desenvolvimento de um fármaco pode ser dividida em:

- Fase I = estudos de toxicidade e dosagem segura
- Fase II = estudos de eficácia em escala reduzida
- Fase III = estudos de eficácia em maior escala
- Fase IV = farmacovigilância, o medicamento continua sendo avaliado após o registro e o lançamento.

Na primeira fase, um pequeno grupo de voluntários (20 a 100 indivíduos saudáveis) recebe doses crescentes de um novo medicamento. Isto é feito com o intuito de avaliar

questões de segurança, tolerância e metabolização dessa nova substância em humanos saudáveis. Possibilitando com isso obter valores aproximados da maior dose tolerável, menor dose efetiva, relação dose/efeito, duração dos efeitos e observação de efeitos colaterais. Assim, ao final dessa etapa podem ser estimados alguns parâmetros farmacocinéticos no ser humano (como metabolismo e biodisponibilidade), dose e posologia da droga.

Fase II = Estudo Terapêutico Piloto

Este é o primeiro estudo do medicamento em indivíduos doentes. O objetivo principal dessa fase é demonstrar a efetividade potencial da medicação, para isso, são selecionandos entre 100 e 200 pacientes para serem submetidos à nova substância. Nessa fase são averiguados parâmetros de segurança em relação à administração, relações dose-resposta e intervalos adequados entre a administração das doses do novo fármaco. Aqui, através de estudos de biodisponibilidade e bioequivalência é possível determinar a dose ótima de administração do medicamento, que é aquela onde se consegue o melhor efeito terapêutico combinado ao menor conjunto de reações adversas.

Fase III = Estudo Terapêutico Ampliado

Esta é a fase na qual o medicamento é administrado em um número grande de pacientes para se avaliar novamente a eficácia e a segurança do produto. O número de pacientes pode variar de dezenas a milhares, dependendo do tipo de patologia. A avaliação é sempre feita de maneira comparativa, utilizando-se outro tratamento de referência, e realizada em condições praticamente normais às de administração. São analisados nessa fase o risco/benefício do princípio ativo em curto prazo, cuidados na utilização, estudo dos eventos adversos, interações medicamentosas, fatores modificadores do efeito tais como sexo, idade e raça.

Fase IV = Farmacovigilância

Também denominada Pesquisa Pós-Comercialização, essa fase é posterior ao registro e ao lançamento do novo medicamento. Quando o medicamento passa a ser comercializado, devido ao grande número de pessoas que passam a se utilizar do me-

dicamento novas indicações, efeitos raros e consequências do uso prolongado do medicamento podem ser detectados. Nas fases anteriores devido às limitações de tempo e número reduzido de pessoas e ao controle de seleção desses indivíduos (não podem estar entre os selecionados aqueles que possam representar riscos para o estudo por apresentarem problemas clínicos ou outras situações que prejudiquem avaliação dos resultados do tratamento como patologias que não sejam alvo direto dos testes, uso concomitante de outras drogas, ou pertencimento a grupos populacionais específicos como grávidas, crianças e idosos .

Portanto, os Estudos de Fase IV, são essenciais ao processo de detecção, acompanhamento e controle de problemas decorrentes do uso de medicamentos legalmente autorizado sendo essenciais aos medicamentos novos, pois proporcionam a avaliação do seu uso em grandes populações. E devem ser desenvolvidos de forma a evitar interesses privados, sendo necessária a implementação de controles para garantir a isenção do processo.

1.12 A crise das Indústrias Farmacêuticas no século XXI

A última década foi marcada por turbulências e mudanças para as indústrias farmacêuticas. Apesar de terem sido anos de muita inovação e descobertas, o cenário do mercado farmacêutico mundial revela incertezas, perdas e falências. Na tentativa de minimizar os prejuízos as grandes empresas vêm fazendo fusões e aquisições de outras empresas. Como consequência direta disso, na última década houve uma reorganização das posições no ranking das 10 maiores empresas farmacêuticas. Das dez maiores empresas em 2001, oito permaneceram no topo do ranking e das dez maiores em 2010, seis realizaram fusões e/ou aquisições. Apesar disso, o valor total dessas dez maiores empresas que em 2001 era de U$ 1,240 bilhões e caiu para U$ 1,110 bilhões em 2011 (124).

Além de fusões e/ou aquisições as indústrias vêm usando outras estratégias como a utilização de capital externo, direcionamento das pesquisas para áreas onde o novo fármaco tem maior chance de ser inovador e maior critério de seleção das moléculas que serão submetidas ao processo de desenvolvimento de fármacos.

Ainda, outro fator preocupante é que nos últimos anos vêm caindo as patentes de medicamentos já consolidados no mercado. Assim, fórmulas às quais se tinham direitos passaram a ser de domínio público e esse capital deixa de ir para a empresa. Uma forma

Tabela 1.7: Ordem das Empresas Farmacêuticas por capitalização de mercado (valor agregado da companhia) adaptada de (124)

Empresa	Capitalização de mercado (US$ bilhões)* em 2011	2011 ranking	2001 ranking	Principais fusões ou aquisições desde 2001
Johnson & Johnson	178	1	2	-
Pfizer	154	2	1	Pharmacia,Wyeth
Novartis	153	3	5	-
GlaxoSmithKline	113	4	3	-
Roche	110	5	10	Genentech
Merck & Co˙	110	6	4	Schering Plough
Sanofi	91	7	9	Aventis, Genzyme
Abbott	85	8	>10	Solvay
AstraZeneca	61	9	8	MedImmune
Bayer	55	10	>10	-
Bristol-Myers Squibb	-	>10	6	-
Lilly	-	>10	7	-

de fugir disso vem sendo o reposicionamento de fármacos, que consiste da utilização de medicamentos antigos em novas aplicações terapêuticas. Como exemplo, temos o Ácido Acetilsalicílico que por muito tempo foi utilizado com um antiinflamatório não esteroidal (AINE) e nas últimas décadas passou a ser utilizado no tratamento de disfunções cardíacas.

Concomitantemente a tudo isso, o mercado farmacêutico vem investindo de forma crescente em Pesquisa e Desenvolvimento (P&D) (**Fig. 1.11**) mas, as indústrias não conseguiram fazer com que a taxa de sucesso de um candidato a fármaco ir ao mercado aumentasse na mesma proporção desse dinheiro investido. De acordo com a *U. S. Food and Drug Administration* (FDA), este problema existe pois o caminho atual para o desenvolvimento de fármacos tem se tornado extremamente desafiador, ineficiente e caro (125). A **Fig. 1.12** mostra a razão entre o lucros obtidos com a venda de novos medicamentos e o capital investido pelas empresas em P&D (126) na última década.

As etapas mais preocupantes são os testes clínicos, principalmente as fases I e II do processo de desenvolvimento de fármacos. No comparativo de 2002 para 2008 houve aumento da taxa de rejeição de novos candidatos a fármaco. Como pode ser observado na **Fig. 1.13**, a probabilidade de um composto chegar ao mercado caiu de 10% para 5% na fase I e de 17% para 11% na fase II em pouco mais de 6 anos. Essas etapas representam as fases onde se confirma a efetividade e se avalia a toxicidade

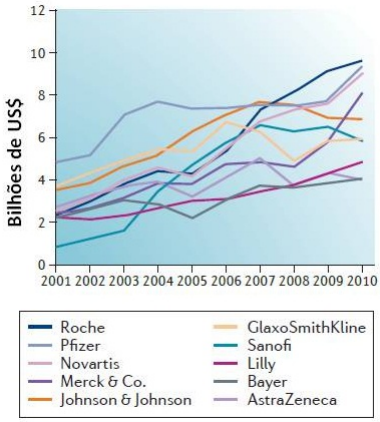


Figura 1.11: Investimento em P&D das principais indústrias farmacêuticas na década passada. Adaptada de (124)

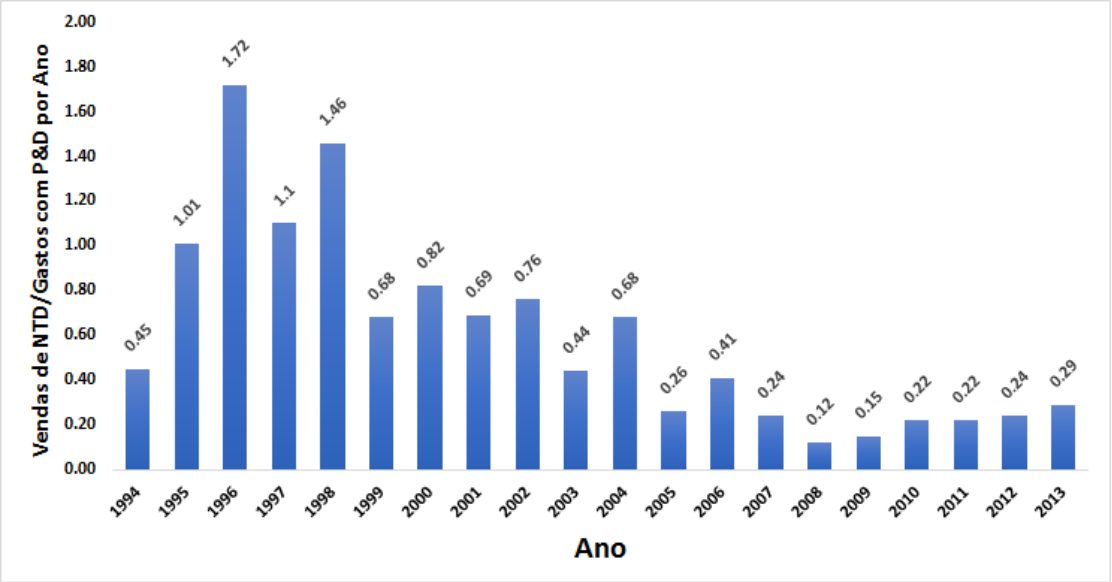


Figura 1.12: Razões entre os lucros obtidos com a venda de novas drogas terapêuticas (New Therapeutic Drugs, NTDs) sobre os gastos com P&D. Adaptada de (126).

das novas moléculas. Isso forçou as indústrias a adotarem novas estratégias como redução das áreas terapêuticas de atuação, questionar prospectos de candidatos que tem a probabilidade de não ser o melhor na área ou um medicamento inovador e focar a linha de pesquisa em poucos candidatos, aqueles que parecerem ser os mais robustos cientificamente (124).

Por isso novas formas de identificar moléculas com potencial de atividade em bases de dados são cada vez mais estudadas e encorajadas. No Brasil, por exemplo, o Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), por meio do Laboratório Na-

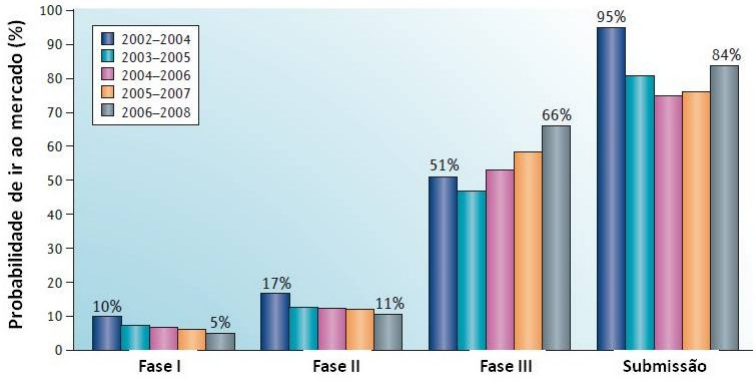


Figura 1.13: Probabilidade de um candidato a fármaco chegar ao mercado dada a fase do desenvolvimento em que o mesmo se encontra. Nota-se a queda constante através dos anos nas Fases I (Toxicologia e II (Eficiência). Adaptada de (124)

cional de Biociências (LNBio) e a Rede Nacional de Métodos Alternativos (RENAMA) lançaram recentemente (janeiro de 2014) uma chamada para seleção de propostas de testes in sílico voltados à predição de propriedades toxicológicas e farmacocinéticas de pequenas moléculas naturais ou sintéticas, candidatas a fármacos e cosméticos. Este tipo de predição computacional representa uma alternativa ou um complemento ao uso de animais para testes biológicos e reduz o tempo e os investimentos requeridos ao desenvolvimento de fármacos e cosméticos. Foram selecionadas 7 propostas, a maioria de indústrias farmacêuticas nacionais (Cristália, Eurofarma, Boticário e Natura), dois centros de pesquisa governamentais (Fundação Oswaldo Cruz e Farmanguinhos), mas apenas uma instituição de ensino e pesquisa (Única Educacional) (127) [1].

[1] **Fonte:** Laboratório Nacional de Biociências. <http://www.brasil.gov.br/ciencia-e-tecnologia/2014/04/projetos-propoem-testes-alternativos-ao-uso-de-animais-em-pesquisas>, acessado em 22/04/2015

Capítulo 2

Objetivos

2.1 Objetivos Gerais

- Desenvolver uma ferramenta que permita realizar análises *in sílico* de similaridade de sítios ativos em larga escala.
- Desenvolver uma ferramenta que permita realizar triagem virtual de ligantes em grandes bases de dados.

2.2 Objetivos específicos

- Aplicação dos conceitos de fingerprints de farmacóforos à analise de similaridade de pequenas moléculas e de sítios ativos.
- Aplicação de métricas de similaridade para estabelecer quais estruturas moleculares em uma base de dados são potencialmente ativas ou não;
- Desenvolver método de modelagem de dados referente a ligantes e sítios ativos;
- Desenvolvimento de um método de validação de modelos que consiga realmente produzir modelos confiáveis e robustos.

Capítulo 3

Metodologia

A elucidação e disponibilização de estruturas proteicas (desde as sequências primárias até suas conformações tridimensionais) são de extrema importância para o desenvolvimento das ciências da vida, pois esses dados constituem uma informação valiosíssima para a determinação das suas funções biológicas e para o esclarecimento dos importantes mecanismos biológicos envolvidos. Por isso, visando a melhor compreensão dessas informações, métodos computacionais de análise de dados vêm sendo cada vez mais utilizados como ferramenta de auxílio aos cientistas na busca da resolução de problemas ligados à saúde.

Nesse sentido, nas últimas décadas, devido à necessidade de uma nova forma de análise de dados biológicos foi desenvolvida a Bioinformática e, nesse pouco tempo já vem trazendo grandes contribuições. Isso tem sido possível devido ao leque do conhecimento aplicado na Bioinformática ser bastante amplo, indo desde o conhecimento biológico até a mais pura estatística e, claro, sendo essencial um bom conhecimento computacional para automatizar cálculos e escrever sofisticados algoritmos. Tal peculiaridade faz com que ela seja utilizada em várias áreas do conhecimento de variadas formas, mas, geralmente, estando associada à resolução de grandes problemas, que envolvem muitas variáveis ou ao tratamento de grandes quantidades de dados.

Como exemplo de aplicação da Bioinformática está a Bioinformática Estrutural que, focada nos aspectos físicos do genoma e das proteínas, pode ser aplicada na análise de estruturas de proteínas em larga escala e, às vezes, tendo que lidar com o proteoma completo de um organismo. A Bioinformática estrutural trabalha principalmente com modelos biológicos tridimensionais obtidos por cristalografia ou ressonância magnética nuclear. Mas, nas últimas décadas, com o aperfeiçoamento das técnicas de modelagem por homologia e outras técnicas, tornou-se possível a construção de mo-

delos tridimensionais de proteínas de alta qualidade utilizando apenas informações de suas sequencias primárias os quais também podem ser utilizados em experimentos de Bioinformática Estrutural. Tais modelos podem ser utilizados para estudos de desenho racional de fármacos e inferência de características e funções biológicas de proteínas o que, de certa forma, também se tornou um fator retro-estimulante à sua criação. Como consequência direta disso houve um grande aumento do número de modelos tridimensionais de proteínas depositados em bases de dados internacionais nas últimas décadas.

Além do conhecimento das estruturas dos receptores biológicos, o conhecimento dos ligantes ativos para um determinado alvo também é muito importante. Esses dados estão disponíveis em grande número e diversidade e estão armazenadas em bases de dados e, portanto, passíveis de serem utilizadas em experimentos computacionais. As industrias farmacêuticas são o exemplo disso, pois há várias décadas utilizam-se de informações de ligantes em experimentos in silico para auxiliar no processo de desenvolvimento racional de novos fármacos.

Hoje, devido à produção rápida e intensa de dados químicos e biológicos, há uma quantidade maior de dados disponível para a realização de estudos computacionais. Entretanto, o fato de existirem tantos dados faz com que a realização de ensaios de bancada para determinação de funções biológicas seja insuficiente para analisar minuciosamente toda essa gama de informações. Por isso, uma alternativa viável são os estudos *in silico*, que permitem a realização rápida de experimentos para análise de grande quantidade de dados, e com o desenvolvimento de algoritmos eficazes que permitem a obtenção de resultados cada vez mais acurados, vêm sendo cada vez mais estimulados tanto no meio educacional quanto nas grandes indústrias.

Tal interesse nesses estudos computacionais devem-se ao fato de eles permitirem um direcionamento nos experimentos de bancada agilizando o processo de descoberta. No âmbito do desenvolvimento racional de fármacos, para aplicações diretas nas indústrias farmacêuticas, as informações estruturais e físico-químicas dos ligantes, aliadas ao conhecimento de estruturas de alvos biológicos, permitem a identificação de compostos potencialmente ativos em grandes bases de dados. Assim, essa triagem *in silico* de pequenas moléculas é de extrema importância para o processo industrial, pois permite a redução dos custos no processo de desenvolvimento racional de fármacos que é altamente dispendioso. Tal possibilidade de economia decorre do descarte precoce de moléculas que não apresentam potencial de atividade ou que não apresentem perfil farmacocinético adequado, reduzindo os custos das etapas posteriores e aumentando as

chances de identificação de novas substâncias bioativas que, com sorte, poderão originar um novo fármaco.

Por isso, propomos neste trabalho uma forma de análise de arquivos proteínas e de pequenas moléculas, que nos permitirá obter resultados de similaridade de sítios ativos e de ligantes de forma rápida e, no futuro, poderá nos permitir fazer inferências mais precisas quanto à classe de uma proteína (obtida por cristalografia ou homologia), mesmo que essa não tenha ainda associação biológica experimentalmente provada, inferir sobre novas funções para proteínas conhecidas e identificar em bases de dados de ligantes aqueles com maior probabilidade de apresentarem alguma atividade biológica relevante.

Deste modo, as metodologias que seguem visam a transformação de dados tridimensionais de ligantes e proteínas de forma a simplificá-los, através de representações vetoriais que permitam comparar e mensurar as suas igualdades e/ou diferenças. Esse tipo de metodologia possibilita a realização de testes *in silico* onde, ao escanear grandes bases de dados tomando como parâmetro certas características ou padrões de informação conhecidos, possamos ter a indicação de estruturas (alvos ou ligantes) potencialmente ativas ou não, com capacidade de desencadear efeitos biológicos ou não.

3.1 Similaridade de Sítios Ativos de proteínas

As proteínas são macromoléculas orgânicas de estruturas complexas formadas por resíduos aminoácidos unidos por ligações peptídicas. Embora elas possuam inúmeras cavidades e reentrâncias em sua superfície, a sua função biológica está geralmente associada a apenas uma delas, aquela que oferece as propriedades necessárias para a ligação a outras proteínas ou ligantes, que é chamada de sítio ativo.

Assim, por sua importância, a similaridade entre sítios ativos é bastante estudada e algoritmos sofisticados para o cálculo de similaridade entre eles têm sido desenvolvidos. As informações de sítios ativos são também disponibilizadas gratuitamente na internet, em grande quantidade e diversidade, contendo dados de entidades que podem estar na forma de estrutura primária, secundária, terciária ou quaternária. Todas essas formas podem ser utilizadas para a comparação de sítios ativos, embora as estruturas 3D (terciária e/ou quaternária) ofereçam maior informação sobre esses sítios.

Neste trabalho, para os cálculos de similaridade de sítios ativos serão utilizadas informações oriundas de estruturas cristalográficas de proteínas depositadas no PDB ou scPDB. Para isso foi desenvolvido no NEQUIM (Núcleo de Estudos em Quimioinfor-

mática - UFMG) um pacote de ferramentas chamado "PharmaSite"que se encarrega de transformar as informações tridimensionais do sítio ativo em uma fingerprint de farmacóforos e, também, de quantificar a similaridade entre esses fingerprints.

As metodologias escolhidas e aplicadas neste trabalho, assim como os métodos tradicionais para a realização de cálculos de similaridade de sítios ativos (principalmente aqueles destinados à realização de estudos em larga escala), prezam por uma representação simplificada dos resíduos da cavidade, o que é feito com a intenção de reduzir a complexidade dos cálculos e o custo computacional.

Além disso, a aplicação do conceito de farmacóforos aliado aos fingerprints torna o método ainda mais abstrato, refletindo em uma redução significativa do custo computacional, principalmente devido à exclusão da necessidade de alinhamento das estruturas dos sítios ou conversão das coordenadas geométricas dos farmacóforos para um mesmo sistema de coordenadas 3D. Esses fatores são os que geralmente mais desabonam a maioria dos programas que calculam similaridade de sítios, por necessitarem cálculos mais complexos e, consequentemente, de mais tempo de processamento.

Para a utilização da metodologia que desenvolvemos, o primeiro passo é a escolha das entidades que serão estudadas. Uma vez escolhidas as estruturas protéicas que constituirão o conjunto de dados, o tratamento e análise de dados para cálculos de similaridade dos sítios ativos passam pelos seguintes passos:

1 Construção do modelo formacofórico do sítio pelo mapeamento dos aminoácidos expostos à cavidade do sítio ativo;

2 Geração da fingerprint de farmacóforos potenciais do sítio;

3 Análise de similaridade dos sítios utilizando os fingerprints de farmacóforos. Esse passo será discutido à frente, pois o mesmo método é utilizado na comparação de fingerprints geradas a partir de ligantes.

3.1.1 Geração dos farmacóforos do Sítio

Os farmacóforos gerados pelo PharmaSite podem ser obtidos por duas formas. Uma com a utilização de um software externo, neste trabalho utilizamos o software comercial THINK (© Treweren Consultants), para o qual conseguimos uma licença acadêmica e outra com a utilização das informações contidas em arquivos PDB.

Através do THINK são gerados pontos nas cavidades dos sítios ativos associados às características farmacofóricas desejáveis para a ótima interação supramolecular de

um ligante ideal. Deste modo, são projetados pontos na cavidade seguindo o conjunto de características fisicoquímicas e estéricas que o ligante perfeito deveria ocupar. A cada uma desses pontos são atribuídas as respectivas propriedades farmacofóricas que podem ser uma ou mais das seis possíveis propriedades farmacofóricas utilizadas pelo THINK :

- Aceptor de Hidrogênio (HACC ou A);
- Doador de Hidrogênio (HDON ou D);
- Positivamente ionizado (POS ou P):
- Negativamente ionizado (NEG ou N);
- Aromático (AROM ou R);
- Alifático (LIP ou H):

Ao final dos cálculos o Think gera um arquivo formatado, que identifica pontos espaciais nos sítios ativos associados às suas características farmacofóricas. Este arquivo será utilizado pelo PharmaSite para a geração das fingerprints.

Caso seja desejado, o PharmaSite pode calcular um conjunto de pontos no sítio ativo e atribuir suas respectivas propriedades farmacofóricas. Para isso é necessário que haja um ligante cocristalizado ou que seja feita uma identificação prévia dos aminoácidos que constituem o sítio de ligação, como os dados depositados no scPDB. Os resíduos de aminoácido de um sítio de ligação são selecionados se qualquer um de seus átomos pesados apresentar uma distância menor ou igual a um dado ponto de corte de qualquer um dos átomos pesados do ligante co-cristalizado. Após essa seleção as características farmacofóricas são atribuídas às coordenadas cartesianas dos carbonos alfa desses resíduos selecionados de acordo com a **Tabela 3.1**. Também, caso seja desejado, os resíduos que constituem o sítio da proteína podem ser selecionados manualmente, ou através de um script externo. Esses resíduos devem ser salvos em um arquivo, em formato PDB, e submentido ao PharmaSite. A atribuição das propriedades farmacofóricas será feita da mesma forma descrita anteriormente.

Os carbonos alfa foram escolhidos como ponto representativo dos resíduos de aminoácido pelo fato de as cadeias peptídicas das proteínas ("backbone") tenderem a ser mais conservadas estruturalmente em relação aos átomos que constituem as cadeias laterais dos resíduos de aminoácido (que tem maior mobilidade). Ainda, sabendo que

a atividade biológica é dependente da manutenção estrutural dos átomos e/ou propriedades físico-químicas no sítio das proteínas, acreditamos que o "backbone"conserve melhor as informações estruturais do sítio ativo e que seja capaz de descrever as diferenças intrínsecas aos sítios de diferentes proteínas.

Tabela 3.1: Tabela adaptada do trabalho de Weill e Rognan, 2010 (97). Descrição das propriedades farmacofóricas atribuídas aos carbonos alfas dos resíduos de aminoácido. Sendo, A: Aceptor de Hidrogênio, R: Aromático, D: Doador de Hidrogênio, H: Hidrofóbico, P: Positivamente carregado N: Negativamente carregado

Aminoácido	Código 3 letras (128)	Código 1 letra (128)	Propriedades Farmacofóricas
Alanina	Ala	A	H
Arginina	Arg	R	H, D, P
Asparagina	Asn	N	D, A
Ácido aspártico	Asp	D	A, N
Cisteína	Cys	C	H
Ácido Glutâmico	Glu	E	D, A
Glutamina	Gln	Q	A, N
Glicina	Gly	G	-
Histidina	His	H	D, A, R
Isoleucina	Ile	I	H
Leucina	Leu	L	H
Lisina	Lys	K	H, P, D
Metionina	Met	M	H
Fenilalanina	Phe	F	R, H
Prolina	Pro	P	H
Serina	Ser	S	D, A
Treonina	Thr	T	H, A, D
Triptofano	Trp	W	D, R
Tirosina	Tyr	Y	D, A, R
Valina	Val	V	H

Os arquivos de saída nas duas abordagens utilizadas no PharmaSite (software externo THINK e carbonos alfa) possuem o mesmo formato. Portanto, a partir desse ponto os dados serão tratados de forma estritamente igual pelo PharmaSite na construção das fingerprints.

3.1.2 Construção das fingerprints

Fingerprints moleculares são vetores binários constituídos por certo número de descritores, dependente do tipo e do número de propriedades que se pretende capturar (129) (130) (131) (132). Existem propostas muito diferentes e com vários graus de complexidade. Nas representações mais simples, cada bit representa a presença ou ausência de uma característica molecular específica. Em representações mais complexas, podem ser capturadas todas as interações entre átomos de uma molécula ou proteína, e podem ser utilizados para armazenar, buscar e recuperar informações de um banco de dados de estruturas (129) (130) (133).

Geralmente as fingerprints de farmacóforos, codificam a informação sobre a presença ou ausência de pontos de farmacóforos potenciais (PPP - Potential Pharmacophore Points) e as distâncias entre eles em um composto ou conjunto de compostos (134) (**Fig. 3.1**). De modo geral, as fingerprints de farmacóforos codificam agrupamentos ("tuplets") de dois, três ou até quatro PPPs. Tripletos de PPPs são os mais usados já que tradicionalmente são considerados os mais efetivos em termos de conteúdo informacional versus complexidade (134). As distâncias euclidianas são discretizadas em intervalos ou "bins"de distância, e a escolha destes intervalos pode ter um impacto significativo no desempenho do método.

Neste trabalho, para gerar as fingerprints serão calculadas todas as possíveis combinações de uma determinada quantidade de PPPs desejada (o PharmaSite oferece a possibilidade de realizar agrupamentos de 2, 3 e 4 pontos). Para evitar as redundâncias, isto é, diferentes representações para o mesmo conjunto geométrico foi utilizado um algoritmo baseado no conceito de árvores de decisões. Onde, a partir das informações iniciais (PPPs e coordenadas espaciais) são tomadas de decisões sequenciais (chamadas de "nós"), até que se chegue a uma solução ou representação única (chamadas de "folhas") que, neste caso, representará uma posição específica na fingerprint. A complexidade das árvores de decisão é diretamente relacionada ao número de PPPs utilizado. Quanto maior for esse número maior será o caminho a ser percorrido na árvore.

Para chegar até a solução que determinará qual posição da fingerprint será acesa, o algoritmo de árvore de decisões retorna um valor que representa a presença de um conjunto de PPPs. Esse valor é gerado com base nas propriedades farmacofóricas de cada farmacóforo potencial de um PPP e das distâncias entre eles. Para isso, todas as distâncias entre os pontos do conjunto de PPPs são calculadas e cada uma delas é classificada tomando-se por base conjuntos de distâncias pré-definidas, formados por

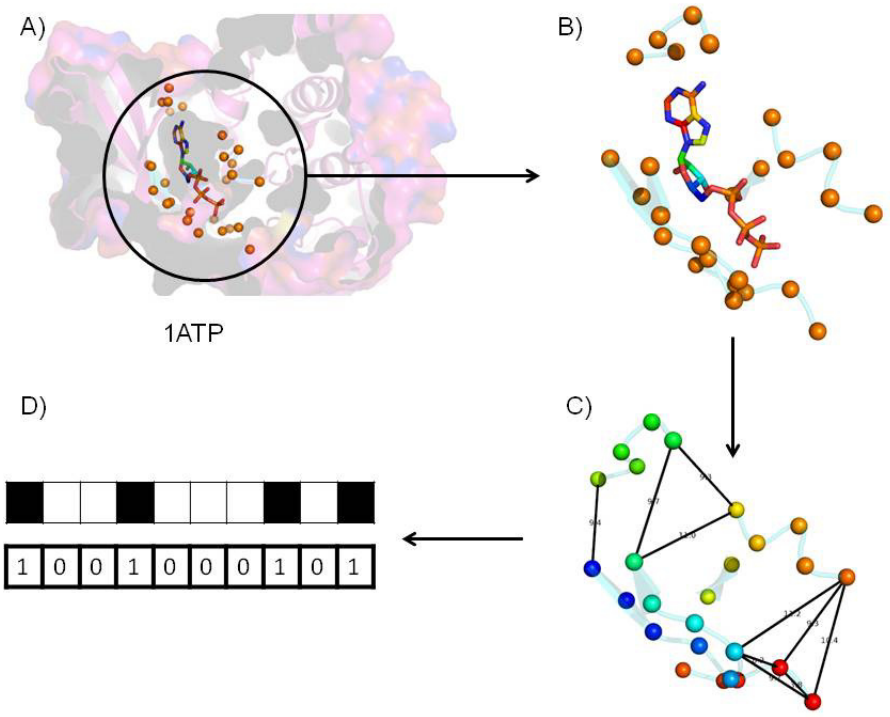


Figura 3.1: Esquema geral para a construção dos farmacóforos e de suas fingerprints. Para a geração das figuras foi utilizada a proteína de pdbID 1ATP co-cristalizada com o ligante ATP. A) A figura a mostra o sítio ativo da proteína 1ATP com o seu ligante, foram selecionados os resíduos de aminoácido que tinham qualquer átomo a uma certa distância de corte de qualquer outro átomo pesado no ligante. B) os carbonos alfa dos resíduos selecionados são separados da proteína. C) São atribuídas as propriedades farmacofóricas e calculados todos os tuplets com 2, 3 ou 4 pontos. D) Os conjuntos de PPPs são codificados em um vetor binário.

intervalos de distâncias ou "bins". O algoritmo atribui um número para a distância calculada se ela estiver compreendida entre os valores máximo e mínimo desse intervalo. Neste trabalho foram utilizados dois conjuntos de distâncias (**Tabela 3.2**). O primeiro deles, chamado de "Normal", possui nove intervalos de distâncias crescentes entre 0 e 21 Å. Esse conjunto foi desenvolvido no NEQUIM como tentativa de promover uma otimização entre a manutenção informacional e a capacidade de diferenciação de conjuntos de PPPs distintos (que podem refletir sítios ativos ou ligantes). O segundo conjunto de distâncias, que foi apresentado no trabalho de citeauthoronlineNathanel-Rognan2010, 2010 (97), é composto por grupo de 5 intervalos de distâncias de 0 a 14.3 Å e, de acordo com o autor, é mais eficiente para a análise de sítios ativos. Também é um dos objetivos deste trabalho a comparação entre estes dois conjuntos de distâncias, para estabelecimento do melhor conjunto, a ser utilizado em trabalhos futuros.

Adicionalmente, implementamos neste programa a possibilidade de uso de

Tabela 3.2: Descrição dos intervalos de distância ("bins") utilizados nos conjuntos "normal"e "rognan".

Conjunto de Distâncias	Intervalos de Distancia ("Bins")								
	1	2	3	4	5	6	7	8	9
Normal	<3	3 - 4,5	4,5 - 6	6 - 8	8 - 10	10 - 12,5	12,5 - 15	15 - 18	18 - 21
Rognan (11)	<4,8	4,8-7,2	7,2 - 9,5	9,5 - 11,9	11,9 - 14,3	-	-	-	-

lógica fuzzy nos cálculos das distâncias entre dois pontos de farmacóforo. Dessa forma, uma distância pode ser classificada em mais de um dos intervalos entre os apresentados na **Tabela 3.2**. Essa lógica fuzzy foi aplica mediante uma função quadrada de pertencimento (**Fig. 3.2**) onde a distância euclidiana final é dada pela função:

$$D_{euclidian} = D_{euclidian} \pm x$$

Nessa fórmula x é um valor de tolerância inserido pelo usuário, para ser aplicada a todas as distâncias nos conjuntos de PPPs. O uso dessa metodologia implica também em um incremento no custo computacional em $3 * n$ operações para 2PPPs, $3^3 * n$ operações em conjuntos de 3PPPs e $3^6 * n$ operações em conjuntos de 4PPPs, onde n é o número de tuplets possíveis em cada uma das metodologias.

A finalidade da aplicação da lógica fuzzy é resolver problemas de distâncias que se situam nas extremidades dos intervalos de distância. Por exemplo, em um intervalo situado entre 2-4 angstrons, se a distancia entre dois PPPs for igual a 3.99 angstrons ela será classificada dentro do intervalo. Entretanto, como as estruturas depositadas no PDB tem uma resolução que indica os possíveis erros de posição de cada átomo, seria correto ser tão restrito nessa classificação? Assim, aplicamos a função acima para tentar resolver esse problema. Também a aplicação dessa fórmula é uma forma de incluir alguma flexibilidade à estrutura cristalográfica rígida de um arquivo PDB, possibilitando que pequenas diferenças conformacionais das proteínas não interfiram significativamente nos cálculos de similaridade.

Ao final o PharmaSite pode gerar seis tipos diferentes de fingerprints através da combinação das três formas de agrupamento de PPPs com os dois conjuntos de distâncias escolhidos ("Normal"e "Rognan") (**Figura 3.3**). Ainda, o algoritmo para a construção dessas fingerprints utiliza um cálculo determinístico para associar um agrupamento de PPPs à posição a ser acessa no vetor da fingerprint, evitando redundâncias e duplicidades. Isso também permite o armazenamento dessas fingerprints para a realização de outros cálculos futuros e facilita as reanálises, caso sejam necessárias.

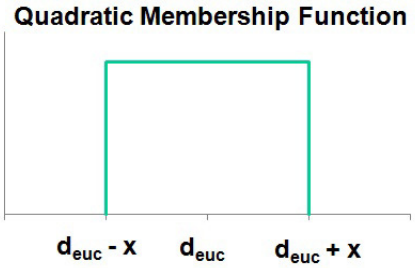


Figura 3.2: Função quadrada de pertencimento ("Quadratic Membership Function"). Uma curva de pertencimento é a curva que define um valor de pertencimento (ou grau de pertencimento) entre 0 e 1 para cada ponto no espaço de entrada. No caso da função quadrada de pertencimento os valores são sempre 0 ou 1, totalmente verdadeiro ou totalmente falso.

Apesar de tudo isso, dependendo do número de configurações possíveis (o que é dependente da possibilidade de agrupamentos de pontos possíveis e das características analisadas) e do número de estruturas do dataset os vetores gerados podem ser muito grandes, demandando um alto custo computacional, tanto em processamento quanto em espaço de armazenamento. Assim, é necessário haver uma racionalização entre o desempenho do programa e a demanda computacional, o que é um fator extremamente relevante nesse tipo de experimento.

Como alternativa para diminuir o espaço necessário para armazenamento dos vetores que são esparsos, as estruturas moleculares são representadas somente pelos bits acesos nas fingerprints, através dos índices da função hash de cada uma das combinações de PPPs e os conjuntos de distâncias. Ao final, cada estrutura é descrita por um conjunto de índices numéricos. Essa metodologia foi escolhida para diminuir o espaço necessário para o armazenamento dos vetores uma vez que, apesar de esparsos e binários, seriam extremamente grandes.

Dessa forma, para agrupamentos de 2 PPPs ou dupletos, sendo os pontos representados por A e B, e a distância entre eles d_{AB}, serão necessários três caracteres, na forma ABd_{AB}. Nos tripletos, serão necessários 6 caracteres para representar os três pontos presentes A, B e C e as três distâncias d_{AB}, d_{AC} e d_{BC}. Já para quadrupletos serão necessários 11 caracteres, sendo 4 para os pontos A, B, C e D, 6 para as distâncias entre eles d_{AB}, d_{AC}, d_{AD}, d_{BC}, d_{BD} e d_{CD} e mais um caractere adicional que represente a quiralidade da combinação, uma vez que a figura geométrica formada pelos quatro vértices (tetraedro) não será plana, e sim espacial (3D).

Ainda, em todas as representações, a nomenclatura A, B, C, D é dada levando em consideração as menores distâncias calculadas em um tuplet. Se houverem empates nos valores de distâncias, serão feitos desempates levando em consideração o tipo de

farmacóforo analisado.

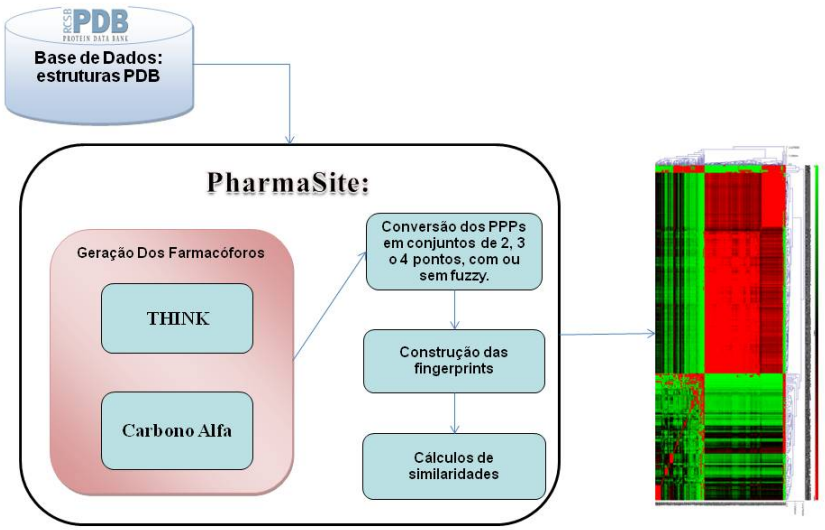


Figura 3.3: Esboço da metodologia de análise de sítios ativos utilizando fingerprints de farmacóforos. A partir das estruturas PDB são gerados os farmamacóforos pelos PharmaSite utilizando o Think ou Carbonos alfa. A partir dos farmacóforos e suas coordenadas são geradas as fingerprints e calculadas as matrizes de similaridade.

3.2 Triagem virtual de pequenas moléculas

Assim como na análise de similaridade de sítios ativos, o primeiro passo para a realização da triagem virtual de ligantes é a escolha da base de dados a ser analisada, que no caso de ligantes, necessita de um tratamento das estruturas moleculares para corrigir possíveis erros. O passo seguinte será a geração dos farmacóforos e suas respectivas fingerprints. Para isso foi desenvolvida no NEQUIM a ferramenta "3D-Pharma". Essa ferramenta, que foi base da tese de doutorado defendida por Bernado Figueredo Magalhães no Programa de Doutorado em Bioinformática da UFMG no ano de 2012, é dedicada unicamente ao tratamento e análise de similaridade de ligantes utilizando tripletos de farmacóforos. Neste trabalho foram sugeridas e aplicadas melhorias baseadas nas experiências com o PharmaSite. Entre as melhorias estão alterações no pré-tratamento das moléculas, métodos diferentes de determinação dos farmacóforos e aplicação de fingerprints com vários agrupamentos de PPPs.

Antes da execução do mapeamento dos farmacóforos nas moléculas dos ligantes, é necessário que as estruturas passem por um pré-tratamento para checagem que envolve

eventuais correções nas estruturas, além de dessalinização. O tratamento propriamente dito inclui o cálculo dos tautômeros mais prováveis e dos possíveis estados de protonação (esta é uma etapa que pode impactar nos resultados, pois dependendo do sistema o pH ideal pode variar acarretando mudanças significativas nas estruturas, entretanto para os cálculos no DUD foi adotado o pH 7), geração da conformação mais estável, cálculo de cargas parciais e geração de múltiplas conformações. Para essas etapas de tratamento foram utilizados os programas QuacPac e Omega (OpenEye) e JChem (Chemaxon). No total foram testados 16 tipos de tratamentos possíveis utilizando combinações das várias opções presentes nesses programas. Iremos analisar todas as moléculas resultantes desses tratamentos a fim de averiguar se os diferentes tratamentos moleculares tem influência significativa nos resultados finais de VS.

Após as etapas de tratamento das estruturas, são calculados os farmacóforos desses ligantes, com a utilização do programa PMAPPER (Chemaxon). Através do PMAPPER um átomo pode ser representado por uma ou mais características farmacofóricas entre seis possíveis e, para a atribuição dessas características, é utilizada como referência a seguinte discriminação:

- Positivamente carregado (P), se um átomo possui uma carga parcial acima de + 0,4;
- Negativamente carregado (N); se um átomo possui uma carga parcial abaixo de - 0,4;
- Doador de hidrogênio (D), se um átomo pode doar hidrogênios para estabelecer uma Ligação de Hidrogênio;
- Aceptor de hidrogênio (A), se um átomo pode receber hidrogênios para estabelecer uma Ligação de Hidrogênio;
- Aromático (R), se um átomo faz parte de um anel aromático;
- Hidrofóbico (H); se um átomo não se encaixa em nenhuma das características anteriores e apresenta carga parcial entre - 0,2 e + 0,2.

Ainda, devido à lacuna existente para a atribuição de características no PMAPPER nos intervalos de cargas de - 0,4 a - 0,2 e de + 0,2 a + 0,4, foram implementadas no 3D-Pharma outras duas possibilidades de atribuição de características Hidrofóbicas, Positivas e Negativas. Em ambas as abordagens foram utilizam as cargas calculadas pelo MolCharge (usando o método AM1BCC) como parâmetro e a classificação como

hidrofóbico somente será aplicada se nenhuma outra característica for atribuída àquele átomo. Na primeira alternativa (que chamamos de ms1) os parâmetros considerados foram iguais aos do PMAPPER entretanto, esperamos que o uso de cargas do MolCharge gere alterações nas análise de similaridade de foram impactar positivamente nos resultados de VS. Na segunda alternativa (chamamdo de ms2) foram considerados como átomos hidrofóbicos aqueles com cargas calculadas pelo MolCharge entre + 0,4 e - 0,4 e os parâmetros para átomos positivos e negativos são iguais aos do PMAPPER.

A seguir, o 3D-Pharma transforma os dados dos arquivos de saída do PMAPPER em um arquivo que contem informações sobre localização espacial dos átomos e suas características farmacofóricas. Cabe salientar que, na prática, os arquivos de farmacóforos do 3DPharma são semelhantes aos arquivos de saída da ferramenta PharmaSite utilizada no tratamento de sítios proteicos. Isso foi feito de forma proposital, com o intuito de facilitar a integração futura de dados de ligantes e alvos moleculares.

A partir desses arquivos serão calculadas as fingerprints e, novamente, podem ser utilizados agrupamentos de 2, 3 e 4 PPPs, além dos intervalos de distância "Normal"e "Rognan"(apesar deste último ser recomendado apenas ao tratamento de alvos moleculares). A metodologia utilizada para os cálculos dessas fingerprints é igual às fingerprints de sítios ativos, descrita anteriormente.

Após a obtenção das fingerprints, os próximos passos são a análise de similaridade e a geração de modelos. Estas etapas podem ser úteis, por exemplo, para a classificação e predição da atividade biológica, diferenciação entre moléculas ativas e inativas em um banco de ligantes ou simplesmente para a montagem de bibliotecas a partir dos modelos produzidos.

3.3 Cálculos de similaridade das Fingerpints de Farmacóforos

As representações vetoriais de entidades biológicas e químicas permitem a realização de cálculos de similaridade com maior rapidez, além do fácil armazenamento de dados. Dessa forma, para calcular a similaridade entre duas moléculas representadas por suas fingerprints, deve-se adotar um coeficiente ou medida que a quantifique. Em geral, as medidas de similaridade buscam medir a quantidade de pontos similares existentes entre os dois vetores. Existem várias métricas disponíveis na literatura, sendo que as mais comuns são os índices de similaridade (veja introdução) ou cálculos de distância entre vetores. Essas medidas possibilitam a construção de matrizes de simi-

laridade (ou de distâncias) que permitem a realização de agrupamentos (clusterização) ou a construção de listas ordenadas pelas medidas de similaridade em relação a uma estrutura de referência, nas quais se espera que as entidades potencialmente ativas presentes na base de dados estejam separadas daquelas consideradas inativas.

Por isso, foram implementados no PharmaSite e no 3DPharma algoritmos para o cálculo de similaridade de fingerprints. Por princípio, esses cálculos são relativamente simples e demandam baixo custo computacional (ordem N^2). Em 2002, Holliday *et al.* (135) realizou um estudo compreensivo entre várias métricas de similaridade de fingerprints 2D. Os resultados encontrados sugeriram que não há um coeficiente que se sobressaia como sendo o mais indicado em todas as situações. Dependendo da aplicação e da implementação das fingerprints, algumas métricas podem ser mais indicadas que outras. Entretanto, o índice de Tanimoto mostrou-se o mais genérico dentre os coeficientes testados, tendo uma performance satisfatória em todos os experimentos e, por isso, foi uma das medidas escolhidas para serem implementadas. O índice de Tanimoto (S_{tan}) pode ser calculado a partir do número de bits acesos em comum entre duas fingerprints (c), cada uma com a e b bits acesos no total:

Fórmula do Coeficiente de Tanimoto:

$$S_{(Tan)} = \frac{a \cup b}{a \cap b}$$

Também foram escolhidas como métricas de similaridade o Coeficiente de Simpson que, de acordo com a Daylight (136), é uma das melhores medidas de similaridade para subestruturas e a distância euclidiana normalizada como medida de dissimilaridade.

Fórmula do Coeficiente de Simpson:

$$S_{(Simp)} = \frac{c}{min(a,b)}$$

Fórmula da Distância Euclidiana Normalizada:

$$d_{E_N}(a,b) = \frac{\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}}{\sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}}$$

3.4 Avaliação da eficiência das medidas de similaridade

A partir dos cálculos similaridade podem ser construídas matrizes de similaridade, quando se utiliza a ferramenta PharmaSite, ou listas ordenadas, quando se utiliza a ferramenta 3DPharma.

Embora as matrizes de similaridade geradas pelo PharmaSite possam ser diretamente utilizadas como arquivo de entrada em procedimentos de clusterização (com a utilização do programa MeV ou outro apropriado para este tipo de abordagem), o foco do programa é a avaliação da eficiência do ranqueamento das estruturas em uma lista onde os valores de similaridade são calculados em relação a uma entidade referência. O mesmo procedimento é adotado no 3DPharma, com a diferença que não é construída uma matriz de similaridade todos contra todos, pois o número de estruturas de pequenos ligantes a ser trabalhada é geralmente muito maior. Nesse programa, apenas as estruturas dos compostos ativos são utilizadas como referência para a construção de listas de similaridade.

À partir das listas de similaridade, ambos os programas podem realizar cálculos de eficiência de ranqueamento. O primeiro passo é ordenar as listas e ranquear (ordenar) as entidades. O ranqueamento é feito pela enumeração das estruturas 1 a n (n é o número total de entidades analisadas). Nesse ranqueamento para evitar que valores empatados influenciem nos resultados finais, caso duas ou mais entidades, entre ativos e inativos, apresentem valores de similaridade iguais, o seu ranking final será dado por uma média ponderada seus rankings individuais. Por exemplo, se no ranqueamento de 5 estruturas, sendo 3 ativos(a, c e d) e dois inativos (b e e), caso b e c tenham o mesmo valor de similaridade, o ranking final será: a = 1, b = 2,5, c = 2,5, d = 4, e = 5, o valor 2,5 é encontrado pela média do ranking 2 e 3, os quais as estruturas b e c deveriam ocupar. Caso não haja nenhum empate no ranking, ou b e c forem ambos ativos ou inativos, o ranking final seria: a = 1, b = 2, c = 3, d = 4, e = 5.

De posse dessas listas ordenadas e ranqueadas, com as respectivas correções, são realizados cálculos para avaliar a qualidade desses ranqueamentos. Para isso, além de medidas de AUCROC, muito comuns em estudos de triagem virtual, uma serie de outros cálculos foram implementados, como BEDROC e Fator de Enriquecimento (EF). Em todas as fórmulas utilizadas, apresentadas abaixo, as variáveis têm os seguintes significados:

- n = número de ativos
- N = número de compostos
- (N-n) = número de inativos ou decoys
- r_i = ranking do ativo i
- χ = fração dos compostos selecionados calculada em relação ao total de compostos para EF ou a fração de compostos inativos para ROCE.

3.4.1 AUC ROC (Area Under the Receiver Operator Characteristics Curve)

As curvas ROC são amplamente utilizadas para avaliar métodos de triagem virtual. Essas curvas são a representação gráfica da taxa de verdadeiros positivos versus a taxa de falsos positivos encontrados à medida que uma base de dados é escaneada. A área formada sobre essas curvas ROC, o AUCROC, representa a probabilidade de um composto ativo ser ranqueado antes de um composto inativo ou decoy, essa medida consiste em um método estatístico muito bem estabelecido e bastante utilizado em VS .

Geralmente, AUCROC é calculado utilizando-se de toda a curva, através da fórmula (também conhecida como teste Mann-Whitney):

$$AUC = \frac{1}{n}\sum_{i=1}^{n}(1 - f_i) \tag{3.1}$$

Onde f_i é a fração de decoys ordenados acima do i-ésimo ativo.

Nesse trabalho, iremos utilizar somente os rankings dos ativos para a realização de todos os cálculos. Assim, iremos utilizar uma fórmula que correlaciona a soma do ranking dos ativos ao AUC (**Eq. 3.2**).

$$AUC = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i - i)}{(N - n)} \tag{3.2}$$

Em um cálculo de AUCROC os valores possíveis estão no intervalo entre 0 a 1, sendo 1 o valor que representa um ranqueamento perfeito, onde todos os ativos estão organizados antes dos inativos. O valor 0.5 representa os valores obtidos quando ordenamos uma lista de modo randômico.

3.4.2 ROCE (Receiver Operator Characteristics Enrichment)

Essa medida representa a fração de compostos de ativos que foram recuperados acima de uma fração de compostos inativos previamente definida. Essa medida é dada por:

$$ROCE(\chi) = \frac{1}{(n * \chi)} \sum_{i=1}^{n} \delta_i$$

$$\delta_i = \begin{cases} 1, & \frac{(r_i - i)}{(N-n)} \leq \chi \\ 0, & \frac{(r_i - i)}{(N-n)} > \chi \end{cases} \tag{3.3}$$

3.4.3 EF (Enrichment Factor)

O fator de enriquecimento pode ser definido como a razão entre a fração de compostos ativos recuperados em uma seleção e a fração de compostos ativos presentes na base de dados. O EF é frequentemente calculado em relação a uma dada porcentagem da base de dados. Por exemplo, EF5% representa o valor obtido quando 5% da base de dados é escaneada.

$$EF(\chi) = \frac{1}{(n * \chi)} \sum_{i=1}^{n} \delta_i$$

$$\delta_i = \begin{cases} 1, & \frac{(r_i)}{(N)} \leq \chi \\ 0, & \frac{(r_i)}{(N)} > \chi \end{cases} \tag{3.4}$$

3.4.4 REF (Fator de enriquecimento relativo)

Devido ao valor de EF ser facilmente influenciado pelo número de ativos na base de dados (137) é recomendado o uso do EF relativo onde o EF é normalizado pelo máximo enriquecimento possível (138) (139) (140).

$$REF(\chi) = \frac{100}{(minimum(n * \chi, n)} \sum_{i=1}^{n} \delta_i$$

$$\delta_i = \begin{cases} 1, & \frac{(r_i)}{(N)} \leq \chi \\ 0, & \frac{(r_i)}{(N)} > \chi \end{cases} \tag{3.5}$$

OBS: arredondar $(n * \chi)$ para o menor número inteiro, já que deve-se comparar com um numero inteiro de compostos selecionados.

3.4.5 Soma dos logaritmos dos ranks (SLR, Sum of logarithms of ranks)

Essa medida é calculada pelo simples somatório do logaritmo dos rankings dos compostos ativos dividido pelo total de moléculas na base de dados. A aplicação logarítmica a esse tipo de avaliação propicia uma ênfase maior às primeiras porções da curva ROC e, dessa forma, enfatiza o reconhecimento precoce de moléculas (140).

$$SLR = \sum_{i=1}^{n} log_{10}(r_i/N) \tag{3.6}$$

Entretanto, assim como EF, o valor do SLR é dependente da quantidade de ativos e decoys na base de dados, o que recomenda o uso de uma normalização. O valor máximo do SLR, onde todos os ativos são encontrados antes dos inativos, pode ser calculado pela seguinte fórmula:

$$SLR_{max} = \sum_{i=1}^{n} log_{10}(i/N) \tag{3.7}$$

Dessa forma, é possível normalizar o SLR. Essa medida normalizada é chamada de NSLR ("Normalised Sum of logarithms of ranks"), que apresentar valores entre 0 e 1, sendo 1 o melhor valor possível.

$$NSLR = \frac{\sum_{i=1}^{n} log_{10}(r_i/N)}{\sum_{i=1}^{n} log_{10}(i/N)} \tag{3.8}$$

3.4.6 AUCpROC

Clark and Clark (139) propuseram a métrica pROC, que realiza uma transformação logarítmica da taxa de falsos positivos, que também faz com que o reconhecimento precoce de ligantes ativos seja mais valorizado. Assim, a área sobre a curva pROC pode ser calculada da seguinte maneira (141).

$$AUCpROC = -\frac{1}{n} \sum_{i=1}^{n} log_{10} \frac{(r_i - i)}{(N - n)} \tag{3.9}$$

3.4.7 BEDROC (Boltzmann-Enhanced Discrimination of ROC)

É uma medida que enfatiza o enriquecimento precoce utilizando o ranking de cada um dos ativos. Ela utiliza uma fator de amplificação que nos permite ajustar os valores obtidos, dando maior ou menor peso ao reconhecimento precoce. Os valores possíveis do BedROC variam no intervalo [0, 1] e pode ser interpretado como a probabilidade de um ativo ser ranqueado antes dos inativos ou decoys.

O RIE (robust initial enhancement) desenvolvido por Sheridan *et al* [14] , que é a base para o calculo do BEDROC, usa um esquema de ponderação exponencial, que da maior peso aos compostos ativos encontrados no inicio das listas (reconhecimento precoce) e os valores obtidos nessa medida dependem da quantidade de moléculas na base de dados.

$$RIE(\alpha) = \frac{N}{n} * \frac{\sum_{i=1}^{n} e^{-\alpha * r_i/N}}{\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}} \tag{3.10}$$

Ainda, é possível calcular o RIE máximo, quando todos os ativos estão situados no início da lista, e o RIE mínimo, quando o ranqueamento não é bem feito e todos os ativos estão no final da lista.

$$RIE_{max} = \frac{N}{n} * \frac{1 - e^{-\alpha * n/N}}{1 - e^{-\alpha}} \tag{3.11}$$

$$RIE_{min} = \frac{N}{n} * \frac{1 - e^{\alpha * n/N}}{1 - e^{\alpha}} \tag{3.12}$$

Com essas duas medidas nós conseguimos calcular o BEDROC, que é dado pela seguinte fórmula:

$$BEDROC(\alpha) = \frac{RIE(\alpha) - RIE_{min}(\alpha)}{RIE_{max}(\alpha) - RIE_{min}(\alpha)} \tag{3.13}$$

ou

$$BEDROC(\alpha) = \frac{\sum_{i=1}^{n} e^{-\alpha * r_i/N}}{\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}} * \frac{sinh(\frac{\alpha}{2})}{cosh(\frac{\alpha}{2}) - cosh(\frac{\alpha}{2} - \frac{\alpha * n}{N})} + \frac{1}{1 - e^{\alpha * (\frac{N-n}{N})}} \tag{3.14}$$

3.4.8 Power Metric

Apesar de existirem muitas métricas para avaliação de um experimento de VS, nós do grupo NEQUIM- UFMG identificamos a necessidade de uma medida que quantificasse a capacidade de predição de um método e não apenas uma maneira de comparação de performances de diferentes métodos. Por isso, buscamos desenvolver uma nova métrica, robusta o suficiente para a avaliação de resultados de VS.

Apesar de métricas bem estabelecidas como as áreas sobre as curvas ROC (AUC) serem capazes de comparar as performances em diferentes métricas, elas são ineficientes em dizer se uma predição apresentada nas soluções dos problemas de VS é satisfatória ou não. As AUC, por exemplo, como apontado por Wald and Bestwick, 2014 (142), podem ser muito inconsistentes para medir a performance de um método. Uma mesma medida de área sobre a curva ROC pode ter várias interpretações em relação ao reconhecimento precoce de ativos.

Já o EF, que é uma das métricas mais usada, tanto em VS como em outras áreas do conhecimento, é muito influenciada pela razão entre ativos e decoys na base de moléculas.

A nova métrica que apresentamos é baseada no teste de poder estatístico, que é a fração de compostos ativos recuperados. Geralmente o poder estatístico é calculado para uma fração fixa de compostos inativos (1% ou 5%, por exemplo). Para incorporar ambas as medidas podemos utilizar a diferença entre as frações de ativos e inativos.

Essa métrica, a diferença entre a fração de ativos e inativos, já foi proposta por vários pesquisadores que a desenvolveram independentemente em várias épocas no passado. O primeiro trabalho a citá-la é de 1884 (143) , mas foi mais de 70 anos depois que ela se tornou conhecida sendo utilizada principalmente em aplicações médicas, e passou a ser chamada de Índice de Youden (144) , devido ao nome do cientista que publicou esse trabalho. Em 2003 essa métrica foi novamente publicada por Powers (145) , chamando-a de "informedness"ou "bookmaker informedness", que pode ser traduzida com a probabilidade de um agente de apostas tomar a decisão correta quando decide aceitar uma aposta.

Apesar de essa métrica ser muito eficiente em avaliar a capacidade de predição de um método o seu uso em estudos de VS não é apropriado, pois ela não é capaz de valorizar a recuperação precoce de compostos ativos. Assim, realizamos uma normalização

dessa métrica, dividindo o Índice de Youden pela soma TPR [1] + FPR [2]. Essa medida variaria de -1 a +1 e, para torná-la associada a uma probabilidade fizemos a medida variar de 0 a +1 por uma simples conversão matemática:

$$\begin{aligned}
\text{Power Metric} &= \frac{TPR - FPR}{TPR + FPR} \\
\text{Power Metric} &= (\frac{TPR - FPR}{TPR + FPR} + 1)/2 \\
\text{Power Metric} &= (\frac{TPR - FPR + TPR + FPR}{TPR + FPR})/2 \\
\text{Power Metric} &= (\frac{2 * TPR}{TPR + FPR})/2
\end{aligned} \tag{3.15}$$

Então, a Métrica Power, com limites entre zero e +1, é dada pela seguinte fórmula:

$$\text{Power Metric} = \frac{TPR}{TPR + FPR} \tag{3.16}$$

3.5 Modelagem de dados

Partindo dos conceitos gerais sobre farmacóforos apresentados anteriormente, que podem ser resumidos na frase de Van Drie, (115) "um farmacóforo é o arranjo espacial de características químicas essencial para a atividade biológica, ou, é um padrão que emerge em um grupo de moléculas bioativas", podemos concluir que dentro das fingerprints de farmacóforos potenciais geradas também há um padrão de informações que podem representar melhor certa classe de estruturas. Assim como em outros métodos de representação de entidades biológicas, as fingerprints de farmacóforos podem trazer dados que representem apenas ruídos, redundâncias ou que simplesmente não trazem ganho informacional, dificultado a correta representação dos alvos moleculares ou ligantes para os quais se pretende avaliar a atividade biológica. Assim, com os arquivos de fingerprints em mãos, ao optar pela realização de modelagem dos dados, pretende-se encontrar uma fingerprint modelo capaz de representar o maior número possível de estruturas de uma classe específica em um dataset.

[1] TRP = "True Positive Rate", Taxa de Verdadeiros Positivos
[2] FPR = "False Positive Rate", Taxa de Falsos Positivos

Com essa finalidade, foi desenvolvida no NEQUIM uma ferramenta capaz de utilizar os métodos bootstrap, validação cruzada e SVM ("Singular Vector Machine") para a realização da modelagem dos dados. Essa ferramenta é a materialização de um novo método desenvolvido no laboratório NEQUIM que propõe uma modelagem de dados classificatórios análoga à metodologia proposta por Tropsha et al (**Figura 3.4**) para aplicação na área de QSAR (Quantitative Structure Activity Relationship ou Relação quantitativa entre estrutura e Atividade) (111) (146). Foi dada a essa ferramenta o nome de ExCVBA (Extensive Cross-Validation e Bootstrap Aplication) e sua metodolotia pode ser resumida na **Figura 3.5**.

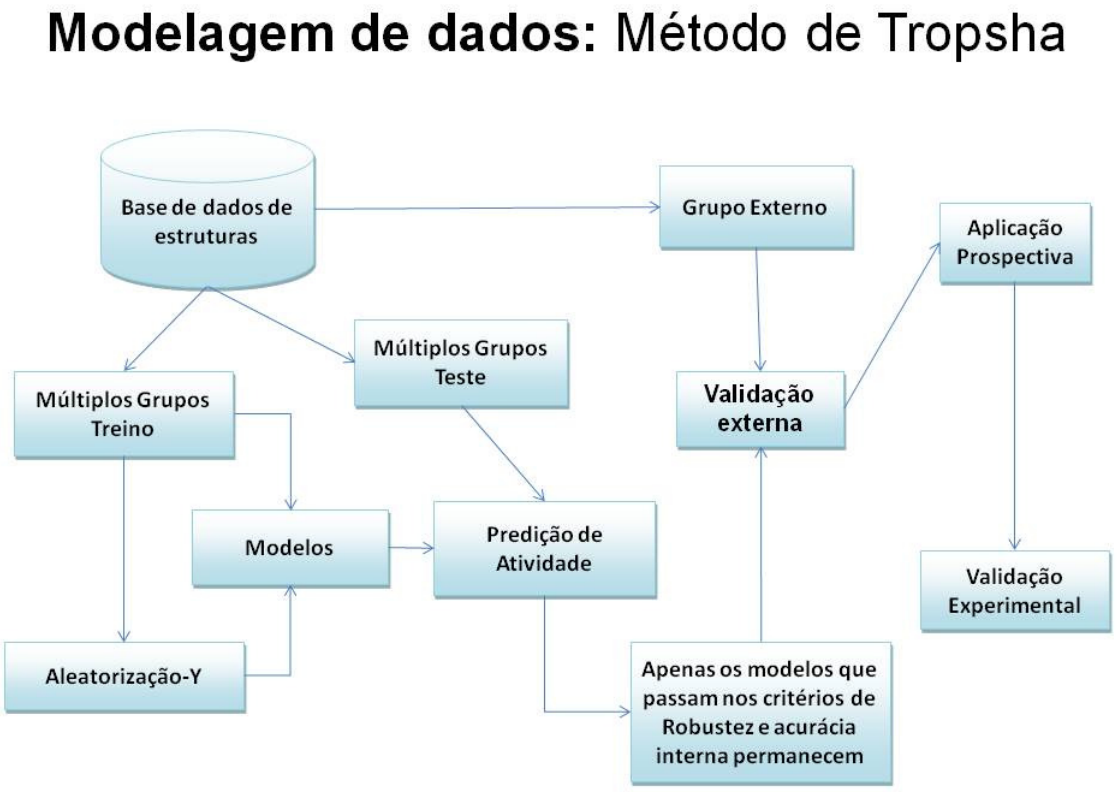


Figura 3.4: Representação do modelo de Tropsha et al (111), fluxograma originalmente proposto à modelagem de dados com utilização de QSAR (Quantitative Structure Activity Relationship) (147) (148). Uma parte do banco de estruturas é retirada, geralmente 20%, para servir como grupo externo, o que resta é divido em grupos treino e teste. Dos grupos treino são gerados modelos reais (que realmente representam os dados contidos no grupo treino) e modelos fruto de aleatorização dos dados. Estes serão confrontados com os grupos teste e os melhores modelos selecionados seguem para a etapa de validação externa. Aqueles modelos que se mostrarem eficientes podem ser utilizados em experimentos com outras estruturas que não estavam contidas no banco de dados.

Para a utilização da ferramenta ExCVBA é necessário ter um data-set onde cada

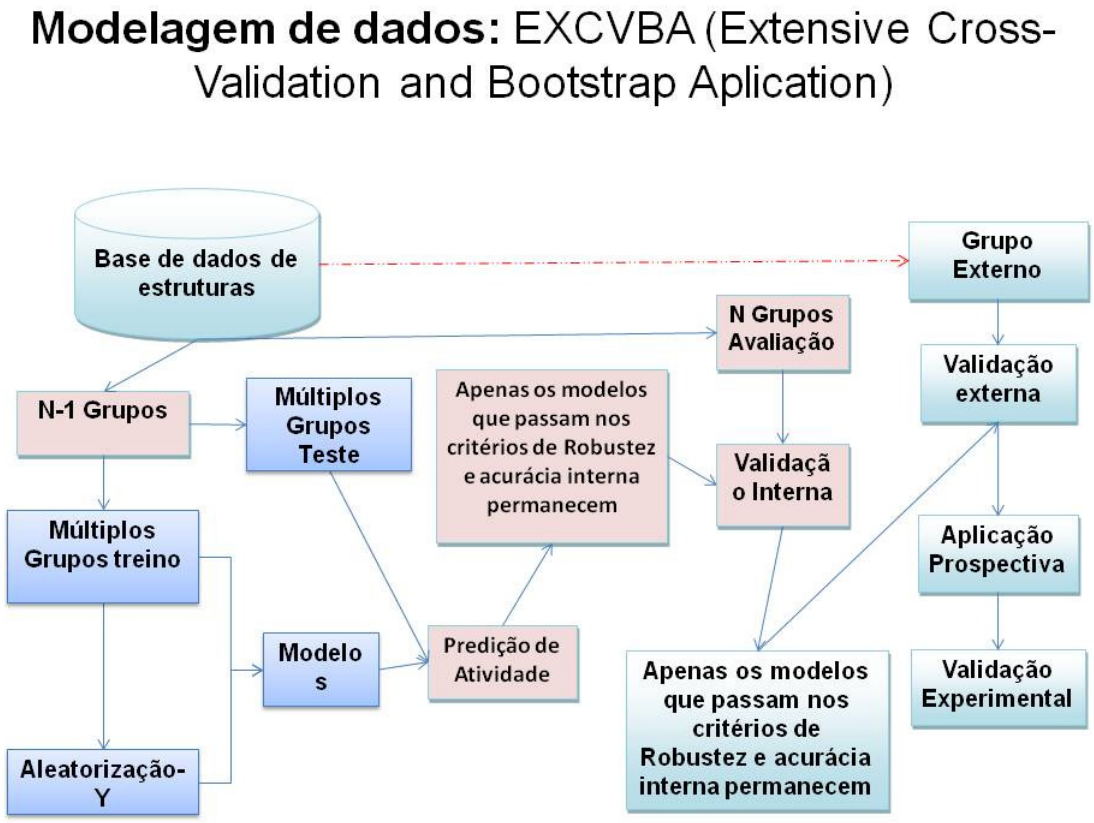


Figura 3.5: Esquema geral do ExCVBA. Caso não haja um grupo de dados para ser utilizado como grupo externo, é retirada uma parte da base de dados para ser utilizada com esse fim. O restante é dividido em n grupos que são submetidos a uma validação cruzada de forma que a cada iteração um grupo é retirado para avaliação e o restante é reagrupado, por uma combinação exaustiva em grupos treino e teste. Para um modelo ser aceito ao final desse protocolo ele deve passar com sucesso por uma validação interna e uma validação externa. Aqueles modelos que se mostrarem eficientes podem ser utilizados em experimentos com outras estruturas que não estavam contidas no banco de dados.

amostra, molécula ou proteína, seja representada por um conjunto de descritores. Estes descritores podem ser resultados de experimentos em bancada, dados físico-químicos ou estruturais, ou quaisquer outros, desde que representem características particulares de cada objeto do dataset. Neste trabalho propomos a utilização das posições das fingerprints geradas com a ferramenta PharmaSite ou 3DPharma como descritores, uma vez que seus farmacóforos representam uma associação de características estruturais e físico-químicas presentes em proteínas e ligantes.

Partindo dessa base de dados, a modelagem será feita da seguinte forma:

- Caso haja um grupo externo e um grupo de análise, o grupo externo é escolhido,

caso contrário é retirado uma porção das estruturas do grupo de análise para servir como grupo externo (por padrão 20%), o restante das estruturas são divididas em n grupos (por padrão n = 10, conhecido como validação cruzada 10-fold).

- Dos n grupos gerados, 1 grupo e retirado por vez para ser usado como grupo de avaliação, e os restantes são divididos entre grupos de treino e teste. Caso n seja igual a 10 por exemplo, com 1 grupo retirado para avaliação sobram nove, destes são calculadas todas as possibilidades de combinação de forma que 3 grupos fiquem no grupo teste e 6 no grupo treino (neste caso são 84 combinações possíveis);

- A partir do grupo treino (formado pela união de seis grupos originais) são gerados um modelo real e um modelo randômico, este é um passo necessário para avaliar se os modelos são realmente representativos e conseguem distinguir ativos de inativos ou, se, devido a dados enviesados, não há diferenças entre estes e modelos gerados de forma aleatória.

- Os modelos são confrontados com os grupos teste e avaliação e os melhores são selecionados;

- Após, retira-se outro grupo (dos n grupos formados) para ser utilizado como grupo de avaliação e o processo é recomeçado, até que o último grupo seja utilizado para avaliação. Ao final teremos vários modelos para serem confrontados com os grupos externo, e o melhor será escolhido.

O "bootstrap" (148) (de utilização opcional), técnica de retirada amostras de uma base de dados (com reposição a cada iteração), pode ser realizado quantas vezes se julgar necessário até que se tenha um bom modelo para o dataset, mas, em geral, é sugerido um número alto de repetições para a realização de análises estatísticas posteriores. Caso não seja utilizado dessa forma, ao menos fornece um grupo externo para reforçar a validação dos modelos gerados internamente.

Capítulo 4

Resultados e Discussão

A capacidade de reter as informações apresentadas por estruturas biológicas, sejam essas oriundas de suas estruturas tridimensionas ou de ensaios de bancada, é um dos fatores que podem explicar o sucesso ou insucesso de uma metodologia aplicada na comparação de estruturas moleculares. Isso significa que essa representação deve conter peculiaridades condizentes com as diferenças e igualdades apresentadas por essas entidades. Assim, espera-se que as estruturas semelhantes gerem representações semelhantes e as desiguais que sejam representadas por padrões informativos diferenciados que permitam a realização de comparações de similaridade de forma eficiente.

Neste trabalho, tanto os sítios ativos das proteínas quanto os ligantes foram representados por fingerprints de farmacóforos. Com isso espera-se que essa metodologia seja robusta o suficiente para a realização de cálculos de similaridade de sítios ativos, na realização de estudos de triagem virtual VS e na geração de modelos preditivos para avaliar probabilidade de atividade biológica de ligantes ou conjuntos de ligante.

Para os estudos de similaridade de sítios ativos utilizaremos bases de dados consideradas *"benchmark"* para esse tipo de estudo, além da base de dados sc-PDB, onde tentamos distinguir proteínas classificadas em diferentes ECs). Para os estudos de triagem virtual utilizaremos a base de dados DUD. E, por último, utilizaremos as bases de dados BBB, AMES e Wisconsin Breat Cancer para a geração de modelos biológicos os quais avaliaremos a capacidade dos famacóforos na geração de modelos preditivos para a previsão de propriedades moleculares (atravessar ou não a barreira hematopoiética) e enfermidades (câncer e câncer de mama).

4.1 Similaridade de Sítios

Para a realização das análises de similaridade de sítios ativos foram utilizados três conjuntos de proteínas contendo estruturas cristalográficas obtidas da base de dados PDB, são elas:

- Aung dataset, uma base de dados com 126 estruturas protéicas que contém sítios ativos co-cristalizados a ligantes estruturalmente semelhantes, mas alguns apresentam o grupo adenina em sua composição e outros não;
- Homogeneous, dataset com 100 estruturas que se ligam a 10 diferentes ligantes com propriedades moleculares similares;
- Base de dados sc-PDB versão 2013 contendo 9.283 entradas sendo, 3.678 proteínas e 5.608 ligantes.

4.1.1 Resultados Dataset Aung

A base de dados Aung é composta de 126 proteínas, sendo que 34 delas apresentam sítios ativos capazes de fazer interação com ligantes derivados da adenina e as outras 92 que não tem essa capacidade de interação, estando cocristalizadas a ligantes similares porém, de outros tipos funcionais **(Tabela 4.1)**.

O estudo onde este conjunto de dados foi proposto (149) é de 2008. Esse trabalho apresentou a ferramenta BSAlign, uma ferramenta que usa da teoria dos grafos através de análises baseadas em isomorfismo de sub-grafos e detecção de cliques máximos para comparar sítios ativos de estruturas protéicas. Como estrutura de referência, para a realização de buscas por outras estruturas semelhantes na base de dados, foi proposta a utilização da proteína de pdb-ID [1] 1ATP, que está co-cristalizada ao ligante ATP. Os resultados apresentados no artigo original mostram uma acurácia de 60%, ou seja, dos quinze primeiros sítios recuperados com a metodologia, nove eram pertencentes ao grupo dos sítios que são capazes de realizar interação com a derivados da adenina.

Em estudo posterior (Weill e Rognan, 2010 (97)), foram apresentados resultados comparativos para os quais foram geradas curvas ROC com a utilização do BSAlign, fornecendo um AUCROC de 0,57. No mesmo trabalho, Nathanaël and Rognan apresentam uma nova metodologia de comparação de sítios ativos, chamada de FuzCav, que usa uma lógica semelhante ao que é feito no PharmaSite, além da apresentação de

[1] pdb-ID = Código utilizado pela base de dados PDB (5) para identificar as proteinas depositadas.

um conjunto de intervalos de distância apropriado para a comparação de sítios ativos. Com o FuzCav foi alcançado um valor de AUCROC de 0,84 nessa base AUNG mas, o melhor resultado foi apresentado pelo programa PocketMatch, com um AUCROC de 0,85 (150).

Por isso, na tentativa de reproduzir o trabalho original, avaliamos a capacidade das ferramentas do PhamaSite em diferenciar entre os dois grupos de sítios presentes nessa base de dados (capazes ou não de realizar interação com derivados da adenina), tendo como estrutura modelo o sítio da proteína 1ATP. E, para isso, tratamos os sítios ativos utilizando as várias combinações possíveis de mapeamento farmacóforos e construção de fingerprints, além de calcularmos a similaridade entre os vetores gerados utilizando diferentes índices de similaridade. Estes estudos iniciais, além de possibilitarem a comparação de resultados com os trabalhos mencionados anteriormente, também servem para direcionar experimentos futuros, pela indicação das melhores opções a serem utilizadas no PharmaSite.

Dessa forma, foram calculados os farmacóforos de todos os sítios das proteínas da base de dados utilizando o THINK e o PharmaSite. Com o THINK foram utilizadas as opções padrão para gerar os farmacóforos. No PharmaSite, como discutido no item 3.1.1 da metodologia, a seleção dos aminoácidos do sítio que serão utilizados para o

Tabela 4.1: Base de dados de 126 proteínas (34 proteínas com capacidade de interação com derivados de adenina e 92 outras) (149).

Functional Type	**Total**	**SCOP Folds**	**PDB IDs**
Adenine-binding proteins	34	18	1a49, 1a82, 1ads, 1atp, 1ayl, 1b4v, 1b8a, 1bx4, 1byq, 1csc, 1csn, 1e2q, 1e8x, 1f9a, 1fmw, 1g5t, 1gn8, 1hck, 1hpl, 1j7k, 1jjv, 1kay, 1kp2, 1kpf, 1mjh, 1mmg, 1nhk, 1nsf, 1phk, 1qmm, 1yag, 1zin, 2src,9ldt
Other proteins	92	21	1a27, 1a52, 1abi, 1acb, 1alq, 1arb, 1azm, 1b56, 1b6o, 1bt5, 1cbs, 1cho, 1com, 1cqq, 1cse, 1csm, 1dbf, 1dcs, 1e6w, 1ecm, 1ela, 1elc, 1equ, 1ere, 1err, 1exm, 1fby, 1fds, 1fem, 1flj, 1fnj, 1fnk, 1ftp, 1g5y, 1ghp, 1gx9, 1hah, 1har, 1hms, 1hne, 1hsg, 1hsh, 1hwr, 1ifc, 1jdO, 1jgl, 1keq, 1kop, 1kqw, 1kzk, 1l2i, 1lhu, 1lib, 1lid, 1lie, 1lvo, 1mbm, 1mdc, 1mml, 1mu2, 1ohO, 1opa, 1opb, 1pek, 1pmp, 1ppf, 1pro, 1q2w, 1qjg, 1qkt, 1rxf, 1sbn, 1sga, 1sgc, 1tgs, 1tyr, 1vrt, 1whs, 1ysc, 1znc, 2alp, 2cbr, 2ifb, 2lbd, 2lpr, 3ert, 3prk, 3sga, 3tec, 4csm, 4sgb, 4tgl
Total	**126**		

Tabela 4.2: Resultados obtidos, em valores de AUC, para a base de dados AUNG com tuplets de 2PPPs. O sítio ativo da proteína 1ATP foi utilizado como referência para o cálculo do índice de Tanimoto e para o ranqueamento dos sítios.

	CA6	**CA7**	**CA8**	**CA10**	**Think**
Normal	0,939	0,908	0,896	0,874	0,752
Rognan	**0,954**	0,917	0,894	0,868	0,736
Fuzzy-0.5normal	0,930	0,915	0,890	0,847	0,745
Fuzzy-0.5rognan	0,937	0,913	0,890	0,857	0,766
Fuzzy-1.0normal	0,933	0,900	0,878	0,852	0,740
Fuzzy-1.0rognan	0,936	0,919	0,901	0,868	0,766
Fuzzy-1.5normal	0,925	0,890	0,869	0,845	0,754
Fuzzy-1.5rognan	0,942	0,913	0,887	0,855	0,758
Fuzzy-2.0normal	0,923	0,893	0,868	0,839	0,751
Fuzzy-2.0rognan	0,942	0,910	0,884	0,851	0,747
Média	0,936	0,908	0,886	0,856	0,752

mapeamento de farmacóforos necessita de um ponto de corte e, por isso, visando avaliar o melhor valor, selecionamos as distâncias de 6, 7, 8 e 10 angstrons para a realização de testes. A seguir, foram calculados os fingerprints com os conjuntos de distâncias "Normal"ou "Rognan"e tuplets de 2, 3 ou 4 pontos, além de poder ser aplicada lógica fuzzy na sua construção. Por fim, para cada uma dessas combinações possíveis foram calculados os valores de similaridade pelas métricas de Tanimoto, Simpson e Distancia Euclidiana Normalizada.

Tabela 4.3: Resultados obtidos, em valores de AUC, para a base de dados AUNG com tuplets de 3PPPs. O sítio ativo da proteína 1ATP foi utilizado como referência para o cálculo do índice de Tanimoto e para o ranqueamento dos sítios.

	CA6	**CA7**	**CA8**	**CA10**	**Think**
Normal	0,917	0,925	0,894	0,868	0,799
Rognan	0,943	0,942	0,920	0,885	0,791
Fuzzy_0.5normal	0,929	0,924	0,897	0,865	0,764
Fuzzy_0.5rognan	0,942	0,939	0,910	0,875	0,775
Fuzzy_1.0normal	0,937	0,923	0,890	0,848	0,746
Fuzzy_1.0rognan	0,949	0,936	0,910	0,875	0,764
Fuzzy_1.5normal	0,933	0,917	0,898	0,849	0,745
Fuzzy_1.5rognan	**0,956**	0,930	0,912	0,873	0,753
Fuzzy_2.0normal	0,925	0,914	0,892	0,842	0,741
Fuzzy_2.0rognan	0,953	0,927	0,912	0,870	0,741
Média	0,938	0,927	0,903	0,865	0,762

A métrica de Tanimoto foi a que apresentou os melhores resultados em todos os experimentos realizados e, por isso, os resultados apresentados nas **Tabelas 4.2, 4.3 e 4.4** são somente da referida métrica. Além disso, a utilização de lógica fuzzy não foi apresentada na **Tabela 4.4** pelo fato de necessitar de um tempo de processamento muito alto, sendo por isso, incompatível com a idéia original desse trabalho de apresentar resultados eficientes e rápidos.

Tabela 4.4: Resultados obtidos, em valores de AUC, para a base de dados AUNG com tuplets de 4PPPs. O sítio ativo da proteína 1ATP foi utilizado como referência para o cálculo do índice de Tanimoto e para o ranqueamento dos sítios.

	CA6	**CA7**	**CA8**	**CA10**	**Think**
Normal	0,878	0,906	0,879	0,860	0,852
Rognan	0,920	**0,936**	0,898	0,887	0,831
Média	0,924	0,929	0,897	0,880	0,841

Os resultados encontrados foram animadores, uma vez que o melhor AUC apresentado na literatura para essa base de dados era de 0,85 (utilizando o programa *PocketMatch* (150), veja a **Tabela 4.5**) e, com o PharmaSite encontramos valores acima de 0,9 em muitas das combinações possíveis para construção de fingerprints. Além disso, pelos resultados apresentados nessas tabelas, também pudemos chegar à indicação de que os melhores valores de corte para seleção de resíduos de aminoácido no sítio de uma proteína estão entre 6 e 7 angstrons (que apresentaram AUCROC médios de 0,93), e que o aumento desse valor pode gerar ruídos que prejudicam os cálculos de similaridade entre sítios de proteínas causando a diminuição dos valores de AUC. Isso pode ser explicado pelo fato de, à medida que aumentamos esse corte, aumenta também a probabilidade de seleção de resíduos sem importância significativa para a explicação dos mecanismos de interação entre a proteína e o ligante.

O método que apresentou os melhores valores de AUCROC foi 2P_rognan [2] e 3P_fuzzy1.5_rognan [3] , onde as áreas sobre a curva ROC chegaram a 0,954 e 0,956 respectivamente (**Figura 4.1**). Também, nessa base de dados, o conjunto de distâncias "Rognan"apresentou melhores resultados de AUC em comparação ao conjunto de distâncias "Normal", o que indica que talvez possa realmente ser mais indicado para a comparação de sítios ativos. Ainda, ficou claro que quando utilizamos as informações dos carbonos alfa do sítio ativo os resultados encontrados foram melhores, mas ainda

[2]Fingerprints construídos com 2PPPs e utilização do intervalo de distâncias "Rognan".

[3]Fingerprints construídos com 3PPPs, utilização do intervalo de distâncias "Rognan"e lógica fuzzy com uma tolerancia de 1,5 Å.

não podemos descartar o uso dos farmacóforos projetados no sítio ativo pelo THINK ou o uso de outros softwares externos para a construção dos farmacóforos.

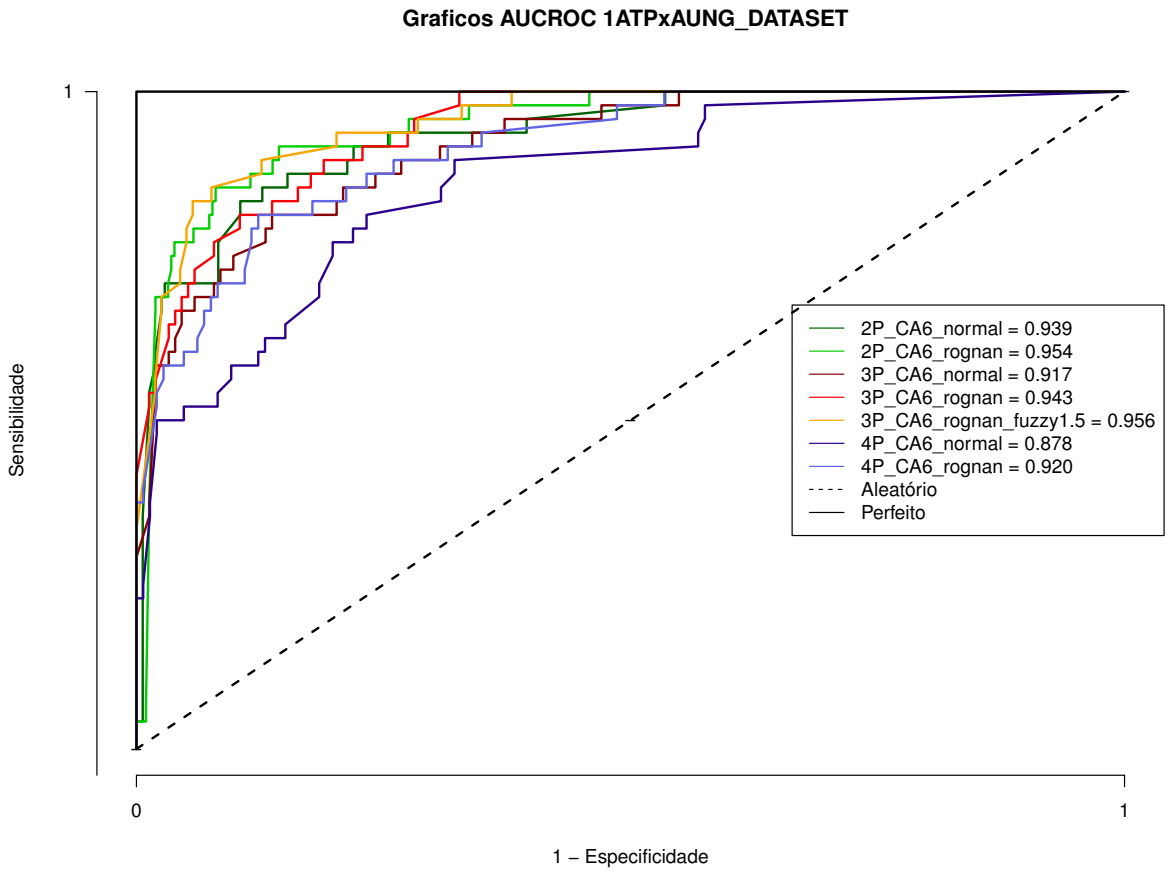


Figura 4.1: Graficos de AUC ROC para o dataset AUNG.

Quanto à utilização de fingerprints fuzzy, em alguns casos houve aumento significativo da recuperação de estruturas ativas mas, e em outros, essa recuperação foi prejudicada. Assim, a sua aplicação deve ser criteriosamente estudada antes da utilização pois, a sua utilização gera um aumento expressivo no tempo necessário de comparação dos sítios. Também, ficou bastante evidente que, devido a esse aumento expressivo do tempo de processamento, a aplicação de fingerprints fuzzy em farmacóforos 4PPPs não é viável computacionalmente. Na prática, o tempo de tratamento de um sítio ativo aumentou de no máximo alguns segundos para, em média, 10 minutos. Isso ocorre pois a aplicação do processo fuzzy necessita da execução de uma série combinatorial de cálculos que gera a inserção 3^6 operações a cada conjunto de 4 PPPs calculado e, isso é agravado pelo fato de a própria resolução do número possível de tuples de pontos de farmacóforo ser também um problema combinatorial.

Tabela 4.5: Resultados de AUC encontrados na literatura onde a base AUNG foi utilizada para a realização de análises de similaridade de sítios.

Método	**AUC médio**
BSAlign (149)	0,57
SiteAlign (96)	0,77
PocketMatch (150)	0,85
FuzCav (97)	0,84
PharmaSite_3PPrognan_CA6Fuzzy1.5	**0,95**

Assim, por ter apresentado melhores resultados que os encontrados na literatura (**Tabela 4.5**) ao utilizar o sítio da proteína 1ATP como referência na busca por estruturas similares na base AUNG, chegamos à indicação de que o PharmaSite pode ser utilizado na análise de similaridade de sítios ativos.

Avaliação do impacto da variação estrutural de sítios nos cálculos de similaridade :

A base AUNG oferece um conjunto de 34 estruturas com sítios ativos capazes de realizar interação com derivados de adenina e, nos estudos anteriores utilizamos apenas um deles como referência para a realização de buscas, por ser esse o modelo apresentado na literatura. Entretanto, poderíamos pensar em utilizar as outras estruturas da base. Dessa forma, visando também avaliar se diferenças individuais entre os sítios, mesmo que de uma mesma classe, influenciam muito nos resultados de similaridade, realizamos outro experimento onde todos os sítios ativos do grupo de proteínas capazes de realizar interação com derivados de adenina foram utilizados, um a um, como estruturas de referência. Os resultados foram avaliados através da média dos valores encontrados, que são mostrados nas **Tabelas 4.6, 4.7 e 4.8** .

Apesar de as diferenças conformacionais terem influenciado negativamente os resultados de AUC, ainda assim, o PharmaSite consegue atingir um nível de recuperação de ativos muito acima das outras metodologias desenvolvidas até agora, chegando a valores de AUCROC de 0,92 e a média de $\sim$ 0,9 para agrupamentos de 2PPPs e 3PPPs. Os melhores valores foram encontrados utilizando 2PPP_rognan e 3PPPFuzzy1.5_rognan.

Novamente, observamos que o aumento do ponto de corte para a seleção de aminoácidos dos sítios prejudica bastante a recuperação de ativos. Também, o aumento no número de pontos de farmacóforo usados para a construção de fingerprints altera

Tabela 4.6: Resultados médios, em valores de AUC, utilizando todos os sítios da base AUNG, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 2 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,918	0,876	0,853	0,838	0,665
Rognan	**0,921**	0,871	0,843	0,842	0,654
Fuzzy_0.5normal	0,905	0,873	0,859	0,826	0,658
Fuzzy_0.5rognan	0,903	0,874	0,848	0,835	0,675
Fuzzy_1.0normal	0,901	0,864	0,848	0,825	0,651
Fuzzy_1.0rognan	0,902	0,879	0,866	0,849	0,670
Fuzzy_1.5normal	0,898	0,862	0,840	0,822	0,657
Fuzzy_1.5rognan	0,903	0,870	0,861	0,843	0,665
Fuzzy_2.0normal	0,893	0,858	0,834	0,816	0,658
Fuzzy_2.0rognan	0,899	0,867	0,861	0,838	0,661
Média	0,904	0,869	0,851	0,833	0,662

Tabela 4.7: Resultados médios, em valores de AUC, utilizando todos os sítios da base AUNG, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 3 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,883	0,897	0,866	0,839	0,722
Rognan	0,904	0,912	0,889	0,862	0,694
Fuzzy_0.5normal	0,896	0,887	0,864	0,840	0,660
Fuzzy_0.5rognan	0,906	0,906	0,881	0,857	0,666
Fuzzy_1.0normal	0,901	0,882	0,858	0,825	0,632
Fuzzy_1.0rognan	0,918	0,902	0,878	0,849	0,650
Fuzzy_1.5normal	0,904	0,882	0,864	0,831	0,630
Fuzzy_1.5rognan	**0,922**	0,891	0,874	0,849	0,639
Fuzzy_2.0normal	0,898	0,877	0,853	0,819	0,626
Fuzzy_2.0rognan	0,921	0,890	0,872	0,843	0,631
Média	0,905	0,893	0,870	0,841	0,655

muito os resultados. Essa variação não está associada, necessariamente, a uma melhora obrigatória nos valores de AUC à medida que são usados mais pontos, exceto para os farmacóforos gerados pelo THINK, onde os melhores valores foram de 0,670, 0,722 e 0,830 com o uso de 2PPPs, 3PPPs e 4PPPs respectivamente.

Tabela 4.8: Resultados médios, em valores de AUC, utilizando todos os sítios da base AUNG, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 4 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,842	0,874	0,847	0,831	0,798
Rognan	0,868	**0,899**	0,869	0,857	0,830
Média	0,855	0,886	0,858	0,844	0,814

4.1.2 Resultados Homogeneous DataSet (HD)

Esta base de dados é composta por estruturas PBD (5) com sítios ativos cujos ligantes co-cristalizados apresentam características moleculares semelhantes, principalmente em tamanho e peso molecular, conforme mostra a **Tabela 4.9**. Ela é constituída por estruturas protéicas cocristalizadas a 10 ligantes diferentes, sendo que para cada ligante existem 10 proteínas na base, em um total de 100 proteínas. Essa base de dados foi proposta em 2010 por Hoffman at al. (151) , e foi idealizada para ser um desafio aos programas de TBVS.

Nessa base de dados o PharmaSite conseguiu, em algumas das combinações de métodos de construção de fingerprint, resultados de AUCROC de 0,728 como mostram as **Tabelas 4.10, 4.11 e 4.12**. Esses resultados, confrontados aos dados da literatura (**Tabela 4.13**), mostram que os valores se encontram dentro do esperado. Pelo fato de ser desafiadora, era esperado que esse dataset realmente fosse difícil de ser tratado. Também, todos os métodos apresentados na **Tabela 4.13** são referentes a um mesmo trabalho (151). Nessa tabela os resultados melhores que os do PharmaSite (sup_CK*) foram encontrados com a utilização de funções kernel que utilizam 4 parâmetos variáveis para otimizar os seus resultados de AUC e, a cada novo experimento, esses parâmetros devem ser recalibrados, demandando bastante tempo. Já o aCSM signature que alcançou um AUC de 0,804, utiliza uma redução de dimensionalidade de dados utilizando SVD ("Singular Value Decomposition") e um método de aprendizagem de máquina supervisionado para alcançar esse valor. Assim, apesar de apresentar resultados de AUC baixos (médias próximas de 0,7), consideramos que o PharmaSite apresentou eficiência comparável às outras metodologias desenvolvidas até aqui, principalmente pelo fato de os resultados obtidos aqui terem sido fruto de uma análise simples de comparação direta dos dados, apesar de poderem ser utilizados também em metodologias de modelagem de dados (**Seção 3.5**) e, talvez, produzir

melhores resultados.

Tabela 4.9: Descrição dos ligantes para o dataset Homogeneous . Adaptada de Hoffman et al. (151)

Ligante	Nº de átomos	Peso molecular	Nº de aceptores de Hidrogênio	Nº de doadores de Hidrogênio	Ligações rotacionáveis
PMP	16	247,17	4	4	4
SUC	23	342,3	11	8	5
LLP	24	361,33	5	6	11
LDA	16	229,4	1	0	11
BOG	20	292,37	6	4	9
PLM	18	255,42	2	0	14
SAM	27	399,45	8	7	7
U5P	21	322,17	8	3	4
GSH	20	306,32	6	6	11
1PE	14	208,25	5	1	11
Average	19,9 ± 4,0	296,4 ± 61,5	5,6 ± 3,0	3,9 ± 2,9	8,7 ± 3,5

Tabela 4.10: Resultados médios, em valores de AUC, utilizando todos os sítios da base Homogeneous, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 2 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,715	0,686	0,669	0,660	0,705
Rognan	0,704	0,673	0,656	0,641	0,700
Fuzzy_0.5normal	**0,718**	0,691	0,682	0,664	0,704
Fuzzy_0.5rognan	0,705	0,678	0,662	0,638	0,706
Fuzzy_1.0normal	0,701	0,684	0,665	0,645	0,705
Fuzzy_1.0rognan	0,710	0,672	0,656	0,627	0,701
Fuzzy_1.5normal	0,712	0,682	0,658	0,640	0,703
Fuzzy_1.5rognan	0,702	0,671	0,657	0,634	0,702
Fuzzy_2.0normal	0,709	0,676	0,654	0,638	0,704
Fuzzy_2.0rognan	0,697	0,667	0,652	0,624	0,701
Média	0,707	0,678	0,661	0,641	0,703

Dentre os resultados encontrados pelo PharmaSite, os melhores valores foram obtidos com ponto de corte para a seleção de resíduos igual a 6 angstrons e uso das

Tabela 4.11: Resultados médios, em valores de AUC, utilizando todos os sítios da base Homogeneous, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 3 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,712	0,706	0,694	0,675	0,706
Rognan	0,714	0,706	0,693	0,675	0,705
Fuzzy_0.5normal	0,723	0,710	0,700	0,674	0,705
Fuzzy_0.5rognan	0,722	0,706	0,689	0,666	0,707
Fuzzy_1.0normal	0,723	0,704	0,687	0,663	0,707
Fuzzy_1.0rognan	0,714	0,697	0,680	0,659	0,702
Fuzzy_1.5normal	**0,728**	0,705	0,683	0,658	0,707
Fuzzy_1.5rognan	0,722	0,695	0,677	0,656	0,703
Fuzzy_2.0normal	0,726	0,700	0,678	0,657	0,707
Fuzzy_2.0rognan	0,716	0,691	0,675	0,653	0,703
Média	0,720	0,702	0,686	0,664	0,705

Tabela 4.12: Resultados médios, em valores de AUC, utilizando todos os sítios da base Homogeneous, um a um, como estruturas de referência para uma busca por estruturas semelhantes. Aqui foram utilizados fingerpritns construídos com tuplets de 4 PPPs e os valores de similaridade foram calculados com o índice de Tanimoto.

	CA6	CA7	CA8	CA10	Think
Normal	0,702	0,701	0,695	0,672	0,703
Rognan	0,697	**0,704**	0,693	0,675	0,700
Média	0,699	0,702	0,694	0,673	0,701

informações dos carbonos alfa para a construção do farmacóforo. Também observamos novamente que à medida que aumentamos os pontos de corte para a seleção de resíduos, os valores de AUC tendem a diminuir, indicando que esse aumento não é favorável à metodologia. Ainda, o uso de 3PPPs apresentou os melhores resultados de AUC em comparação às outras combinações e, nessa base de dados, os valores de AUC encontrados para os conjuntos de distancia "Normal"e "Rognan"não apresentaram uma diferença significativa.

Tabela 4.13: Resultados de AUC médios encontrados na literatura para proteínas do dataset Homogeneous.

Método	AUC médio
aCSM signature	0,804
sup-CKL-Vol	0,766
sup-CKL	0,752
PharmaSite 3P-normal CA6fuzzy1,5	0,728
sup-CK-Vol	0,722
sup-CK	0,710
sup-PI	0,702
MultiBind	0,690
Princ-Axis	0,650
Vol	0,648
Sequence	0,577

4.1.3 scPDB

O scPDB é uma base de dados que reúne proteínas drogáveis do PDB (5). Essas proteínas drogáveis são aquelas que interagem com pequenas moléculas com características similares aos fármacos, sendo que uma da formas mais utilizadas para a identificação dessas pequenas moléculas é através da regra dos 5 de Lipinsky [4] (6). Essa base de dados foi criada cerca de 10 anos atrás, com a finalidade de avaliar o desempenho de abordagens, *in silico* e baseadas em estrutura, para o desenvolvimento de novos fármacos. O sc-PDB disponibiliza os seus arquivos em formato MOL2 separados por ligante, sítio ligante e a(s) correspondente(s) cadeia(s) do sítio da proteína. Na sua última versão, v.2013 que foi publicada em 13 de fevereiro de 2014, há estruturas de quase 4000 proteínas (152) (153).

Os sítios das proteínas no sc-PDB são representados por monômeros (aminoácido, íon, cofator, grupo prostético) que apresentam pelo menos um de seus átomos pesados a uma distância de, pelo menos, 6,5 angstron do ligante cocristalizado e, como essas informações dos sítios ativos já estão organizadas e em arquivos individualizados, essas

[4] Regra dos 5 de Lipinsky: Sugere que moléculas com características farmacóforicas relevantes possuem algumas características moleculares com valores múltipos de 5, como:

- Número de hidrogênios doadores menor que 5;
- Menos de 10 hidrogênios aceptores;
- Peso atômico menor que 500 daltons;
- Coeficiente de partição óleo/água (logP) menor que 5.

podem ser utilizadas diretamente como informação de entrada para o PharmaSite.

Nosso objetivo ao submeter o sc-PDB a testes com a ferramenta PharmaSite é avaliar o desempenho dessa na busca de proteínas com características similares a uma estrutura de referência. Para isso utilizaremos as classes EC ("Enzime Classification") das enzimas do sc-PDB como parâmetro para avaliação do desempenho da nossa ferramenta.

O esquema de classificação EC existe há muitos anos e foi idealizada pela IUBMB (International Union of Biochemistry and Molecular Biology). No topo dessa classificação as enzimas são divididas em 6 categorias que generalizam as reações de catálise de cada uma das classes de enzima, sendo elas:

1 - oxiredutases, realizam oxidação ou redução;

2 - transferases, transfere um grupo químico (como metil, etil, glicosil, etc) de um substrato para um produto;

3 - hidrolases, cliva ligações por reação de hidrólise (carbono-carbono, carbono-nitrogênio, carbono-oxigênio e outros);

4 - liases, eliminam ligações duplas ou anéis, em que reações que não do tipo hidrólise ou oxidação;

5 - isomerases, produz alterações isoméricas (geométricas ou estruturais) em moléculas;

6 - ligases, realizam a ligação de duas estruturas, geralmente é necessária hidrólise de uma ligação pirofosfato, o que produz uma ligação altamente energética.

Ainda, as proteínas desses 6 grupos podem ser divididas em subgrupos, chamados subníveis de EC, onde a cada subnível a classificação de uma proteína se torna mais específica.

Assim, partindo da versão completa do scPDB v.2013, selecionamos todas as proteínas cuja informação sobre a sua classe EC ("Enzime Classification") esteja disponível. No total foram selecionadas 6566 estruturas, as quais foram submetidas a experimentos de comparação de similaridade dos sítios (todos contra todos) utilizando diferentes métodos de construção de fingerprints. Para a avaliação das similaridades, conforme sugeridos pelos resultados anteriores, utilizamos o índice de similaridade Tanimoto.

No trabalho de Weill e Rognan (97), para a verificação de eficácia da ferramenta FuzCav, também foi utilizado a base sc-PDB, em sua versão 2008. Uma parte desse grupo, de proteínas cinases, pertencentes às classes ECs 2.7.10.- [5], 2.7.11.- [6] , 2.7.12.- [7], 2.7.13.- [8] foi tomada como grupo de ativos, e o restante da base como grupo de decoys e, como estruturas de referência, foram usados os sítios ativos das proteínas do proto-oncogene humano de serina/treonina quinase Pim-1 (pdb-ID 1yi4, 3cy3 and 1yhs [9]). Dessa forma, utilizamos esse conjunto de dados como tentativa de avaliar o desempenho da nossa ferramenta no sc-PDB, através da comparação dos nossos resultados com os obtidos por Weill e Rognan (97).

Também foi realizada uma comparação de sequências. Para esse procedimento utilizamos o OpenBabel para converter todas as cadeias dos sítios presentes no sc-PDB em arquivos de formato FASTA e o Clustal Omega para a realização dos alinhamentos de todos os pares possíveis, fornecendo como resultado desses alinhamentos os valores de identidade de sequências.

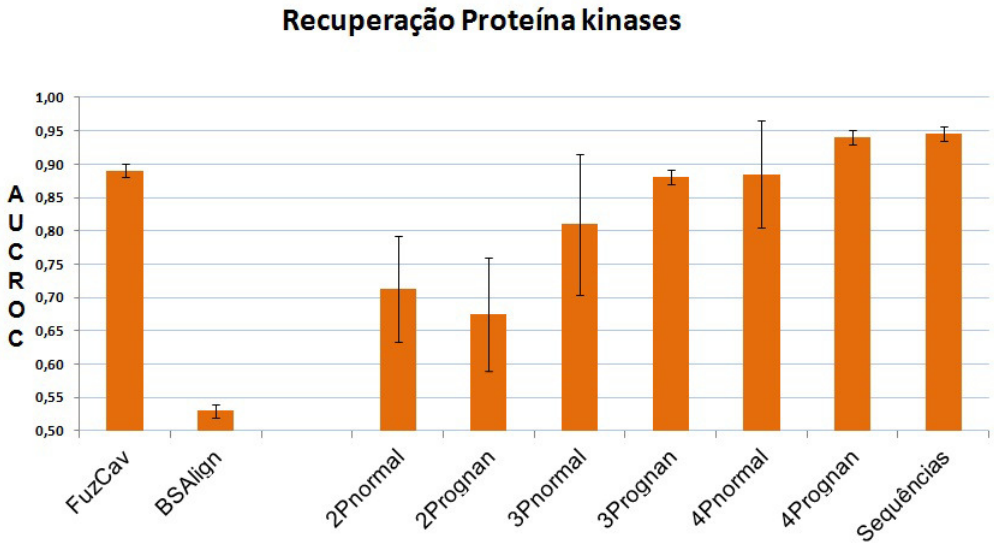


Figura 4.2: Resultados de AUCROC comparados com os dados da literatura para a recuperação de proteína cinases de ECs 2.7.10.-, 2.7.11.-, 2.7.12.-, 2.7.13.- na base sc-PDB.

Os resultados do alinhamento de sequências nos surpreendeu, pois o valor de AUC foi de 0,945 ± 0,011, o que pode indicar uma grande influência dessa similaridade na

[5] Fosfotransferases, subclasse de proteínas tirosino-quinases

[6] Fosfotransferases, subclasse de proteínas serina/treonina quinases

[7] Fosfotransferases, subclasse de proteínas com dupla especificidade, atuando em resíduos de tirosina e serina/treonina

[8] Fosfotransferases, subclasse de proteínas histidino-quinases

[9] Aproteína de pdb-ID 1yhs foi excluída do sc-PDB na versão v2013, por isso não pudemos utilizá-la em nossos cáculos

classificação de ECs. Entretanto, os resultados do PharmaSite também foram muito bons e, com a utilização de 4PPPs e conjunto de distâncias "Rognan", foi encontrado um AUC de 0,941 ± 0,011 (**Figura 4.2**). Utilizando 4PPPs e distância "Normal"(AUC 0,885 ± 0,80) e 3PPPs com distância "Rognan"(AUC 0,881 ± 0,11) os resultados se aproximaram dos do FuzCav (AUC de 0,890 ± 0,010), nossa referência na literatura. Ainda, os melhores resultados encontrados com o PharmaSite foram obtidos com a utilização de 3PPPs e 4PPPs, ficando muito acima dos valores encontrados quando usamos 2PPPs.

Com a obtenção desses valores positivos, resolvemos então analisar toda a base sc-PDB utilizando como critério de separação as classes EC de todas as proteínas presentes. Assim, para todas 6566 estruturas selecionadas do sc-PDB, aquelas com informação de EC definida, foram construídos fingerprints de farmacóforos potenciais com tuplets de 3 ou 4 pontos e esquemas de distância "Normal"ou "Rognan". Todos esses fingerprints foram comparados 2 a 2 utilizando o índice de similaridade de Tanimoto, levando a um total superior a 21 milhões de comparações para cada método de construção de fingerprint usado. Ainda, para efeito de comparação, além da análise utilizando os fingerprints de farmacóforos potenciais, também foi feita novamente a análise utilizando a similaridade das cadeias de aminoácidos pertencentes aos sítios. Cada estrutura de sítio ativo depositada no scPDB, em formato mol2, foi convertida em formato fasta com o uso do OpenBabel e, novamente essas estruturas foram então alinhadas 2 a 2 com o programa Clustal Omega (utilizando os parâmetros padrão).

Para os cálculos de AUCROC, todas as enzimas foram agrupadas de acordo com as suas classes EC considerando, separadamente, apenas os dois primeiros níveis (EC2) ou apenas os 3 primeiros níveis (EC3) de EC. Dentro desses grupos, as enzimas foram comparadas por seus valores de similaridade e, a seguir, submetidas a um processo de fusão de dados (*data fusion*), preservando-se apenas o valor máximo (MAX) ou o valor médio (MED) de similaridade em relação aos outros sítios do grupo. Também, foram utilizados filtros para a similaridade das sequências com a finalidade de permitir uma comparação justa e eliminar possíveis viéses surgidos quando os sítios pertencentes à mesma enzima ou a enzimas muito semelhantes são comparados. Com esse passo, foram eliminadas do processo de análise por fusão de dados as estruturas cujas sequências primárias apresentavam uma similaridade de 90% ou de 95% entre si, de acordo com clusteres disponibilizados no site do PDB. Aqui, foram eliminadas não apenas as estruturas redundantes, nas quais a mesma enzima encontra-se cristalizada com diferentes ligantes, mas também outras proteínas com alta similaridade de sequência.

Após os processos de filtragem foram considerados apenas os grupos EC2 e EC3 que continham pelo menos 5 estruturas para a realização dos cálculos de AUC. Com isso, o número de classes EC analisadas, que eram 45 para EC2 e 106 para EC3 antes das filtragens, caíram, após as filtragens a 95% ou 90%, para 41 e 86, respectivamente, com uma média de 158,2 sítios por EC2 (máximo de 2041 sítios para EC:2.7.x.x) e de 72,7 sítios por EC3 (máximo de 1088 sítios para EC:2.7.11.x). Dentro de cada um dos grupos EC2 ou EC3, todos os sítios das enzimas do sc-PDB foram ordenados e seus rankings foram utilizados para a construção de uma curva ROC e para o cálculo da área sob a curva (AUC), medida essa utilizada como critério de performance da classificação em nossos experimentos.

Na **Figura 4.3** são apresentados os valores de AUC médios obtidos quando utilizada a fusão de dados para o ranqueamento dos sítios enzimáticos do scPDB a partir da similaridade média ou máxima com todos os integrantes de uma determinada classe. Para cada classe enzimática analisada, os seus sítios são considerados ativos e todos os demais sítios do scPDB são considerados inativos, isso é feito com todas as classes, de forma que a cada passo da iteração, uma classe diferente seja tomada como grupo de ativos. Ainda, nessa figura, para cada metodologia de construção de fingerprints utilizada são apresentados os resultados obtidos com os filtros de 90% e 95% de similaridade de sequência, bem como o resultado sem o uso de filtros.

De forma geral, nota-se uma grande diminuição nos valores de AUC com a aplicação dos filtros de similaridade de sequência. O melhor resultado de AUC para EC3 foi obtido com a fusão de dados MAX usando a comparação das sequências dos sítios (alinhamento com o CLustalQ), o valor encontrado aqui foi de 0.93, sem o uso de filtro, e de 0.86 e 0.85 , com os filtros de 95% e 90% respectivamente. Para o PharmaSite, o melhor resultado para EC3, também encontrado com o uso da fusão de dados MAX, foi AUC de 0,90 sem o filtro e 0,78 com o uso dos filtros.

Para os grupos EC2, o melhor resultado de AUC foi obtido com a fusão de dados MAX usando a comparação das sequências dos sítios, esse valor foi de 0,92 sem o uso de filtro e, de 0,86 e 0,84, com os filtros de 95% e 90% respectivamente. Já no PharmaSite, o melhor resultado para EC2, também com a fusão de dados MAX, foi AUC de 0,90 sem o filtro e 0,75 e 0,74 com o uso dos filtros de 90% e 95%.

O fato de a comparação das sequências dos sítios apresentar resultado mais elevado e também ser menos afetado pelos filtros demonstra que há uma maior conservação dos sítios ativos das enzimas do que na sequência completa. É fato conhecido na biologia molecular que as regiões da proteína associadas à sua função biológica, o sítio ativo

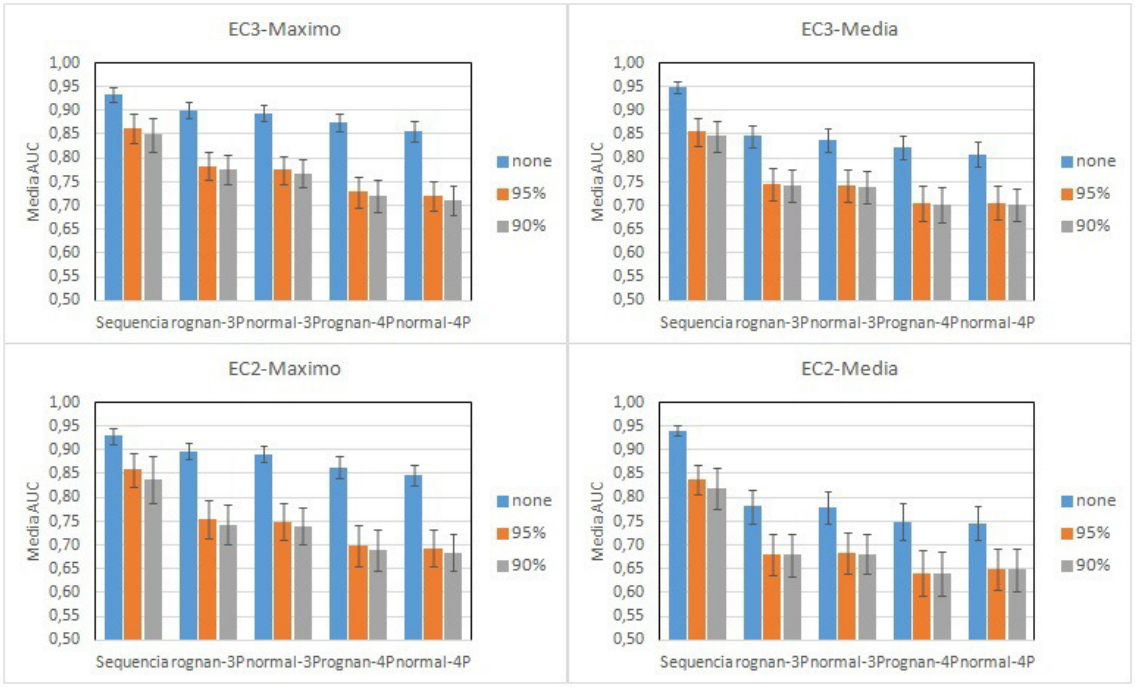


Figura 4.3: Resultados de AUC médios obtidos em análises do sc-PDB para a recuperação de enzimas de mesmo nível de EC. A figura mostra ainda os resultados quando se utilizam métodos de fusão de dados baseados nos valores de similaridade máxima encontrada e nos valores médios.

no caso das enzimas, são mais conservados e menos sujeitas a mutações que as outras regiões, já que mutações nos sítios podem, com frequência, levar à inativação das enzimas.

No entanto, tendo em vista que o PharmaSite foi capaz de produzir resultados próximos ao obtido com as sequências, isso indica que o método utilizado é capaz de reter a maior parte das informações relevantes dos sítio, apesar de ser um método 3D, bem mais complexo que a comparação direta das sequências dos aminoácidos dos sítios, que usa informações em apenas 1 dimensão (1D).

De um modo geral, nota-se que o uso da fusão de dados em sua forma MAX produz os melhores resultados para o PharmaSite, tanto para os grupos EC2 como EC3. No entanto, dada a grande variabilidade de resultados entre os vários grupos, o que resulta em um intervalo de confiança a 95% mais largo, as diferenças são insuficientes para declarar que este método (MAX) é melhor que o outro (MED). O mesmo pode ser afirmado dos esquemas de distâncias adotados na geração dos farmacóforos ("Rognan"e "Normal") e, também, em relação ao número de PPPs adotado (3 ou 4 pontos). No primeiro caso, do esquema de distâncias, as diferenças são praticamente inexistentes,

enquanto que no caso do número de PPPs há uma vantagem clara quando se utilizam farmacóforos de 3 pontos, que é superior ao de 4 pontos, porém ainda dentro do limite do intervalo de confiança a 95%. Por outro lado, o elevado custo computacional para o uso dos farmacóforos de 4 pontos contribui para que os farmacóforos de 3 pontos seja os mais indicados neste caso.

Por tudo isso, o PharmaSite mostrou-se muito robusto apresentando uma boa performance para a comparação dos sítios ativos das enzimas do sc-PDB, mesmo quando foi utilizado o filtro para similaridade de sequência a 90%. Para os 86 grupos EC3 analisados foi obtido um AUC médio de 0,78 e, para os 41 grupos EC2, foi obtido um AUC médio do 0,74. Considerando o grande número de grupos analisados e o grande número de estruturas em cada grupo (158 em média em cada EC2 e 72 em média em cada EC3), podemos afirmar que o PharmaSite é uma ferramenta viável para ser utilizada com a finalidade de auxiliar na atribuição de funções de novas proteínas, para as quais ainda não haja informações suficientes para uma atribuição definitiva.

4.2 Similaridade de Ligantes

A base de dados DUD ("A Directory of Useful Decoys") (154) que é muito utilizada para a validação estudos de VS contém estruturas de uma ampla gama de alvos biológicos farmaceuticamente relevantes tendo sido, por isso, escolhida para avaliar o desempenho do 3DPharma em estudos de LBVS.

Originalmente essa base de dados foi desenvolvida para estudos de TBVS, mas Good and Oprea (155), em 2008, realizaram uma filtragem tornando-a apta à aplicação em estudos de LBVS (**Tabela 4.14**). Essa filtragem teve por finalidade reduzir o número de estruturas semelhantes à molécula utilizada como referencia e, dessa forma, reduzir o viés de enriquecimento causado por estruturas análogas, que é comum em estudos de LBVS e fornecem um enriquecimento acima do que seria correto. Para este trabalho utilizamos uma versão mais nova dessa base de dados adaptada a estudos de VS (DUD LIB VS 1.0) que foi tratada por Jahn em 2009 (154).

São 40 alvos biológicos representados nessa versão do DUD, sendo que para cada alvo existe um número de estruturas reconhecidamente ativas e para cada uma dessas, existem aproximadamente 36 estruturas consideradas potencialmente inativas (decoys) e com características semelhantes aos ativos. No DUD, o grande desafio para as novas ferramentas desenvolvidas é justamente o discernimento entre as estruturas ativas e os decoys.

Tabela 4.14: Base de dados DUD.

Alvo	Código PDB	Versão Filtrada		DUD original	
		ativos	decoys	ativos	decoys
ACE	1o86	46	1796	49	1797
AChE	1eve	99	3859	107	3891
ADA	1ndw	23	927	39	927
ALR2	1ah3	26	986	26	995
AmpC	1xgj	21	786	21	786
AR	2ao6	68	2848	79	2854
CDK2	1ckp	47	2070	72	2074
COMT	1h1d	11	468	11	468
COX-1	1q4g	23	910	25	911
COX-2	1cx2	212	12606	426	13289
DHFR	3dfr	190	8350	410	8366
EGFr	1m17	365	10303	475	15996
ERagonist	112i	63	2568	67	2570
ERantagonist	3ert	18	1058	39	1448
FGFr1	1agw	71	3462	120	4550
FXa	1f0r	64	1633	146	5743
GART	1c2t	8	155	40	879
GPB	1a8i	52	2135	52	2947
GR	1m2z	32	2585	78	2947
HIVPR	1hpx	4	9	62	2038
HIVRT	1rt1	34	1494	43	1519
HMGR	1hw8	25	1423	35	3478
HSP90	1uy6	23	975	37	979
InhA	1p44	57	2707	86	3266
MR	2aa2	13	636	15	634
NA	1a4g	49	1713	49	1874
P38	1kv2	137	6779	454	9140
PARP	1efy	31	1350	35	1351
PDE5	1xp0	26	1697	88	1977
PDGFrb	model	124	5603	170	5980
PNP	1b8o	25	1036	49	1036
PPARg	1fm9	6	40	85	3117
PR	1sr7	22	920	27	1041
RXRa	1mvc	18	575	20	750
SAHH	1a7a	33	1346	33	1346
SRC	2src	98	5679	159	6319
Thrombin	1ba8	23	1148	72	2456
TK	1kim	22	891	22	891
Trypsin	1bju	9	718	49	1664
VEGFr2	1vr2	48	2712	88	2906
TOTAL		2266	98956	3960	127200

4.2.1 Resultados 3DPharma para o DUD

Nos testes iniciais, foram utilizadas as estruturas presentes no DUD sem a realização de tratamentos para os cálculos de conformações, protômeros ou tautômeros. Todas as estruturas dos 40 alvos da versão filtrada do DUD foram submetidas apenas a uma etapa manual de dessalinização, correção de estruturas e cálculo de farmacóforos com o PMAPPER. A seguir essas estruturas foram submetidas ao 3D-Pharma para a construção das fingerprints e a realização dos cálculos de similaridade. Várias combinações de farmacóforos e fingerprints foram testadas no primeiro momento. Assim, utilizamos tuplets com 2, 3 e 4 pontos de farmacóforo com utilização de conjuntos de distâncias "Normal"e "Rognan"e, também, nos cálculos de similaridade foram utilizadas as métricas de Tanimoto e Simpson. Esses testes tem por objetivo avaliar a performance do 3DPharma em estudos de LBVS e direcionar trabalhos futuros. Esses resultados são mostrados na **Figura 4.4**.

Para a recuperação das moléculas do dataset, foram utilizados como modelo os ligantes que estão cocristalizados nos PDBs de referência para cada alvo. Essas estruturas são disponibilizadas no próprio site do DUD no formato Smile ou SDF.

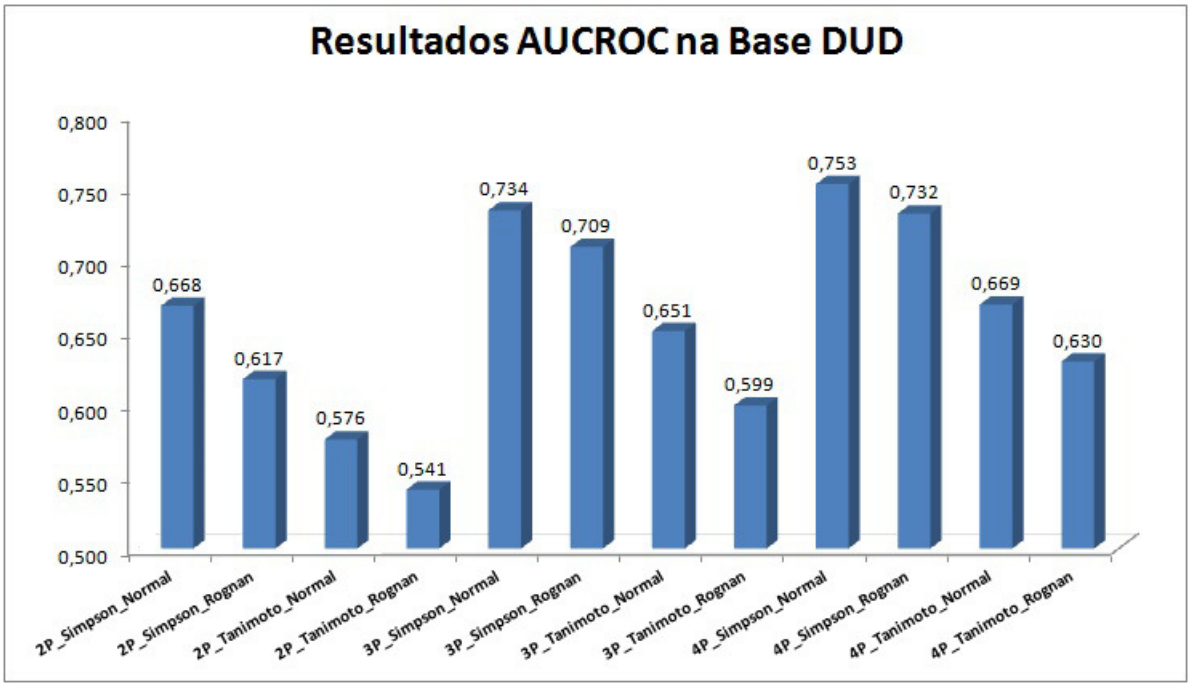


Figura 4.4: Gráfico com os valores de AUCROC obtidos com o 3DPharma no dataset DUD.

Em uma análise geral, os resultados indicam que a utilização da métrica Simpson com 3PPPs e 4PPPs são as melhores abordagens a serem utilizadas para os cálculos de similaridade em ligantes. Ainda, como esperado, os melhores resultados foram encon-

trados com a utilização do conjunto de distâncias "Normal", uma vez que o conjunto de distâncias "Rognan"foi idealizado para a comparação de sítios ativos. Também, esses valores mostraram que a métrica e o conjunto de distâncias utilizados para a realização de cálculos de similaridade interferem bastante nos resultados.

Entre todas as combinações analisadas, aquela em que foram utilizados tuplets de 4 PPPs com conjunto de distâncias "Normal"e a métrica de similaridade Simpson claramente apresentou os melhores resultados. Nessa abordagem os resultados de AUCROC foram melhores em 19 dos 40 alvos quando comparados aos outros métodos de construção de fingerprints (vide material auxiliar, **Tabela B.1**). Por esse critério, de melhores resultados por alvo, as abordagens que obtiveram os segundos melhores resultados foram 4PPP com distância "Rognan"e métrica Simpson, 3PPP com distância "Normal"e métrica Simpson além do 2PPP com distâncias "Rognan"e métrica Tanimoto, todas elas foram melhores em 4 do 40 alvos, mas, dentre esses, o melhor resultado médio foi observado por aquela que utilizava 3PPPs.

Além disso, superficialmente, podemos inferir que a melhor métrica de similaridade a ser utilizada nos cálculos de similaridade de ligante é a métrica Simpson, que em 31 dos 40 alvos apresentou os melhores resultados (vide material auxiliar, **Tabela B.1**).

Os resultados encontrados com o 3DPharma se mostraram muito promissores quando comparados aos resultados apresentados na literatura (**Tabela 4.15**). Dentre as metodologia que analisaram essa base anteriormente, apenas o SHAFTS apresentou melhores resultados médios que os conseguidos aqui. Para todos os outros programas, o 3DPharma com utilização de tuplets de 4 PPPs apresentou resultados melhores e, com a utilização de tupets de 3PPPs somente o ECFP_2 (além do SHAFTS), que utiliza um método 2D de construção de fingerprints, apresentou melhores resultados que o 3DPharma.

O ROCS (159) (Rapid Overlay of Chemical Structures), programa que usa funções Gaussianas centradas nos átomos para representar a estruturas moleculares, em estudo apresentado por Venkatraman et al (2010) (140), apresentou AUCROC médio igual a 0,69 e, fazendo uma comparação dos melhores resultados alvo a alvo para cada um dos 40 alvos do DUD, o 3D-Pharma apresentou melhores resultados em 25 deles. Isso indica que o 3D-Pharma apresenta resultados que fundamentam considerar a sua utilização em estudos de VS.

No ROCS também possível a geração de múltiplas conformações do ativo para a montagem de um modelo, e o mesmo trabalho de Venkatraman et al (2010) mostra que os resultados tendem a ser melhores por esta metodologia. Nele, quando são

Tabela 4.15: Resultados de AUC encontrados na literatura para análises de LBVS sobre o dataset DUD.

Programa	AUC médio	Alvos do DUD em que o 3D-Pharma 4PPPs teve AUC maior
3D-Pharma 4P	0,753	-
SHAFTS (72)	0,767	16
PharmMapper (72), (156)	0,572	31
(2D) ECFP_2 (157)	0,745	19
Shape_ele (157)	0,727	19
ShaEP (158)	0,643	32
ROCS (158)	0,690	26
(2D) CF (158)	0,716	21

utilizadas 10 conformações (AUC = 0,70), 100 conformações (AUC = 0,72) e 1000 conformações (AUC = 0,72) os resultados de AUC tenderam a melhorar. Essa pode ser uma abordagem interessante, pois visa abranger o máximo possível do espaço químico das moléculas analisadas.

Além desses resultados acima, a **Tabela 4.15** também apresenta outros obtidos da literatura. Nessa tabela são apresentados os valores médios de AUC e o número de alvos em que o 3D-Pharma apresentou AUCROC mais alto na recuperação de ativos.

Por fim, considerando o tempo gasto para a realização dos cálculos e a quantidade de espaço necessária para o armazenamento de dados, apesar de os melhores resultados de AUCROC terem sido encontrados com 4PPPs, o custo computacional necessário nessa abordagem foi muito elevado, necessitando de 500ms e 950 ms por comparação, além de cada fingerprint precisar de 4 a 5 MegaBytes de espaço para ser armazenado. Como os métodos de LBVS são desenvolvidos para lidar com bases de dados de ligantes que pode conter milhões de entidades, o tempo gasto e o espaço em disco necessário seriam extremamente elevados. Nesse ponto certamente a maior rapidez foi encontrada quando foram utilizados 2PPPs na construção das fingerprints mas, os resultados de AUCROC médios foram os piores desse experimento. Por isso, a melhor relação entre o custo computacional e a acurácia é encontrada com a utilização de abordagens que usam tuplets de 3PPPs. Esses fingerprints gastam entre 10ms e 20ms para a realização de comparações de similaridade e apresentaram bons resultados de AUC.

4.2.2 Avaliação da correlação entre múltiplas conformações de ligantes e acurácia em VS

Alguns programas utilizam uma análise do espaço químico das moléculas através da geração de multiconformeros para a realização de estudos VS, o ROCS por exemplo possui uma metodologia capaz gerar várias conformações de uma molécula e utilizar esses dados para a realização comparações de similaridade. O 3DPharma, apesar de ainda não ser capaz de gerar esses conformeros, consegue trabalhar com várias conformações de moléculas. Assim os estudos realizados nessa seção tem por finalidade a análise de influência da utilização de múltiplas conformações de ligantes nos resultados de VS realizado com o 3DPharma sobre a base DUD.

Inicialmente, submetemos os ligantes da base DUD a uma série de tratamentos moleculares, no total foram realizadas 16 combinações de tratamento de cargas, conformações, protômeros e tautômeros (**Figura 4.5**) utilizando os programas QuacPac e Omega (OpenEye) e JChem (Chemaxon), conforme descrito no item 3.2. Ainda, os farmacóforos podem ser gerados pelo PMAPPER (programa da ChemAxon) ou utilizando outras duas variações que utilizam uma classificação diferente para grupos hidrofóbico, positivo e negativo.

No total, as moléculas foram submetidas a 48 possibilidades de tratamento molecular sendo que, para todas as moléculas geradas em suas várias conformações, foram mapeados os respectivos farmacóforos e calculados os fingerprints. Após a construção dos fingerprints uma moléculas será representada pelo conjunto dos fingerprints de todas as suas conformações, através da realização de uma união (computacionalmente conhecido pelo termo "*merge*") desses fingerprints. Por essa metodologia, o fingerprint resultante é uma miscelânea de todas as conformações uma molécula que contêm informações do espaço conformacional analisado. Por fim, são realizados os cálculos de similaridade e os cálculos de medida de eficiência na recuperação de ativos.

Devido à grande quantidade de moléculas geradas em alguns desses tratamentos moleculares (em alguns casos o número de entidades, que originalmente eram 98.956, passou da casa dos milhões) e ao número de tratamentos diferentes abordados nesse estudo, foi necessário escolher uma abordagem que maximizasse a relação custo computacional versus eficiência para os cálculos de similaridade de moléculas. Como vimos anteriormente (**Item 4.2.1**), a combinação de métodos que propiciou essa melhor correlação nos experimentos iniciais foi a que utilizava tuplets de 3PPPs com conjunto de distâncias normal sendo, por isso, a metodologia escolhida para a construção das fingerprints de farmacóforo e a posterior realização dos cálculos de similaridade.

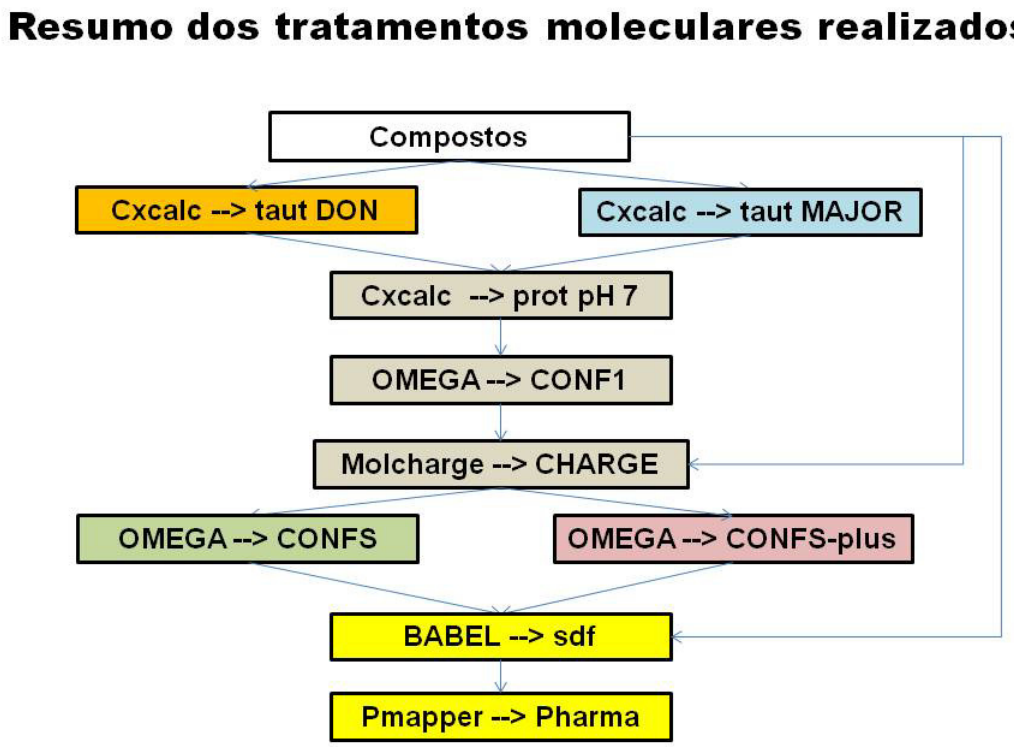


Figura 4.5: Possibilidades de tratamento e cálculos de protômeros e tautômeros utilizando programas da ChemAxon e OpenEye.

Com a utilização do o índice de Simpson o 3DPharma conseguiu melhores resultados de AUCROC em todos os tratamentos moleculares realizados na base DUD. Esses valores (**Figura 4.6**) indicam novamente que a métrica Simpson é mais recomendada para os cálculos de similaridade dessas moléculas, ao contrário do que aconteceu na análise de similaridade de sítios ativos, com o PharmaSite, onde o índice de Tanimoto apresentava melhores resultados.

Ainda, percebemos que o tipo de cargas e quantidade conformações das moléculas, pode impactar os resultados de AUCROC de forma signicativa, uma vez que houve uma uma grande diferença nos valores encontrados. Isso nos leva a inferir que um experimento desse tipo pode ser muito prejudicado por uma má escolha da forma de realização de tratamento das estruturas de uma determinada base.

Nesse quesito (AUCROC), os melhores resultados foram encontrados quando se realizou o tratamentos ms1_6 e ms2_6 com 0,759 de AUCROC médio. Também deve ser dado crédito ao tratamento ms2_30 que alcançou o AUCROC de 0,753. Cabe

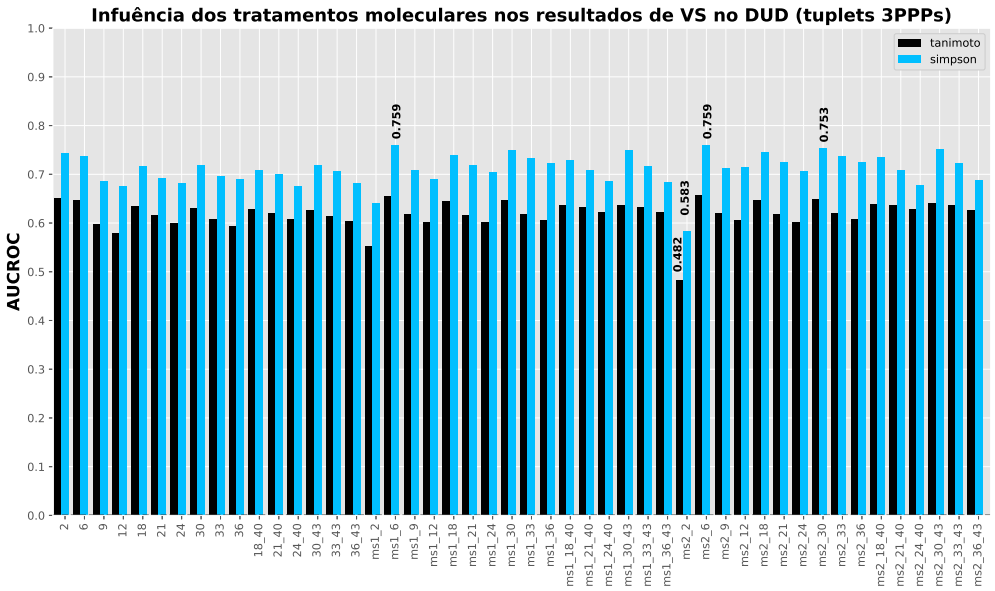


Figura 4.6: Resultados das médias dos valores de AUCROC para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

lembrar que, sem a utilização de qualquer tipo de tratamento, o melhor valor de AUC encontrado era de 0,753 utilizando fingerprints de 4PPPs. Com 3PPPs e conjunto de distancias "Normal"o melhor valor encontrado era de 0,734 e, com a utilização de tratamentos moleculares, subiu para 0,759, que representa um aumento de 3,5 %.

Apesar de terem sido encontrados bons resultados de AUC, outras medidas de avaliação de eficiência em VS são necessárias, principalmente aquelas que priorizam o reconhecimento precoce de ligantes que é uma das análises mais importantes para a aplicação prática, uma vez que, a capacidade de um método recuperar precocemente moléculas ativas é muito importante para aplicações práticas, principalmente para o descobrimento de novos fármacos. Várias métricas para essa finalidade foram implementadas no algoritmo do 3D-Pharma, dentre elas estão o Fator Enriquecimento Relativo (REF), o NSLR e a Power Metric, este último desenvolvido pelo grupo NEQUIM.

Os gráficos de REF a 1% (**Figura 4.7**) mostram que, os melhores resultados foram apresentados com o tratamento ms1_30 e o ms2_30 , com valores de 49,13% e de 49,74% respectivamente. O mesmo aconteceu com os resultados de REF a 5% (**Figura 4.8**) onde os valores obtidos com o tratamento ms1_30 foi de 41,274% e o de ms2_30 foi de 42,105%. Entretanto, podemos perceber que nos gráficos de REF a 5% há uma menor diferença nos resultados em relação aos outros métodos de tratamento molecular,

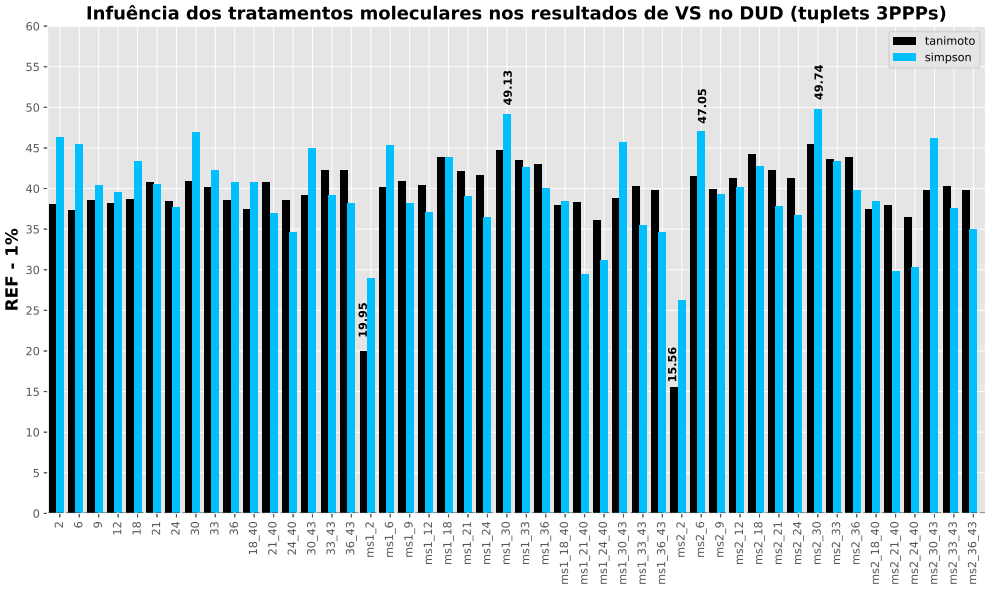


Figura 4.7: Resultados das médias dos valores de REF 1% para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

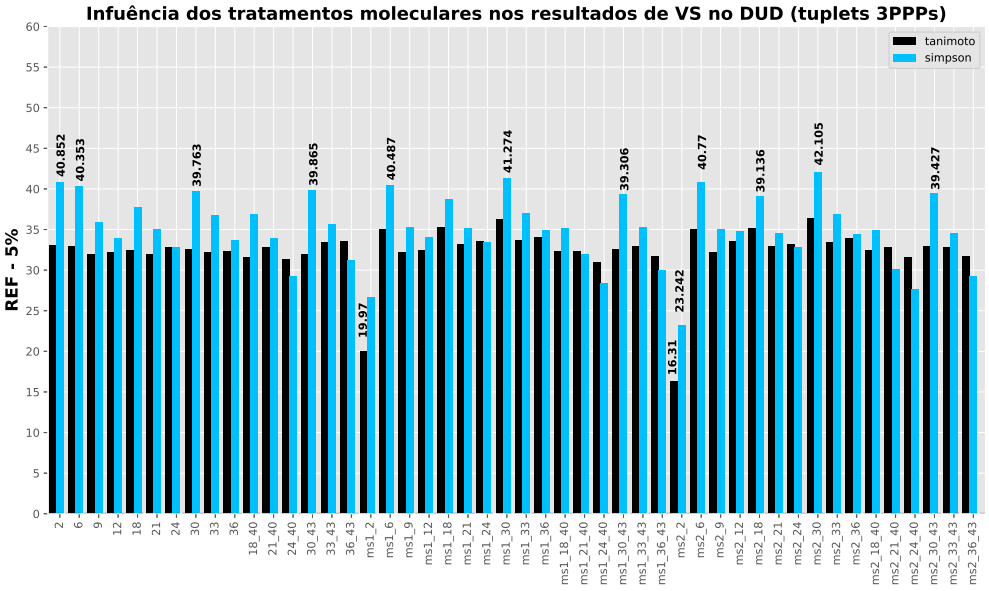


Figura 4.8: Resultados das médias dos valores de REF 5% para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

principalmente aqueles apontados como melhores nos gráficos de AUCROC (ms1_6 e ms2_6).

Esses resultados indicam que que, realmente, o método de AUC não fornece um

bom resultado em relação à recuperação precoce de ligantes. Houve uma alteração significativa na ordem dos tratamentos que fornecem os melhores resultados e, ainda, os valores obtidos com o índice de Tanimoto tiveram, em alguns casos, valores similares e até maiores que os obtidos com o índice de Simpson. Mas, no geral, a métrica Simpson produziu os melhores resultado de Fator de Enriquecimento Relativo, apesar de seus valores apresentarem uma maior variação.

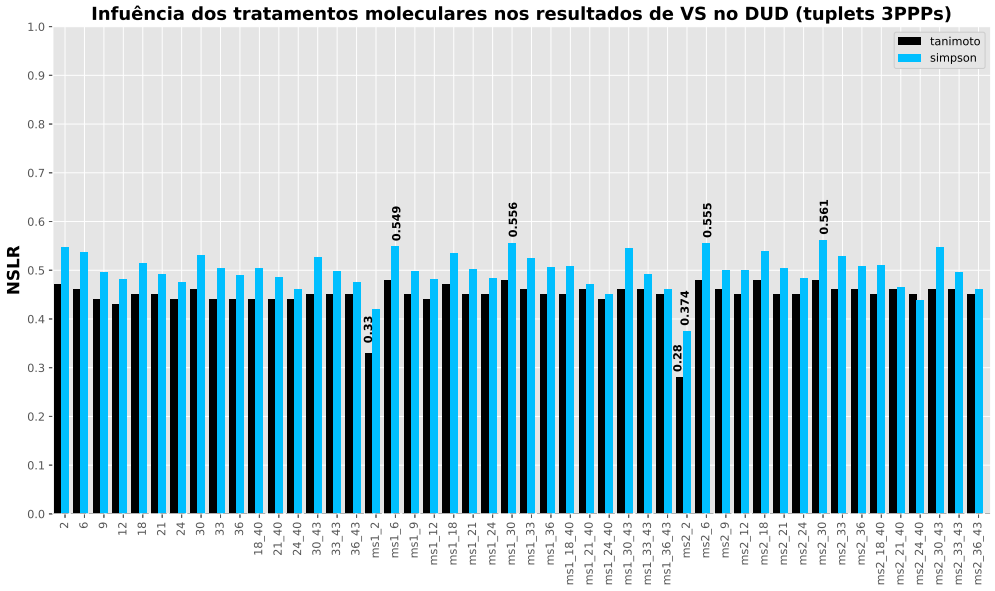


Figura 4.9: Resultados das médias dos valores de NSLR para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

Utilizando a métrica NSLR (**Figura 4.9**), que usa os valores dos logaritmos dos rankings dos compostos ativos recuperados, os melhores valores também foram encontrados usando os métodos de tratamento de moléculas ms1_30 e o ms2_30 , com valores de 0,556 e 0,561 respectivamente. Também percebemos nesses cálculos que o valores obtidos com a métrica de Simpson, que variam bastante, foram os melhores que quase todos aqueles encontrados com a métrica Tanimoto. A exceção ficou por conta do método ms2_24_40, onde a métrica Tanimoto obteve melhores resultados que o índice Simpson.

Por fim, os gráficos nas **Figuras 4.10, 4.11, 4.12 e 4.13** mostram os resultados obtidos com a métrica desenvolvida no NEQUIM, Power Metric, que tem valor máximo de 1.0 para um reconhecimento perfeito de ativos, onde todos são encontrados antes dos ativos. Cada um desses gráficos tem valores de corte diferentes que representam uma porção dos ativos da base. Por essa métrica os métodos apontados anteriormente

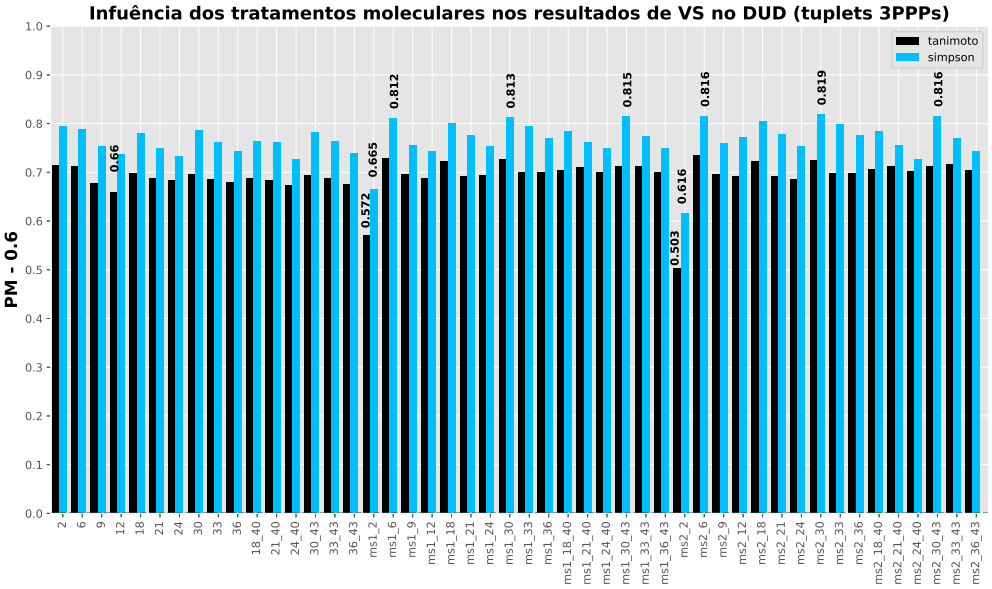


Figura 4.10: Resultados das médias dos valores de PM 0.6 para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

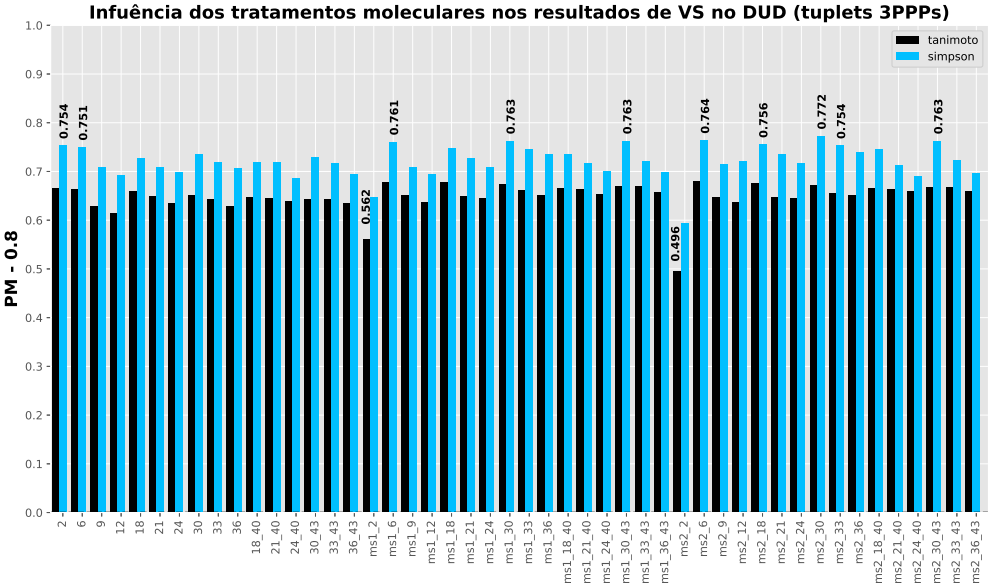


Figura 4.11: Resultados das médias dos valores de PM 0.8 para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

(ms1_6 , ms1_30, ms2_6 e ms2_30) também obtiveram os melhores resultados na avaliação reconhecimento precoce.

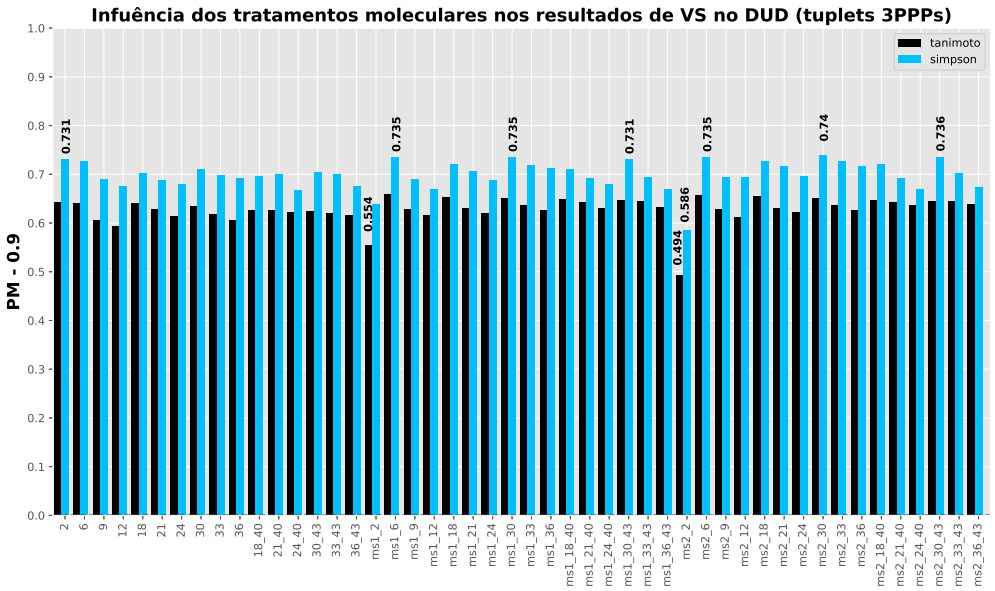


Figura 4.12: Resultados das médias dos valores de PM 0.9 para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

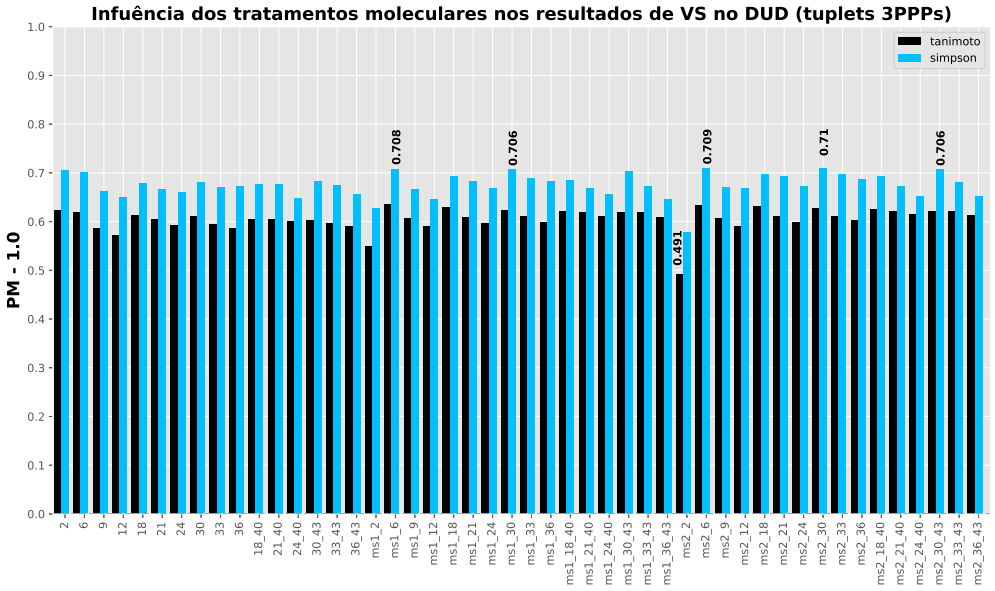


Figura 4.13: Resultados das médias dos valores de PM 1.0 para os 40 alvos do DUD testados em cada um dos 48 tratamentos moleculares aos quais a base foi submetida.

4.3 Modelagem de dados

Utilizando a metodologia de geração de modelos proposta no item 3.5, alguns experimentos foram realizados com bases de dados indicadas como benchmark para esses

estudos de modelagem. Tais bases foram baixadas do site libSVM (160), um programa que utiliza SVM ("Support Vector Machine") para a gerar e avaliar modelos. Essas bases são disponibilizadas em formato de tabelas com a descrição de entidades, suas classes e seus respectivos descritores. Também utilizamos nossa metodologia de geração de modelos em bases de dados biologicamente relevantes onde os fingerprints das moléculas foram gerados com o 3DPharma e utilizadas como descritores para o SVM.

Descritores moleculares utilizados para a representação de alvos e ligantes são, geralmente, obtidos de propriedades calculadas à partir da estrutura molecular ou de informações oriundas de ensaios de bancada. O conjunto desses descritores que melhor caracterizam uma determinada atividade ou efeito biológico compartilhada por um grupo de entidades presentes em uma determinada base de dados pode ser caracterizado como um modelo biológico. Dessa forma, podemos considerar o usos dos fingerprints calculados com o 3DPharma e com o PharmaSite para a construção de modelos SAR ("Strtucture Activity Relationship", Relação Estrutura atividade) pois, esses fingerprints tem a capacidade de representar uma moléculas de forma inequívoca. Também, teoricamente, estruturas parecidas ou de mesmas funções gerarão fingerprints também semelhantes. Dessa forma, os modelos resultantes deverão fornecer as posições desses fingerprints que melhor representem classes de entidades presentes em bases de dados.

4.3.1 Dataset Breast Cancer (Wisconsin Breast Cancer)

Esse conjunto de dados reflete 683 amostras de cânceres de mama benignos e malignos. Nas descrições dos dados é mostrado o tipo de câncer de uma amostra e um conjuntos de 10 descritores (observação de mitoses, células com formato ou tamanhos uniformes, adesão marginal entre outros) que marcam as características observadas em cada amostra.

Primeiramente é necessário ser realizada uma otimização do custo a ser aplicado à função de classificação do SVM (**Figura 4.14**), para otimizar os resultados preditivos e eliminar etapas desnecessárias de cálculo. O custo nesse tipo de função visa aperfeiçoar a separação dos dados através de aumento nas penalizações aplicadas pelo SVM para classificação. O LibSVM, possui quatro funções kernel implementadas em seu algoritmo sendo elas: linear, sigmoide, polinomial e exponencial. Em todas essas funções existe a variável custo e, em algumas (sigmoide, polinomial e exponencial) também podemos otimizar a variável gamma. Os resultados dessa otimização mostram valores de acurácia próximos a 100% em todos os Kernel analisados **Figura 4.14**.

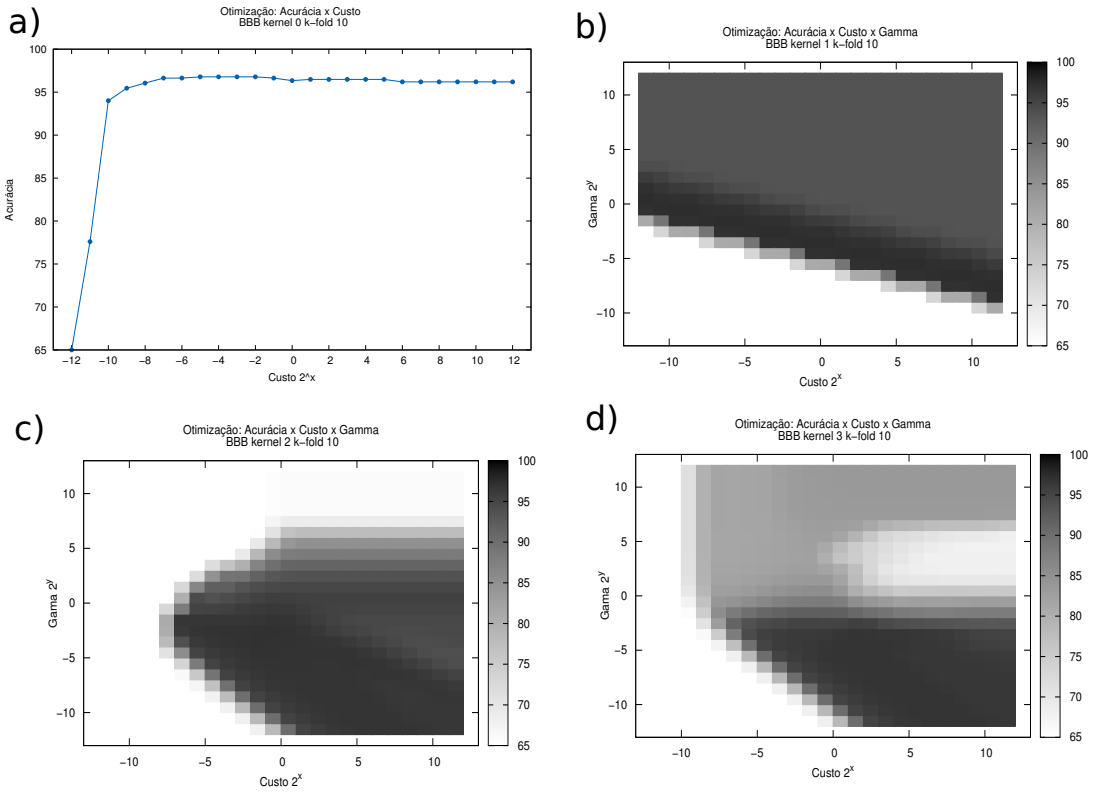


Figura 4.14: Otimização da Relação Custo x Acurácia para o dataset Breast Cancer.

Escolhidos os melhores valores das variável custo e gamma para cada kernel, a base foi submetida a análise pelo ExCVBA. Após, todos os modelos (840 randômicos e 840 gerados a partir dos grupos treino) gerado, foram analisados pelo cálculo do seu poder de predição, ou seja, a capacidade desse modelo em prever corretamente a classe ou atividade de uma determinada entidade, ou seja, prever corretamente se amostra analisada é de um câncer maligno ou benigno. Nos resultados mostrados na **Figura 4.15** foi usado o kernel linear para a avaliação desses modelos. Esses gráficos apresentam a quantidade de modelos encontrados (o valor máximo é de 840) em uma determinada faixa de valores de acurácia .

É possível observar que os resultados de recuperação gerados pelos grupos treino são melhores que os randômicos em todos os casos, entretanto no custo 1024 as curvas se tornam um pouco confusas, as caudas se misturam, e a normalidade da distribuição não se aplica a algumas delas. Para as outras (custos 1 a 8), testes T estatísticos demonstraram que há diferenças significantes entre essas curvas, o que ficou comprovado pelos resultados p-valor menores que 2e-16 em todos os casos, com uma confiança de 95% .

4.3.2 Data-set BBB

Essa base de dados é composta por moléculas que tem ou não a capacidade de atravessar a barreira hemato-encefálica (BHE ou BBB), uma camada de células epiteliais nas paredes dos vasos sanguíneos justapostas ("tight junctions"), que restringe a passagem de solutos. Todas as moléculas dessa base de dados foram convertidas em fingerprints de farmacóforos com a utilização do 3DPharma e as posições das fingerprints foram usadas como conjunto de descritores a ser submetido à metodologia de modelagem descrita no item 3.5.

A otimização da correlação custo por acurácia, realizada inicialmente mostrou que os descritores podem levar à geração de bons modelos para as moléculas desse dataset, pois os resultados de acurácia foram próximos de 100% (conforme gráfico 4.16) em todos os custos analisados. Conforme pode ser observado, todas as metodologias fornecem resultados de acurácia parecidos, assim escolhemos o kernel linear que tem um numero menor de variáveis para serem controladas e realizamos testes com vários custos utilizando o ExCVBA para a modelagem dos dados. Na **Figura 4.17** são mostrados os resultados de acurácia obtidos com todos os modelos (840 randômicos e 840 gerados a partir dos grupos treino).

Na figura apresentada é notória a separação entre as curvas geradas por modelos dos grupos treino ("normais") e aleatórios. Para confirmar essa observação realizamos um teste t e o valor do p-valor foi menor que 2e-16, o que confirma a significância estatística em dizer os modelos gerados dos grupos treino são melhores que os gerados aleatoriamente.

4.3.3 Dataset Ames

A base AMES é amplamente empregada em ensaios biológicos para avaliar o potencial mutagênico de compostos químicos. Nesse estudo, os modelos SVM (Singular Vector Machine) foram construídos com a base de dados AMES Bursi (161) que contem 4284 compostos, sendo 2383 mutagênicos e outros 1901 não mutagênicos. Nossa abordagem foi empregada para a construção de modelos SAR que poderão ser usados para predizer a potencial mutagenicidade de pequenas moléculas orgânicas.

Todas as moléculas do AMES Bursi foram submetidas a tratamento molecular onde foram gerados múltipos tautomeros dominates ou o tautômero principal no pH = 7 com o Calculator Plugin da ChemAxon e como o software OMEGA da Openeye, onde foram geradas as múltiplas conformações. Os farmacóforos foram calculados com

o PMAPPER (da ChemAxon) e foram então submetidos ao 3DPharma para a geração dos fingerprints, onde cada molécula é então representada por sequências numéricas que refletem todos os tuplets de 3PPPs encontrados. Ainda, a característica hidrofóbica foi atribuída aos átomos por duas formas diferentes que são baseadas na carga do átomo, em um primeiro método consideram-se hidrofóbicos os átomos com cargas entre -0,2 e + 0,2, chamado de "ms1", e o método onde os hidrofóbicos estão entre -0,4 e + 0,4 chamamos de "ms2".

Assim, realizamos esse experimento para avaliar a viabilidade de construção de modelos SAR (gerados com o software LibSVM) utilizando fingerprints de farmacóforos como descritores e, também avaliar o impacto de utilização de múltiplos conformeros e múltiplas espécies na construção desses modelos.

Primeiramente, com todas as moléculas do dataset, foram realizados experimentos para avaliar quais os kernel (função para a realização da separação dos conjuntos pelo SVM) e custo (**Figura 4.18**), seriam melhores para a classificação das entidades presente na base AMES e, baseado nos resultados obtidos, apenas o kernel radial não obteve valores de acurácia bons para o intervalo de custos testado de 2^{-12} a 2^{12}. A seguir, foram gerados modelos produzidos com os grupos de treino (48% dos compostos) e avaliados contra grupos externos (20% dos compostos) selecionado aleatoriamente, que obtiveram uma acurácia de 74,4 $\pm$ 2.4 % a um nível de confiança de 99,7% (**Figura 4.19**).

A validação final foi feita com a utilização de uma nova base de dados contendo 2278 compostos (1281 mutagenicos e 997 não mutagênicos), LMMD (162), que será utilizada como conjunto externo. Foram calculados todos os fingerprints dessas moléculas e usados como descritores dessas. Para avaliar o poder de predição foi gerado um único modelo usando a base de Bursi completa (4284 moléculas) e o resultados mostraram uma acurácia de 81% em alguns métodos, próxima do limiar de reprodutibilidade do AMES (80-85%) (163) **Tabela 4.16**, resultado máximo possível nessa base.

Tabela 4.16: Resultados de predição encontrados quando se utiliza a base LMMD como grupo externo e o modelo gerado com o AMES.

Kernel/Custo	ms1_don	ms1_major	ms2_don	ms2_major
linear / 2^{-6}	79,7%	79,3%	**80,7%**	80,4%
polinomial / 2^{12}	60,2%	60,2%	60,2%	60,1%
radial / 2^{8}	72,9%	72,7%	73,1%	73,1%
sigmóide / 2^{8}	71,4%	71,2%	71,3%	71,4%

Assim, com esse estudo de caso, novamente demonstramos a possibilidade de que essa abordagem seja usada para construir modelos válidos para a predição de atividade biológica e até propriedades ADMET para pequenas moléculas orgânicas.

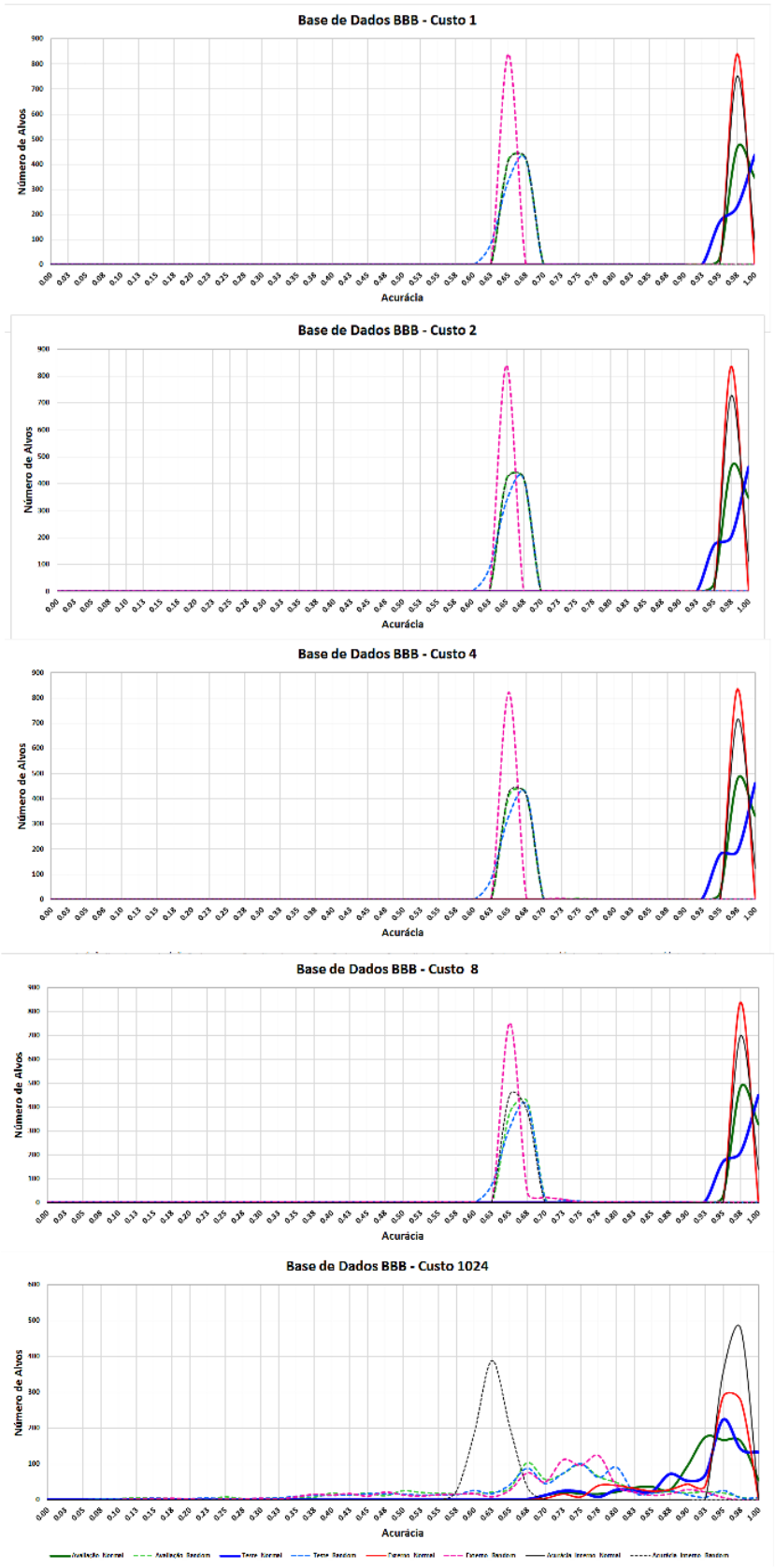


Figura 4.15: Perfis de distribuição de resultados de acurácia, para diferentes custos (1, 2, 4, 8 e 1024) na função do SVM, resultantes da metodologia de modelagem de dados do dataset Breast Cancer.

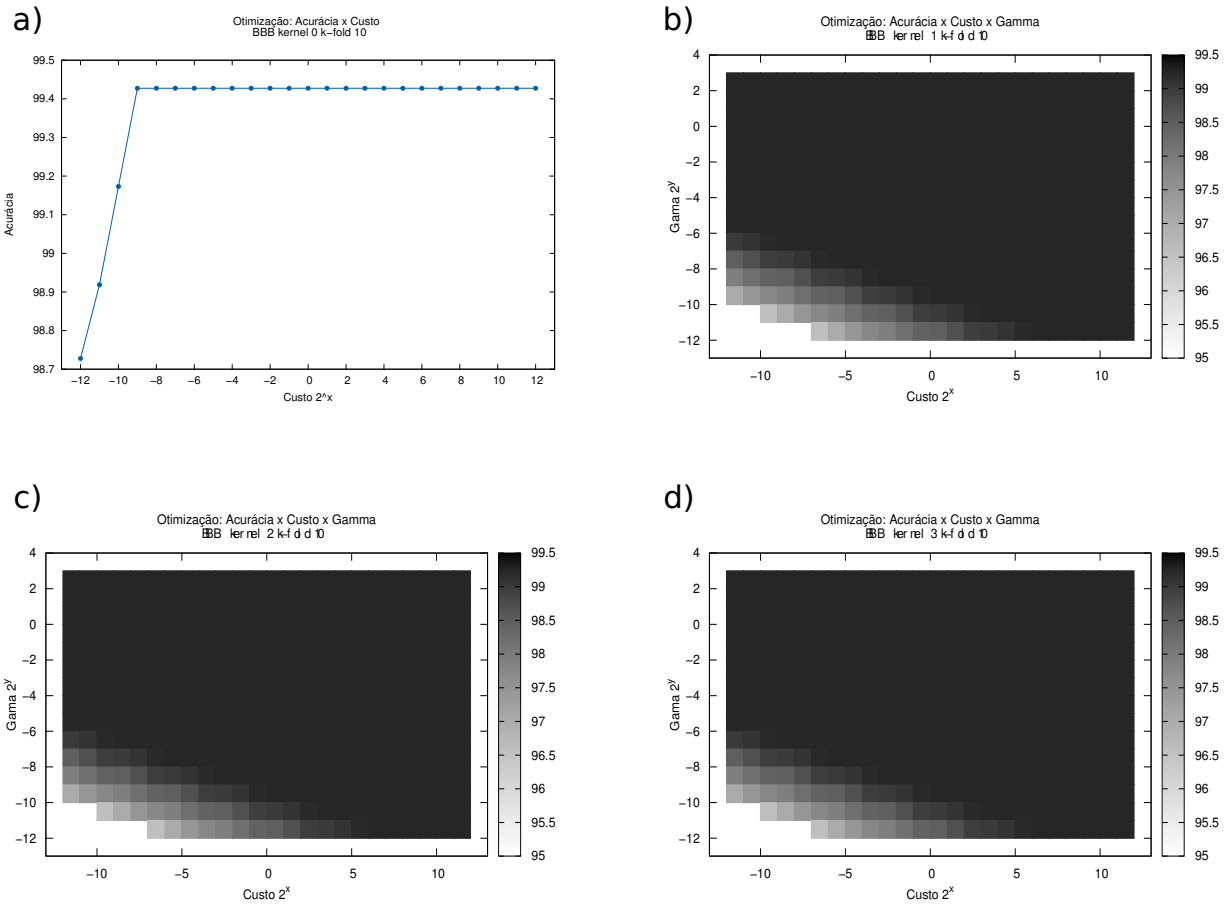


Figura 4.16: Resultados dos testes de otimização de custo versus acurácia para os Kernel linear, polinomial, exponencial e sigmóide.

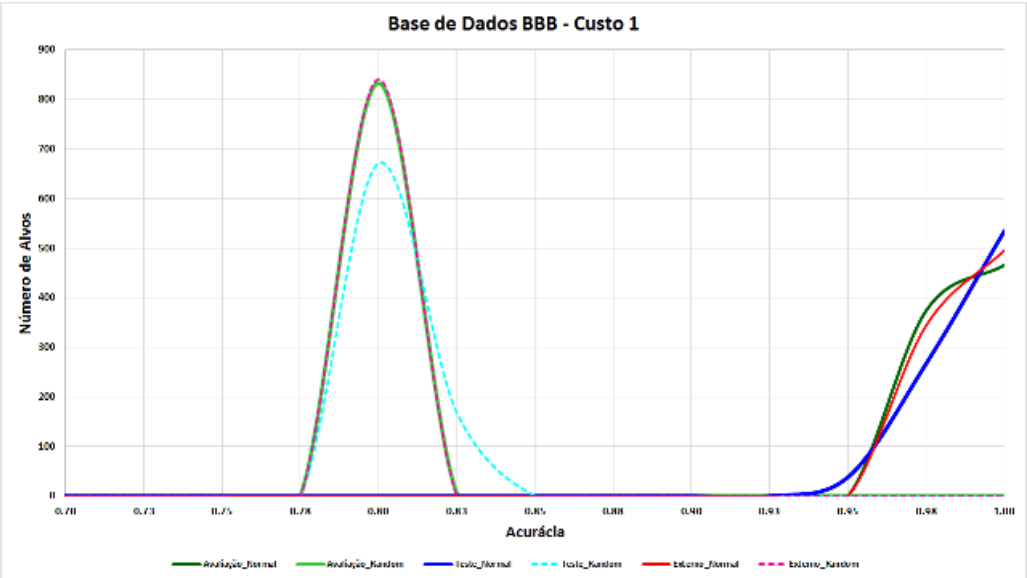


Figura 4.17: Perfil de distribuição de resultados de acurácia, com custo 1, resultante da metodologia de modelagem de dados aplicada ao dataset BBB.

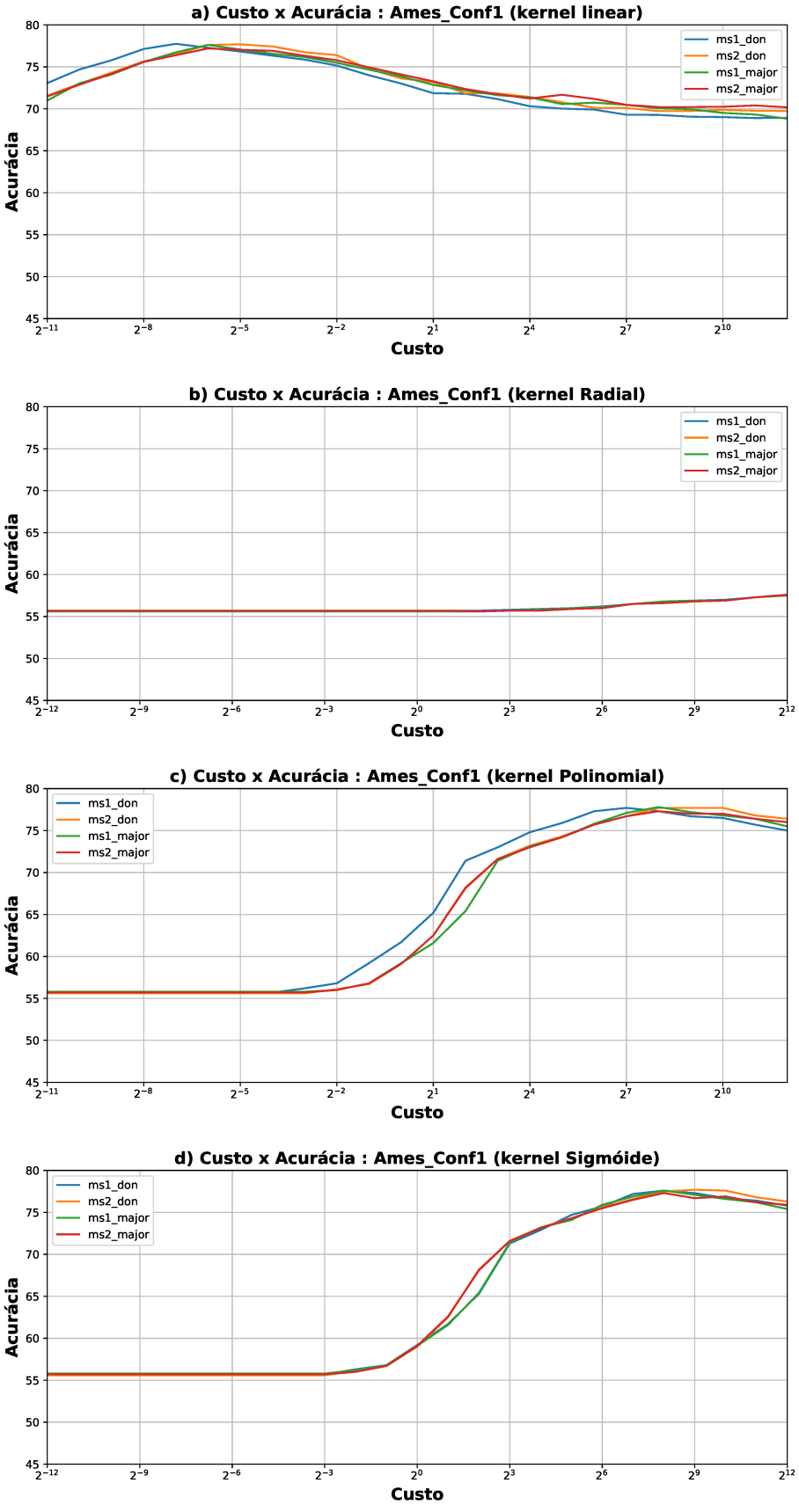


Figura 4.18: Resultados dos testes de otimização de custo versus acurácia para os Kernel linear (a), radial (b), polinomial (c) e sigmóide(d).

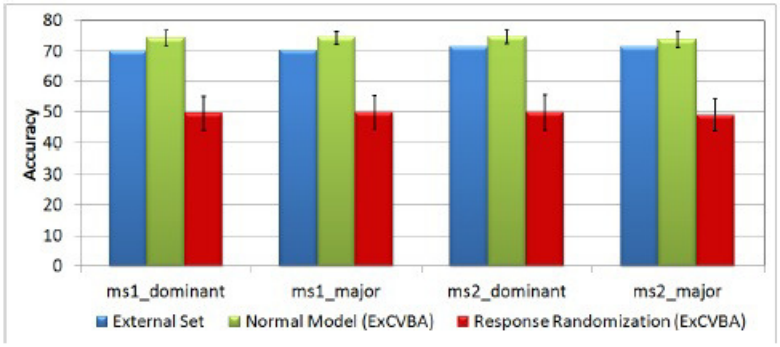


Figura 4.19: Resultados predição dos modelos gerados com o ExCUBA sobre os grupos de avaliação de externo, os resultados são calculados com um nível de 99,7%.

Capítulo 5

Conclusão

Neste trabalho foi apresentada uma metodologia que utiliza fingerprints e farmacóforos para a representação de estruturas biológicas e químicas. Essa metodologia pode ser utilizada em estudos de similaridade de sítios ativos, similaridade entre ligantes e, ainda, para a geração de modelos. Também, para cada uma dessas aplicações, foram produzidas ferramentas computacionais que se mostraram competitivas e, em alguns casos, até melhores que outras ferramentas disponíveis atualmente.

Nas análise de sítios ativos obtivemos dados que indicam ser viável à realização de análises de similaridade através do uso fingerprints de farmacóforos e o uso de informações dos carbonos alfa dos resíduos de aminoácido de sítios ativos para o mapeamento desses farmacóforos. Também, observamos que a seleção de aminoácidos que caracterizam o sítio ativo não devem estar muito distantes dos ligantes cocristalizados, uma vez que esses aminoácidos podem não adicionar informação relevante para explicar a interação proteina-ligante e prejudicarem a obtenção de bons resultados. Pelo que observamos, o melhor ponto de corte para a seleção de aminoácidos dos sítios de ligação está entre 6 e 7 angstrons. Ainda, observamos que a utilização de tuplets com 3PPPs propiciaram os melhores resultados nesse tipo de análise e que o uso de lógica fuzzy aplicada à construção desses fingerprints pode produzir bons resultados. Por último, dentre as métricas de similaridade analisadas, a que apresentou os melhores resultados para a comparação de sítios protéicos foi o índice de Tanimoto.

Já para a análise de similaridade entre ligantes, os métodos que usam 3PPPs possibilitaram uma melhor correlação entre o custo computacional e a eficiência e, por isso, deve ser a metodologia de escolha para lidar com grandes bases de dados. Entretanto, os métodos que apresentaram melhores resultados foram os que utilizavam 4PPPs para a construção de fingerprints mas, esses métodos também tem a maior

demanda computacional. Ainda, o índice de similaridade Simpson propicia melhores resultados em comparação ao índice de similaridade Tanimoto para lidar com ligantes. Também, através da geração de múltiplos tautômeros e conformeros para os ligantes da base DUD, concluímos que o tratamento molecular influencia bastante nos resultados de VS, deve ser escolhido com cuidado pelo pesquisador e pode promover um aumento expressivo dos resultados em uma análise de VS.

Por fim, a nossa metodologia para a geração de modelos se mostrou eficiente e foi testada em algumas bases de dados, apresentando em todos eles bons resultados. Essa ferramenta constitui uma peça essencial para a continuação futura deste trabalho uma vez que ele converge para um pacote de ferramentas capaz de realizar a comparação entre alvos biológicos e ligantes de forma integrada, ou seja, dado um sítio qualquer, ao realizar uma pesquisa em um banco de ligantes pretende-se retornar aqueles com maior afinidade e vice-versa.

Referências Bibliográficas

1 WATSON, J. D.; WITKOWSKI, J.; MYERS, R. *DNA Reconbinante: Genes e Genomas , 3a Ed.* [S.l.]: Artmed, Porto Alegre, 2009. 273-303 p.

2 VENTER, J. C. t. genome sequencing consortium. "initial sequencing and analysis of the human genome.". *Nature*, v. 409, p. 860–921, 2001.

3 GENBANK. `http://www.ncbi.nlm.nih.gov/genbank/`. Acessado em 18/07/2015.

4 EMBL-EBI, European Bioinformatics Institute, Swiss-Prot and TrEMBL. `http://web.expasy.org`. Acessado em 18/07/2015.

5 RCSB Protein Data Bank (RCSB PDB). `http://www.rcsb.org/`. Acessado em 19/05/2015.

6 BIELSKA, E. Virtual screening strategies in drug design - methods and applications. *Journal of Biotechnology*, v. 92, p. 249–264, 2011.

7 KLEBE, G. Virtual ligand screening: strategies, perspectives. *Drug Discov. Today*, v. 11, p. 580–594, 2006.

8 WILLET, P. Similarity-based virtual screening using 2d-fingerprints. *Drug Discovery Today*, v. 11, p. 1046–1053, 2006.

9 MOBILIO, D. t. A protein relational database and protein family knowledge bases to facilitate structure-based design analyses. *Chem. Biol. Drug. Des.*, v. 76, p. 142–153, 2010.

10 TODD, A. E.; ORENGO, C. A.; THORNTON, J. M. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, v. 307, p. 1113–1143, 2001.

11 ROGNAN, D. Structure based approaches to target fishing and ligand proffiling. *Molecular Informatics*, v. 29, p. 176–187, 2010.

12 MORRIS, G. et al. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, v. 19, p. 1639–1662, 1998.

13 OSTERBERG, F. et al. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins*, v. 46, p. 34–40, 2002.

14 TROTT, O.; OLSON, A. Autodock vina: improving the speed and accuracy of docking with a new scoring func-tion, efficient optimization, and multithreading. *J. Comput. Chem.*, v. 31, p. 455–461, 2010.

15 LUETHI, E. et al. Identification of selective norbornane-type aspartate analogue inhibitors of the glutamate transporter 1 (glt-1) from the chemical universe generated database (gdb). *J. Med. Chem.*, v. 53, p. 7236–7250, 2010.

16 PARK, H. et al. Discovery of novel cdc25 phosphatase inhibitors with micromolar activity based on the structure-based virtual screening. *J. Med. Chem.*, v. 51, p. 5533–5541, 2008.

17 KOVAC, A. et al.) discovery of new inhibitors of dalanine: D-alanine ligase by structure-based virtual screening. *J. Med. Chem*, v. 51, p. 7442–7448, 2008.

18 STEFFEN, A. et al. Improved cyclodextrin-based receptors for camptothecin by inverse virtual screening. *Chemistry*, v. 13, p. 6801–6809, 2007.

19 EWING, T. et al. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aid. Mol. Des.*, v. 15, p. 411–428, 2001.

20 KUNTZ, I. et al. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, v. 161, p. 269–288, 1982.

21 LANG, P. et al. Dock 6: combining techniques to model rna-small molecule complexes. *RNA*, v. 15, p. 1219–1230., 2009.

22 MOUSTAKAS, D. T. et al. Development and validation of a modular, extensible docking program: Dock 5. *J. Comp. Aid. Mol. Des.*, v. 20, p. 601–619, 2006.

23 OSHIRO, C. M. et al. Flexible ligand docking using a genetic algorithm. *J. Comp. Aid. Mol. Des.*, v. 9, p. 113–130, 1995.

24 CHEN, C. S. et al. Structure-based discovery of triphenylmethane derivatives as inhibitors of hepatitis c virus helicase. *J. Med. Chem.*, v. 52, p. 2716–2723, 2009a.

25 LIU, Z. et al. Virtual screening of novel noncovalent inhibitors for sars-cov 3clike proteinase. *J. Chem. Inf. Model.*, v. 45, p. 10–17, 2005.

26 MOZZICONACCI, J. C. et al. Optimization and validation of a docking-scoring protocol; application to virtual screening for cox-2 inhibitors. *J. Med. Chem.*, v. 48, p. 1055–1068, 2005.

27 RAREY, M. et al. A fast flexible docking method using an incremental construction algorithm. *Journal Molecular Biology*, v. 10, p. 470–489, 1996.

28 MORO, W. B. et al. Virtual screening to identify lead inhibitors for bacterial nad synthetase (nads). *Bioorg. Med. Chem. Lett*, v. 19, p. 2001–2005, 2009.

29 WU, S. et al. In silico screening for ptpn22 inhibitors: active hits from an inactive phosphatase conformation. *ChemMedChem*, v. 4, p. 440–444, 2009.

30 J., K. et al. Identification of novel hcv rna-dependent rna polymerase inhibitors using pharmacophore-guided virtual screening. *Chem. Biol. Drug. Des.*, v. 72, p. 585–591, 2008.

31 CHO, Y. et al. Discovery of novel nitrobenzothiazole inhibitors for mycobacterium tuberculosis atp phosphoribosyl transferase (hisg) through virtual screening. *J. Med. Chem.*, v. 51, p. 5984–5992, 2008.

32 R., K. et al. Discovery of novel human histamine h4 receptor ligands by large-scale structure-based virtual screening. *J. Med. Chem.*, v. 51, p. 3145–3153, 2008.

33 FRIESNER, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, v. 47, p. 1739–1749, 2004.

34 CHENG, J. et al. Combination of virtual screening and high throughput gene profiling for identification of novel liver x receptor modulators. *J. Med. Chem.*, v. 51, p. 2057–2061, 2008.

35 G., G. B. et al. Search for nonnucleoside inhibitors of hiv-1 reverse transcriptase using chemical similarity, molecular docking, and mm-gb/sa scoring. *J. Chem. Inf. Model*, v. 47, p. 2416–2428, 2007.

36 VERDONK, M. L. et al. Improved protein-ligand docking using gold. *Proteins*, v. 2003, p. 609–623, 52.

37 VERDONK, M. L. et al. Modeling water molecules in protein-ligand docking using gold. *J. Med. Chem.*, v. 48, p. 6504–6515, 2005.

38 LALONDE, J. M. et al. Design, synthesis and biological evaluation of small molecule inhibitors of cd4-gp120 binding based on virtual screening. *Bioorg. Med. Chem.*, v. 19, p. 91–101, 2011.

39 KURCZAB, R. et al. The development and validation of a novel virtual screening cascade protocol to identify potential serotonin 5-ht(7)r antagonists. *Bioorg. Med. Chem. Lett.*, v. 20, p. 2465–2468, 2010.

40 XU, W. et al. Identification of a submicromolar, non-peptide inhibitor of β - secretase with low neural cytotoxicity through in silico screening. *Bioorg. Med. Chem. Lett.*, v. 20, p. 5763–5766, 2010.

41 DEYE, J. et al. Structure-based virtual screening for novel inhibitors of the sarco/endoplasmic reticulum calcium atpase and their experimental evaluation. *Bioorgan. Med. Chem.*, v. 17, p. 1353–1360, 2009.

42 NERES, J. et al. Discovery of novel inhibitors of trypanosoma cruzi trans-sialidase from in silico screening. *Bioorg. Med. Chem. Lett.*, v. 19, p. 589–596, 2009.

43 JAIN, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, v. 46, p. 499–511, 2003.

44 JAIN, A. Surflex-dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledgebased search. *J. Comp. Aid. Mol. Des.*, v. 21, p. 281–306, 2007.

45 HOLT, P. A. et al. Discovery of novel triple helical dna intercalators by an integrated virtual and actual screening platform. *Nucl. Acids Res.*, v. 37, p. 1280–1287, 2009.

46 FEDER, M. et al. Virtual screening and experimental verification to identify potential inhibitors of the ermc methyltransferase responsible for bacterial resistance against macrolide antibiotics. *ChemMedChem*, v. 3, p. 316–322, 2008.

47 MUSMUCA, I. et al. Combining 3-d quantitative structure-activity relationship with ligand based and structure based alignment procedures for in silico screening of new hepatitis c virus ns5b polymerase inhibitors. *J. Chem. Inf. Model.*, v. 50, p. 662–676, 2010.

48 DATABASE PubChem. http://pubchem.ncbi.nlm.nih.gov. Acessado em 18/07/2015.

49 CHEMBL, The European Bioinformatics Institute. https://www.ebi.ac.uk/chembl/. Acessado em 18/07/2015.

50 CHEMSPIDER, The free chemical database. http://www.chemspider.com/. Acessado em 18/07/2015.

51 MDL Drug Data Report. http://www.akosgmbh.de/Symyx/software/databases/mddr.html. Acessado em 18/07/2015.

52 WORLD of Molecular BioAcTivity. http://www.sunsetmolecular.com. Acessado em 18/07/2015.

53 KALYAANAMOORTHY, S.; CHEN, Y.-P. P. Structure-based drug design to augment hit discovery. *Drug Discovery Today*, v. 16, p. 831–839, 2011.

54 RIPPHAUSEN, P.; NISIUS, B.; BAJORATH, J. State-of-the art in ligand-based virtual screening. *Drug Discovery today*, v. 16, p. 372–376, 2011.

55 CROSS, S. et al. Flap: Grid molecular interaction fields in virtual screening. validation using the dud data set. *Chem. Inf. Model.*, v. 50, p. 1442–1450, 2010.

56 PERRUCCIO, F. et al. Flap: 4-point pharmacophore fingerprints from grid. in cruciani, g. (ed.). molecular interaction fields: Applications in drug discovery and adme prediction. *Weinheim, Germany: Wiley-VCH*, chapter 4, 2006.

57 BARONI, M. et al. A common reference framework fo analyzing/comparing proteins ans ligands. fingerprints for ligands and proteins (flap): Theory and application. *J. Chem Info Model*, v. 47, p. 279–294, 2007.

58 CROSS, S. S. J. Improved flexx docking using flexs-determined base fragment placement. *J. Chem Info Model*, v. 45, p. 993–1001, 2005.

59 LANGER, T.; HOFFMANN, R. D. *Pharmacophores and Pharmacophore Searches. v. 32.* [S.l.]: Wiley-VCH, Weinheim, Germany, 2006.

60 LEACH, A. R. et al. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.*, v. 53, p. 539–558, 2010.

61 BAJORATH, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, v. 1, p. 882–894, 2002.

62 WILLET, P. Similarity searching using 2d structural fingerprints. In: _____. *Chemoinformatics and Computational Chemical Biology*. [S.l.: s.n.], 2011. cap. 5, p. 133–158.

63 MAGGIORA, G. M. On outliers and activity cliffs: Why qsar often disapoints. *Journal of Chemical Information and Modeling*, v. 46, p. 1535–1535, 2006.

64 PELTASON, L.; BAJORATH, J. Sar index: Quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry*, v. 50, n. 23, p. 5571–5578, 2007.

65 BAJORATH, J.; ECKERT, H. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, v. 12, p. 225–233, 2007.

66 TROPSHA, A.; GOLBRAIKH, A. Predictive qsar modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*, v. 16, p. 357–369, 2002.

67 LAVECCHIA, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, v. 00, p. 1–14, 2014.

68 SCREENMD. `https://docs.chemaxon.com/display/screen/ScreenMD`. Acessado em 09/07/2015.

69 KALáSZI, A. et al. Screen3d: A novel fully flexible high-throughput shape-similarity search method. *J. Chem. Inf. Model.*, v. 54, n. 4, p. 1036–1049, 2014.

70 CHEESERIGHT, T. J. et al. Fieldscreen: Virtual screening using molecular fields. application to the dud data set. *J. Chem. Inf. Model.*, v. 48, p. 2108–2117, 2008.

71 SCHNEIDMAN-DUHOVNY, D. et al. Pharmagist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Research*, v. 36, p. 223–228, 2008.

72 LIU, X.; JIANG, H.; H., L. Shafts: A hybrid approach for 3d molecular similarity calculation. 1. method and assessment of virtual screening. *J. Chem. Inf. Model.*, v. 51, p. 2372–2385, 2011.

73 LIU, X. et al. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics*, v. 10, n. 1, p. 101, 2009.

74 OPENEYE Scientific Software. ROCS. 2005-2015. `http://www.eyesopen.com/rocs`. Acessado em 09/07/2015.

75 LóPEZ-RAMOS, M.; PERRUCCIO, F. Hppd: Ligand- and target-based virtual screening on a herbicide target. *J. Chem. Inf. Model.*, v. 50, p. 801–814, 2010.

76 SASTRY, G. M.; DIXON, S.; W., S. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J. Chem. Inf. Model.*, v. 51, p. 2455–2466, 2011.

77 JAHN, A. et al. 4d flexible atom-pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening. *Journal of Cheminformatics*, v. 3, n. 1, p. 23, 2011.

78 JAHN, A. et al. Probabilistic modeling of conformational space for 3d machine learning approaches. *Molecular Informatics*, v. 29, n. 5, p. 441–455, 2010.

79 JAHN, A. et al. Boltzmann-enhanced flexible atom-pair kernel with dynamic dimension reduction. *Molecular Informatics*, v. 30, n. 4, p. 307–315, 2011.

80 REDDY, A. S. Virtual screening in drug discovery: A computational perspective. *Current Protein and Peptide Science*, v. 8, p. 329–351, 2007.

81 GEPPERT, H.; VOGT, M.; BAJORATH, J. urgen. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.*, v. 50, p. 205–216, 2010.

82 TRUCHON, J.; BAYLY, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.*, v. 47, p. 488–508, 2007.

83 NICHOLLS, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.*, v. 22, p. 239–255, 2008.

84 CARUGO, O. Recent progress in measuring structural similarity between proteins. *Curr. Prot. Pept. Sci.*, v. 8, p. 219–241, 2007.

85 KOLODNY, R.; PETREY, D.; HONIG, B. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol.*, v. 16, p. 393–398, 2006.

86 KELLENBERGER, E. et al. How to measure the similarity between protein ligand-binding sites? *Current Computer-Aided Drug Design*, v. 4, p. 209–220, 2008.

87 PROSDOCIMI, F. t. Bioinformática: Manual do usuário. *Biotecnologia Ciência & Desenvolvimento*, v. 29, p. 12–25, 2002.

88 SHULMAN-PELEG, A.; NUSSINOV, R.; WOLFSON, H. Recognition of functional sites in protein structures. *J. Mol. Biol.*, v. 339, p. 607–633, 2004.

89 MINAI, R. et al. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins: Structure, Function, and Bioinformatics*, v. 72, p. 367–381, 2008.

90 SCHMITT, S.; KUHN, D.; KLEBE, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, v. 323, p. 387–406, 2002.

91 KINOSHITA, K.; FURUI, J.; NAKAMURA, H. J. Identification of protein functions from a molecular surface database, ef-site. *Journal of Structural and Functional Genomics*, v. 2, p. 9–22, 2001.

92 JAMBON, M. et al. A new bioinformatic approach to detect common 3d sites in protein structures. *Proteins: Structure, Function, and Bioinformatics*, v. 52, p. 137–145, 2003.

93 POWERS, R. et al. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins: Structure, Function, and Bioinformatics*, v. 65, p. 124–135, 2006.

94 BRAKOULIAS, A.; JACKSON, R. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: An automated all-against-all structural comparison using geometric matching. *Proteins: Structure, Function, and Bioinformatics*, v. 56, p. 250–260, 2004.

95 GARDINER, E.; ARTYMIUK, P.; WILLETT, P. Clique-detection algorithms for matching three-dimensional molecular structures. *Journal of Molecular Graphics and Modelling*, v. 15, p. 245–253, 1997.

96 SCHALON, C. et al. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Structure, Function, and Bioinformatics*, v. 71, p. 1755–1778, 2008.

97 WEILL, N.; ROGNAN, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.*, v. 50, p. 123–135, 2010.

98 PEROT, S. et al. Druggable pockets and binding sitecentric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, v. 15, p. 656–667, 2010.

99 EF-SEEK. `http://ef-site.hgc.jp/eF-seek/top.do`. Acessado em 05/06/15.

100 FINDSITE. `http://cssb.biology.gatech.edu/skolnick/files/FINDSITE/`. Acessado em 09/06/15.

101 SUMO. `http://sumo-pbil.ibcp.fr`. Acessado em 05/06/15.

102 Q-SITEFINDER. `http://www.modelling.leeds.ac.uk/qsitefinder/`. Acessado em 08/06/15.

103 THINK. `http://www.treweren.com`. Acessado em 21/04/15.

104 KIRCHMAIR, J. et al. Evaluation of the performance of 3d virtual screening protocols: Rmsd comparisons, enrichment assessments, and decoy selection - what can we learn from earlier mistakes? *J. Comput. Aided. Mol. Des.*, v. 22, p. 213–228, 2008.

105 AGRAFIOTIS, D. K. et al. Conformational sampling of bioactive molecules: A comparative study. *J. Chem. Inf. Model.*, v. 47, p. 1067–1086, 2007.

106 SCHWAB, C. H. Conformations and 3d pharmacophore searching. *Drug Discovery Today: Technologies*, v. 7, p. 245–253, 2010.

107 SMELLIE, A. et al. Poling: promoting conformational variation. *J. Comput. Chem.*, v. 16, p. 171–187, 1995.

108 FERGUSON, D.; RABER, D. A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *J. Am. Chem. Soc.*, v. 111, p. 4371–4378, 1989.

109 HAWKINS, P. et al. Conformer generation with omega: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.*, v. 50, p. 572–584, 2010.

110 MOHAMADI, F. et al. Macromodel : an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.*, v. 11, p. 440–467, 1990.

111 TROPSHA, A. Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, v. 29, p. 476–488, 2010.

112 LANGER, T. Pharmachophores in drug research. *Molecular Informatics*, v. 29, p. 470–475, 2010.

113 THE International Union of Pure and Applied Chemistry (IUPAC). `http://www.chem.qmul.ac.uk/iupac/medchem/ix.html#p7`. Acessado em 18/03/2015.

114 SEIDEL, T. et al. Strategies for 3d pharmacophore based virtual screening. *Drug Discovery Today: Technologies*, v. 7, p. 221–228, 2010.

115 DRIE, J. H. V. History of 3d pharmacophore searching: commercial, academic and open-source tools. *Drug Discvery Today: Technologies*, v. 7, p. 255–262, 2010.

116 DRIE, J. H. V. Monty kier and the origin of the pharmacophore concept. internet electron. *J. Mol. Des.*, v. 6, p. 271–279, 2007.

117 KIER, L. B. Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone . *Mol. Pharmacol.*, v. 3, p. 487–494, 1967.

118 KIER, L. B. Receptor mapping using mo theory. In: *Fundamental Concepts in Drug-Receptor Interactions by Danielli, J. F., Moran, J. F., Triggle, D. J.* [S.l.]: Academic Press: New York, 1970.

119 KIER, L. B. *MO Theory in Drug Research.* [S.l.]: Academic Press: New York, 1971. 164-169 p.

120 HERT, J. et al. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J Chem Inf Comput Sci*, v. 44: 3, p. 1177–1185, 2004.

121 XU, J. et al. *Applications of Fuzzy Logic in Bioinformatics.* [S.l.]: Imperial College Press, London, England, 248 pages, 2008.

122 WOOLF, P. J.; WANG, Y. A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics*, v. 3, p. 9–15, 2000.

123 ASHBURN, T. T.; THOR, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, v. 3, p. 673–683, 2004.

124 ARROWSMITH, J. A decade of change . *Nature Rev Drug Discov*, v. 11, p. 17–18, 2012.

125 SHAFIEI, N. et al. Transformation in the pharmaceutical industry - a systematic review of the literature. *PDA J Pharm Sci and Tech*, v. 67, p. 105–122, 2013.

126 SCHULZE, U. et al. R&d productivity, on the comeback trail. *Nature Reviews Drug Discovery*, v. 11, p. 17–18, 2014.

127 LABORATóRIO Nacional de Biociências. `http://www.brasil.gov.br/ciencia-e-tecnologia/2014/04/projetos-propoem-testes-alternativos-ao-uso-de-animais-em-pesquisas`. Acessado em 22/04/2015.

128 HAUSMAN, R. E.; COOPER, G. M. *The cell: a molecular approach.* [S.l.]: ASM Press, Washington D.C., 2004. 51 p.

129 XUE, M.; JIANG, H.; SHEN, J. Bssf: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinformatics*, v. 11, p. 1–11, 2010.

130 XUE, L.; GODDEN, J.; BAJORATH, J. Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory. *SAR and QSAR in Environmental Research*, v. 14, p. 27–40, 2003.

131 MELVILLE J. L.; RILEY, J. F.; HIRST, J. D. Similarity by compression. *J. Chem. Inf. Model.*, v. 47, p. 25–33, 2007.

132 FELDMAN, H. J.; LABUTE, P. Pocket similarity: Are carbons enough? *J. Chem. Inf. Model.*, v. 50, p. 2010, 1466–1475.

133 KOGEJ, T. et al. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.*, v. 46, p. 1201–1213, 2006.

134 POPTODOROV, K.; LUU T; HOFFMANN, R. Pharmacophore model generation software tools. In: _____. *Pharmacophores and Pharmacophore Searches.* [S.l.]: Wiley-VCH, 2006.

135 HOLLIDAY, J. et al. Grouping of coefficients for the calculation of inte-molecular similarity and dissimilarity using 2d fragment bit-string. *Combinatorial Chemistry and Hight Throughput Screening*, v. 5, p. 155–156, 2002.

136 DAYLIGHT Theory: Fingerprints. `http://www.daylight.com/dayhtml/doc/theory/theory.finger.html`. Acessado em 29/06/2015.

137 KORFF, M. von; FREYSS, J.; SANDER, T. Comparison of ligand- and structure-based virtual screening on the dud data set. *J. Chem.Inf. Model.*, v. 49, p. 209–231, 2009.

138 TRUCHON, J.; BAYLY, C. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information Modeling*, v. 47, p. 488–508, 2007.

139 CLARK, R.; WEBSTER-CLARK, D. Managing bias in roc curves. *Journal of Computer-Aided Molecular Design*, v. 22, p. 141–146, 2008.

140 VENKATRAMAN, V. et al. Comprehensive comparison of ligand-based virtual screening tools against the dud data set reveals limitations of current 3d methods. *J. Chem. Inf. Model.*, v. 50, p. 2079 – 2093, 2010.

141 ZHAO, W. et al. A statistical framework to evaluate virtual screening. *BMC Bioinformatics*, v. 10, p. 1–13, 2009.

142 WALD, N. J.; BESTWICK, J. Is the area under an roc curve a valid measure of the performance of a screening or diagnostic test? *Journal of Medical Screening*, v. 21, p. 51–56, 2014.

143 PEIRCE, C. S. The numerical measure of the success of predictions. *Science*, v. 4, n. 93, p. 453–454, 1884.

144 YOUDEN, W. J.; CONNOR, W. S. The chain block design. *Biometrics*, v. 9, n. 2, p. 127–140, 1953.

145 POWERS, D. M. W. *International Conference on Cognitive Science (ICCS)*. 2003. `http://david.wardpowers.info/BM/index.htm`. Acessado em 29/06/2015.

146 TROPSHA, A.; GOLBRAIKH, A. Predictive qsar modeling workflow, model applicability domains, and virtual screening. *Current Pharmaceutical Design*, v. 13, p. 3494–3504, 2007.

147 CORTéS-CABRERA, A.; GAGO, F.; MORREALE, A. A reverse combination of structure-based and ligand-based strategies for virtual screening. *Journal of Computer-Aided Molecular Design*, v. 26, p. 319–327, 2012.

148 EFRON, B.; TIBISHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: John Wiley and Sons, New York, 1993. 642 p.

149 AUNG, Z.; TONG, J. C. A rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *Genome Inf.*, v. 21, p. 65–76, 2006.

150 YETURU, K.; CHANDRA, N. Pocketmatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.*, v. 9, p. 543, 2008.

151 HOFFMANN, B. et al. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d application to ligand prediction. *BMC Bioinf.*, v. 11, n. 99, p. 1–16, 2010.

152 KELLENBERGER, E. et al. sc-pdb: an annotated database of druggable binding sites from the protein data bank. *J. Chem. Inf. Model.*, v. 46, p. 717–727, 2006.

153 SC-PDB Home Page. `http://cheminfo.u-strasbg.fr:8080/scPDB/2011/db_search/acceuil.jsp?uid=3807447192798425088`. Acessado em 01/08/2015.

154 JAHN, A. et al. Optimal assignment methods for ligand-based virtual screening. *Journal of Cheminformatics*, v. 1, p. 1–14, 2009.

155 A.C., G.; T.I., O. Optimization of camd techniques 3. virtual screen- ing enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.*, v. 22, n. 3-4, p. 169–178, 2008.

156 LIU, X. et al. Pharmmapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Research*, v. 38, p. 609–614, 2010.

157 HU, G. et al. Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *J. Chem. Inf. Model.*, v. 52, p. 1103–1113, 2012.

158 VAINIO, M. J.; PURANEN, J. S.; JOHNSON, M. S. Shaep: Molecular overlay based on shape and electrostatic potential. *J. Chem. Inf. Model.*, v. 42, n. 2, 2009.

159 GRANT, J. A.; GALLARDO, M. A.; PICKUP, B. T. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *J. Comput. Chem.*, v. 17, p. 1653–1666, 1996.

160 LIBSVM. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. Acessado em 09/07/2015.

161 BURSI, R.; KAZIUS, J.; MCGUIRE, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, v. 48, p. 312–320, 2005.

162 XU, C. et al. In silico prediction of chemical ames mutagenicity. *J. Chem. Inf. Model.*, v. 52, n. 11, p. 2840–2847, 2012.

163 MCCARREN, P.; SPRINGER, C.; WHITEHEAD, L. An investigation into pharmaceutically relevant mutagenicity data and the influence on ames predictive potential. *J. Cheminf.*, v. 3, p. 51, 2011.

164 HALPERIN, I. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct. Funct. and Gene.*, v. 47, p. 409–443, 2002.

165 JONES, G.; WILLETT, P.; GLEN, R. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal Molecular Biology*, v. 10, p. 43–53, 1995.

166 DANISHUDDIN, M.; KHAN, A. Structure based virtual screening to discover putative drug candidates: Necessary considerations and successful case studies. *Methods*, v. 71, p. 135–145, 2015.

167 MORRIS, G. M. et al. Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function. *Journal of Computational Chemistry*, v. 19, p. 1639–1662, 1999.

168 KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, v. 3, p. 935–949, 2004.

169 MCGANN, M. Fred pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, v. 51, p. 578–596, 2011.

170 MITEVA, M. A. et al. Fast structure-based virtual ligand screening combining fred, dock, and surflex. *J. Med. Chem.*, v. 48, p. 6012–6022, 2005.

Apêndice A

Algoritmos aplicados ao estudo de Triagem Virtual Baseada em Alvos Biológicos (TBVS)

A.1 Classificação Algoritmos TBVS - Docking

Dependendo da metodologia aplicada, os algoritmos de docking para TBVS [1] podem superficialmente classificados em três categorias (**Figura A.1**):

a - algoritmos que realizam a busca do espaço conformacional durante o docking;

b - algoritmos que realizam a busca do espaço conformacional antes do docking;

c - docking incremental.

a) A primeira categoria de algoritmos realiza a otimização conformacional e orientacional das pequenas moléculas na cavidade do sítio ligante do receptor, mas, devido à complexidade deste problema a sua aplicação é dificultada para uso em grandes bases de dados. Como é sabido as moléculas podem apresentar diferentes conformações em solução e a quantidade de conformações possíveis pode ser dada pela seguinte fórmula:

$$\text{Número de Conformações} = \prod_{i=1}^{n} \frac{360^{\circ}}{\Theta} \tag{A.1}$$

[1] TBVS = *"Target Based Virtual Screening"*, Triagem Virtual Baseada em Alvos Biológicos

Onde n representa o número de ligações rotacionáveis e Θ o ângulo incremental utilizado, ou seja, o ângulo que irá ser usado para girar a ligação a cada passo. Por isso, muitas vezes são utilizados algoritmos estocásticos, resultados gerados a partir de condições aleatórias buscando as conformações mais estáveis possíveis nas posições de mínimos de energia. Entre os algoritmos estocásticos mais utilizados estão os métodos de Monte Carlo, "Simulated Annealing" e os algoritmos genéticos (80).

Os métodos de Monte Carlo (MC) são baseados na utilização de técnicas de amostragem e geração aleatórias de estados conformacionais de baixa energia. Para isso, o sistema realiza movimentos aleatórios nas moléculas e a nova conformação é aceita ou rejeitada tendo por base o cálculo estatístico de que essa nova conformação seja melhor que a anterior, geralmente é aplicada uma função de distribuição de probabilidade de Boltzmann. Essa metodologia está entre as técnicas de otimização estocástica mais consolidadas e é uma das mais utilizadas atualmente. Já o "Simulated annealing" pode ser considerado um método de MC generalizado, onde o estado inicial do sistema é dado por movimentos termais aleatórios dentro de um campo de força específico, e a temperatura dos sistemas (também referenciado como grau de liberdade) é diminuída com o tempo até que seja obtida a posição ancorada mais estável.

Já os Algoritmos Genéticos (GA) aplicam buscas heurísticas adaptativas baseadas nos conceitos evolucionários de mutação genética e seleção natural. Assim, eles são projetados para simular os processos de seleção natural indispensáveis para a evolução. Estes algoritmos mantêm parâmetros para forçar uma pressão seletiva na tentativa de conduzir a uma solução ótima. A princípio uma população de soluções é submetida a transformações de mutação e de cruzamento. As soluções recém-geradas passam por seleção, direcionadas pela grau de adaptação ao ambiente, sendo selecionadas somente as melhores. Entre os programas de ancoramento molecular em existe um GA implementado estão o GOLD, AutoDock e DARWIN. Os programas de ancoragem molecular com base em procedimentos estocásticos, tem o potencial de apresentar boas soluções para complexos proteina-ligante mesmo com ligantes muito grandes e flexíveis. Entretanto a aplicação prática desses métodos apresenta problemas em relação às outras duas categorias apresentadas à frente. A principal delas é o custo computacional, pois mesmo que a execução de uma ancoragem seja relativamente rápida esse processo deve ser repetido várias vezes para que se considere a estrutura prevista confiável (164).

b) Uma segunda classe de algoritmos separa a busca conformacional de moléculas pequenas da sua ancoragem no sitio de ligação. Em primeiro lugar é realizada uma análise conformacional, e todas as conformações relevantes de baixa energia são rigidamente ancoradas ao sítio. Para resolver o problema de explosão conformacional, geralmente são considerados ligantes com até seis graus de liberdade para rotação e translação. Os programas Slide e Fred utilizam essa metodologia de ancoragem.

c) A Terceira metodologia é baseada no desmonte e na construção incremental dos ligantes. Primeiro as ligações rotacionáveis dos ligantes são quebradas, gerando fragmentos que são rigidamente ancorados nas suas várias posições favoráveis no sítio ativo, os que apresentarem melhores interações são selecionados e passam a ser considerados como fragmentos base. A seguir os outros fragmentos são adicionados em várias posições do sítio e essas são então ranqueadas. Este processo é repetido até que todo o ligante seja montado. Dentre as ferramentas de ancoragem a usar esta metodologia estão o FlexX, Surflex, HOOK e um componente do DOCK4.0 (80).

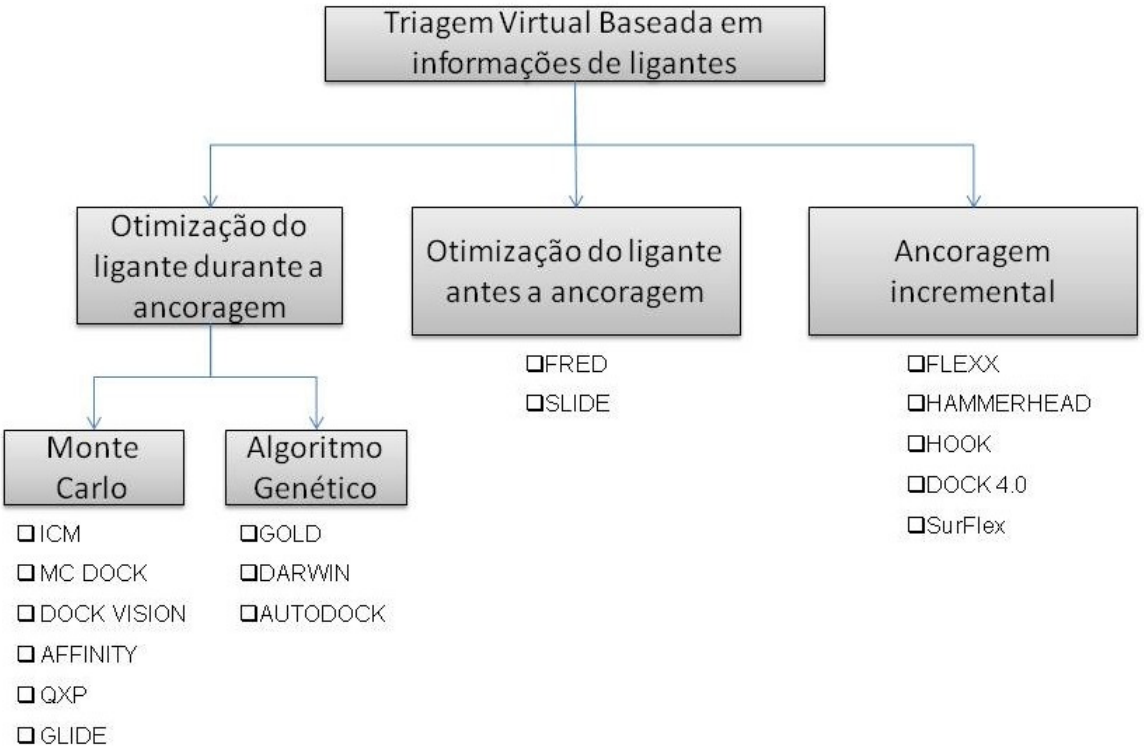


Figura A.1: Classificação dos vários métodos de TBVS. Adaptada de (80).

A.2 Principais programas para ancoragem molecular ("Molecular Docking")

Atualmente existem vários programas disponíveis para a realização de estudos de ancoragem molecular, cada uma deles com suas vantagens e desvantagens como mostra a tabela **Tabela 1.2**. Apesar disso, poucos são conhecidos por apresentarem histórias de sucesso no descobrimento de novos fármacos. Abaixo serão listadas as características dos principais deles.

GOLD (Genetic Optimization of Ligand Doking) (165)

GOLD é um programa de ancoragem automatizada que aplica um algoritmo evolucionário que visa à exploração conformacional do ligante através da aplicação de uma flexibilidade parcial nas proteínas (166), isso o torna uma poderosa ferramenta para a triagem e identificação de novos compostos líder. Ainda, os parâmetros do seu algoritmo genético podem ser modificados e otimizados para variadas aplicações de triagem virtual (80). A medida de adaptação de cada molécula avalia as conformações dos ligantes dentro do complexo e é calculada tomando por base três termos de energia (energia de ligações de hidrogênio, energia de interação estérica, energia de ligação intramolecular). Outra vantagem deste método é a capacidade de lidar facilmente com moléculas de água e íons no sítio ativo de proteínas e enzimas (166).

FlexX

FlexX usa um algoritmo de construção incremental onde os ligantes são quebrados nas ligações rotacionáveis e a ancoragem é iniciada com essa base de fragmentos (27) . A melhor solução é avaliada pela entropia, ligações de hidrogênio, lipofilicidade, ionicidade e aromaticidade por uma função de Bohm modificada. Assim o FlexX é capaz de predizer a geometria do complexo proteína-ligante e estimar afinidade da ligação (80). Este programa é conhecido por predizer localizações potenciais de moléculas de água preferenciais em posições cristalográficas (166) .

AutoDock (167)

Audodock é baseado em algoritmo de ancoragem que calcula termos de flexibilidade automaticamente e busca as melhores posições dos ligantes aleatoriamente. Ele altera as posições dos ligantes ou só uma parte deles, rotacionando suas ligações e prediz a energia de interação entre pequenas moléculas e um receptor com estrutura tridimensional conhecida. Para isso ele utiliza de algoritmos de Monte Carlo, genético evolucionário e genético Lamarquiano. Esta ferramenta é conhecida por ser o programa mais usado entre os pesquisadores na área de desenvolvimento computacional de fármacos, mostrando várias histórias de sucesso no decorrer do tempo. Além de ser possível sua utilização com linhas de comando o que torna possível a automatização através de scripts ele ainda possui um interface gráfica - o Autodock Tools (ADT)-, muito eficiente e simples de utilizar, tornando o trabalho mais agradável para aqueles avessos aos terminais de controle do Linux ou do Window (DOS).

AutoDock Vina (14)

Desenvolvido pelo mesmo grupo que desenvolveu o Autodock 4, *"The Scripps Research Institute"* - Califórnia, o AutoDock Vina (http://vina.scripps.edu) usa um sofisticado método de otimização de gradiente em seu procedimento de otimização local que dá um "senso direcional" para a avaliação das ancoragens. Além disso o Vina uma função de scoring mais simples que o AutoDock e permiti o uso de processamento paralelo em máquinas que possuem vários processadores, o que possibilita que ele seja mais rápido que o Autodock em aproximadamente duas ordens de grandeza. Além disso, esse programa apresenta uma maior acurácia na predição do modo de ligação de moléculas, além de ser mais eficiente que o Autodock 4 em lidar con ligantes com mais de 20 ligações flexíveis. Esse programa pode ser usado com a ajuda do AutoDockTools (ADT), através de terminal de comandos ou, ainda, através de scripts (o que possibilita a automatização de estudos de triagem virtual em bases de dados maiores).

DOCK (168)

As estratégias de busca incluem construção incremental e busca aleatória por conformações além de utilizar funções Coulombica e de Lennard-Jones para ranqueamento. Ele descreve os ligantes e os sítios de ligação geometricamente através de um conjunto de esferas, após isso ele tenta encaixar cada composto da base de dados dentro do sítio de ligação (80). A versão atualizada do DOCK usa energias de interações estéricas, eletrostáticas e de mecânica molecular para calcular a eficácia do

complexo receptor-ligante (166).

GLIDE (Grid-based Ligand Docking with Energetic)

O algoritmo Glide com seus filtros hierárquicos aproximam uma busca sistemática conformacional, orientacional e do espaço posicional do ligante no sítio de ligação do alvo. A melhor conformação é selecionada basendo-se em uma função de ranqueamento que combina termos empíricos e baseados em campos de força. O passo de geração de um grid requer a inserção dos arquivos com as estruturas, tanto dos ligantes quanto do sítio, com a inclusão dos seus átomos de hidrogênio (166).

FRED (Fast Rigid Exhaustive Docking)

O FRED executa a ancoragem em duas etapas: adaptação da forma e (2) otimização. O FRED requer diferentes conformeros dos ligantes. Bibliotecas de conformeros de ligantes podem ser geradas pelo uso do OMEGA (Open Eye Scientific Software). Durante a simulação da ancoragem cada ligante é colocado dentro de uma caixa de grid específica com todos os átomos do sítio ativo usando um potencial Gaussiano suavizado (166), sendo que tanto ligante quanto o sítio são tratados como estruturas rígidas na maior parte do tempo. A estratégia do FRED é calcular todas as posições de cada ligante exaustivamente dentro do sítio. Esta busca exaustiva é baseada em realização de rotações e translações em cada um dos compostos, o que evita amostragens associadas a métodos estocásticos usadas por outros programas de ancoragem (166). Funções de ranqueamento como: ranqueamento pela forma Gaussiana, ChemScore, ScreenScore, Chemical Gaussian Overlay (CGO) e Chemical Gaussian Tanimoto (CGT) são usados na etapa de otimização para ranquear os ligantes (169) . Outra característica do FRED, e que o torna atrativo à triagem virtual em grandes bases de dados é a facilidade de ser distribuído em múltiplos processadores, reduzindo bastante o tempo de execução (170).

DARWIN

O DARWIN utiliza um algoritmo genético para otimizar as conformações moleculares guiado por um fator de seletividade dado pelo calculo de energia potencial

do complexo calculado por mecânica molecular através do programa CHARMM. O DARWIN combina o algoritmo genético com uma uma estatégia de busca baseada na minimização do gradiente energético (80).

HAMMERHEAD

Hammerhead é útil para a triagem em grandes bases de dados de moléculas flexíveis para serem ancoradas a um sítio de ligação de uma proteína de estrutura tridimensional conhecida. Ele fornece resultados precisos para uma ampla gama de ligantes flexíveis conhecidos e gasta em média poucos segundos para cada composto. A sua abordagem é totalmente automatizada, desde a elucidação dos sítios ativos da proteína, passando pela ancoragem da molécula até a seleção dos compostos que formam complexos mais estáveis (80).

SURFLEX

Atualmente o Surflex é um módulo de ancoragem disponível na plataforma de modelagem molecular Sybyl (Tripos). O algoritmo do Surflex constrói um sítio ativo idealizado chamado de "protomol". O protomol é produzido a partir de resíduos que constituem o sítio ativo da proteína utilizando um arquivo mol2 da proteína contendo os hidrogênios (86). O arquivo de entrada do ligante deve estar no formado mol2, cada ligante é fragmentado gerando entre 1 e 10 fragmentos, que podem conter ligações rotacionáveis. Cada fragmento é submetido a uma busca conformacional e uma reconstrução incremental para gerar as conformações. A ancoragem dos ligantes e montagem das conformações pode ser feita pala superposição dos fragmentos ao protomol ou de moléculas inteiras e os resultados são dados pela combinação de funções empíricas de ranqueamento do Hammerhead com métodos de similaridade molecular (43).

Apêndice B

Resultados do 3DPharma em análises no de VS no DUD.

Tabela B.1: Resultados, em valores de AUCROC, utilizando várias metodologias de construção de fingerprints para a recuperação de ativos na base de dados DUD.

Alvo	2P				3P				4P			
	Tanimoto		Simpson		Tanimoto		Simpson		Tanimoto		Simpson	
	Normal	Rognan	Normal	Rognan	Normal	Rognan	Normal	Rognan	Normal	Rognan	Normal	Rognan
ace	0,523	0,510	0,810	0,795	0,596	0,519	0,795	0,779	0,582	0,531	0,760	0,776
ache	0,768	0,651	0,857	0,649	0,770	0,699	0,902	0,825	0,758	0,730	0,905	0,875
ada	0,678	0,523	0,680	0,516	0,745	0,614	0,828	0,689	0,519	0,675	0,619	0,770
alr2	0,541	0,521	0,778	0,742	0,646	0,581	0,843	0,819	0,657	0,563	0,866	0,831
ampc	0,598	0,518	0,886	0,827	0,735	0,699	0,861	0,861	0,746	0,735	0,876	0,864
ar	0,584	0,579	0,672	0,728	0,662	0,683	0,689	0,751	0,664	0,665	0,699	0,720
cdk2	0,555	0,510	0,505	0,434	0,604	0,526	0,616	0,508	0,636	0,574	0,672	0,598
comt	0,403	0,409	0,257	0,332	0,418	0,397	0,552	0,448	0,408	0,400	0,550	0,436
cox1	0,562	0,501	0,541	0,456	0,523	0,564	0,484	0,478	0,585	0,619	0,526	0,533
cox2	0,707	0,639	0,878	0,861	0,881	0,873	0,901	0,894	0,853	0,869	0,879	0,881
dhfr	0,592	0,554	0,563	0,445	0,821	0,701	0,734	0,592	0,877	0,771	0,783	0,660
egfr	0,458	0,353	0,637	0,593	0,803	0,594	0,753	0,682	0,879	0,688	0,773	0,707
er _ agonist	0,692	0,534	0,674	0,525	0,733	0,682	0,725	0,622	0,726	0,636	0,738	0,647
er _ antagonist	0,791	0,637	0,907	0,744	0,966	0,926	0,965	0,925	0,981	0,973	0,968	0,949
fgfr1	0,342	0,214	0,346	0,241	0,248	0,210	0,332	0,257	0,362	0,358	0,415	0,323
fxa	0,673	0,736	0,700	0,721	0,485	0,471	0,644	0,642	0,525	0,523	0,659	0,687
gart	0,571	0,819	0,618	0,677	0,535	0,602	0,451	0,474	0,432	0,479	0,360	0,403
gpb	0,847	0,792	0,859	0,567	0,881	0,849	0,926	0,881	0,896	0,872	0,947	0,924
gr	0,673	0,639	0,665	0,644	0,843	0,844	0,867	0,848	0,881	0,875	0,901	0,898
hivpr	0,000	0,056	0,250	0,250	0,083	0,028	0,250	0,250	0,111	0,028	0,333	0,250
hivrt	0,560	0,471	0,526	0,449	0,686	0,606	0,715	0,683	0,730	0,674	0,791	0,748
hmga	0,911	0,897	0,912	0,911	0,910	0,912	0,918	0,918	0,906	0,906	0,913	0,919
hsp90	0,554	0,586	0,570	0,536	0,720	0,730	0,766	0,791	0,827	0,759	0,813	0,806
inha	0,398	0,390	0,635	0,615	0,466	0,460	0,700	0,684	0,517	0,477	0,778	0,737
mr	0,571	0,556	0,945	0,939	0,779	0,630	0,960	0,957	0,889	0,761	0,945	0,944
na	0,904	0,937	0,889	0,913	0,851	0,890	0,871	0,906	0,817	0,848	0,842	0,875
p38	0,497	0,496	0,471	0,494	0,445	0,467	0,454	0,508	0,404	0,423	0,500	0,521
parp	0,266	0,330	0,367	0,455	0,507	0,434	0,636	0,625	0,500	0,444	0,693	0,658
pde5	0,537	0,368	0,728	0,690	0,769	0,640	0,792	0,795	0,813	0,753	0,835	0,838
pdgfrb	0,277	0,243	0,745	0,651	0,438	0,328	0,768	0,732	0,533	0,410	0,783	0,747
pnp	0,658	0,615	0,781	0,748	0,749	0,639	0,804	0,783	0,762	0,699	0,839	0,828
ppar_gama	0,140	0,070	0,702	0,513	0,219	0,162	0,886	0,838	0,254	0,211	0,956	0,912
pr	0,524	0,606	0,594	0,565	0,576	0,479	0,824	0,847	0,697	0,519	0,845	0,836
rxr_alpha	0,887	0,822	0,832	0,856	0,933	0,924	0,891	0,884	0,944	0,937	0,913	0,892
sahh	0,807	0,809	0,678	0,692	0,940	0,886	0,950	0,919	0,939	0,906	0,951	0,944
src	0,286	0,280	0,446	0,455	0,517	0,357	0,645	0,566	0,692	0,542	0,746	0,702
thrombin	0,649	0,694	0,582	0,526	0,392	0,399	0,405	0,519	0,396	0,375	0,441	0,435
Tk	0,740	0,585	0,750	0,505	0,771	0,716	0,779	0,713	0,803	0,753	0,808	0,770
trypsin	0,882	0,767	0,910	0,831	0,913	0,787	0,880	0,815	0,787	0,758	0,835	0,786
vegfr2	0,414	0,420	0,584	0,600	0,471	0,457	0,616	0,653	0,477	0,470	0,643	0,653
Nº Melhor resultado	0	4	2	1	1	0	4	2	3	1	19	4
Média	0,576	0,541	0,668	0,617	0,651	0,599	0,734	0,709	0,669	0,630	0,753	0,732

Apêndice C

Formato de arquivo de saída do PharmaSite e do 3DPharma

##Data_set_information
#Actives_file = ../listasativos_scpdb_EC3numbers_3level/ativos_scpdb2013_5_1_1.txt
Size = 12

#Similarity_file = normal4P_tanimoto_scpdb2013/rankedfiles/1ftx_EPC_1_site.fgp.u.txt
Size = 6565

##Actives_matrix_information
#molecule #similarity #rank ##REPETE #active? #cluster
1ftx_EPC_1_site.fgp.u 1.000000e+00 1 1 YES -
1epv_DCS_1_site.fgp.u 7.983330e-01 2 1 YES -
1bd0_IN5_1_site.fgp.u 7.302227e-01 3 1 YES -
1xqk_PMH_2_site.fgp.u 3.951954e-01 4 1 YES -
3e6e_DCS_1_site.fgp.u 3.946285e-01 5 1 YES -
1vft_DCS_2_site.fgp.u 2.746162e-01 6 1 YES -
2rjh_DCS_3_site.fgp.u 2.668078e-01 7 1 YES -
1amu_AMP_2_site.fgp.u 1.854144e-01 1791 1 YES -
2w4i_VGA_2_site.fgp.u 1.665160e-01 2642 1 YES -
2ohv_NHL_1_site.fgp.u 1.560314e-01 3114 1 YES -
2jfz_003_2_site.fgp.u 9.664150e-02 5594 1 YES -
4b1f_KRH_2_site.fgp.u 7.058560e-02 6093 1 YES -
###

#ROC_AUC
0.756040490360649

#AUCpROC
4.56862988186012

#SLR
56.1911

#NSLR
0.6573

##PM scores
#PM_values #(TPR+FPR)
1.000000 0.001
1.000000 0.005
1.000000 0.010
1.000000 0.020
1.000000 0.050
1.000000 0.100
1.000000 0.200
1.000000 0.250
1.000000 0.500
0.978735 0.600
0.872410 0.700
0.792667 0.800
0.730644 0.900
0.690623 1.000

##EF and REF scores
#EF_values #REF_values #(Fraction_Dataset)
547.083333 100.000000 0.001
116.907932 58.453966 0.005
58.607296 58.607296 0.01
29.456979 58.913957 0.02
11.966788 59.833941 0.05
6.136725 61.367246 0.1
3.221693 64.433857 0.2
2.638686 65.967162 0.25
1.677991 83.899530 0.5

##ROCE
#ROCE_scores #Fraction_Decoys
583.639606 0.001
116.972939 0.005
58.639606 0.01
29.472939 0.02
11.972939 0.05
6.139606 0.1
3.222939 0.2
2.639606 0.25
1.678264 0.5

##BEDROC_Scores
#Alpha_Values #BEDROC_Scores
1 0.720161558821276 9.86978503225768 0.582508921458107 1.58053175948606
2 0.687644998035236 8.74354340478002 0.313608172799848 2.30881249329835
4 0.639755681794701 7.72618420247139 0.0749029341910008 4.05976985145288
8 0.599833783532708 7.14290162194917 0.00270432609888121 7.94445713570818
16 0.588139254205082 6.94699370440425 1.82715119880587e-06 15.7682980590262
32 0.590443841639707 6.86533125837093 4.17339838335778e-13 31.082111226715
50 0.594410248557717 6.79075242002324 1.00981763239743e-20 47.7832011774728

100 0.605419392496849 6.58928588905491 4.08176541872616e-42 91.3929352144731
250 0.637933343214065 6.02842485279103 8.45904274628459e-107 200.67010529097
500 0.689807807797688 5.22177573719847 5.8237022385863e-215 327.734438473811

##ROC_CURVE
#1-Specificity #Sensitivity
0.0000000000 0.0000000000
0.0000000000 0.5833333333
0.2720891195 0.5833333333
0.2720891195 0.6666666667
0.4018007020 0.6666666667
0.4018007020 0.7500000000
0.4736761788 0.7500000000
0.4736761788 0.8333333333
0.8519761941 0.8333333333
0.8519761941 0.9166666667
0.9279719213 0.9166666667
0.9279719213 1.0000000000
1.0000000000 1.0000000000
###

##Dataset_matrix_information
#molecule #similarity #rank #REPETE #active? #cluster
1ftx_EPC_1_site.fgp.u 1.000000e+00 1 1 YES -
1epv_DCS_1_site.fgp.u 7.983330e-01 2 1 YES -
1bd0_IN5_1_site.fgp.u 7.302227e-01 3 1 YES -
1xqk_PMH_2_site.fgp.u 3.951954e-01 4 1 YES -
3e6e_DCS_1_site.fgp.u 3.946285e-01 5 1 YES -
1vft_DCS_2_site.fgp.u 2.746162e-01 6 1 YES -
2rjh_DCS_3_site.fgp.u 2.668078e-01 7 1 YES -
2c29_DQH_2_site.fgp.u 2.626752e-01 8 1 NO -
4nk5_NAD_4_site.fgp.u 2.624329e-01 9 1 NO -
1f0x_FAD_2_site.fgp.u 2.607700e-01 10 1 NO -
2wyv_NAD_2_site.fgp.u 2.606070e-01 11 1 NO -
1guf_NDP_2_site.fgp.u 2.587171e-01 12 1 NO -
1hsk_FAD_1_site.fgp.u 2.550320e-01 13 1 NO -
3qgi_33F_1_site.fgp.u 2.542273e-01 14 1 NO -
4euf_NAD_1_site.fgp.u 2.533845e-01 15 1 NO -
2nnl_ERD_1_site.fgp.u 2.528345e-01 16 1 NO -
2j3k_NAP_2_site.fgp.u 2.518547e-01 17 1 NO -
4c77_N01_1_site.fgp.u 2.512118e-01 18 1 NO -
3jxe_TYM_2_site.fgp.u 2.508780e-01 19 1 NO -
1c0k_FAD_1_site.fgp.u 2.506997e-01 20 1 NO -
3vqs_JT1_3_site.fgp.u 2.492249e-01 21 1 NO -
2jbs_FMN_3_site.fgp.u 2.491193e-01 22 1 NO -
1eq2_ADQ_6_site.fgp.u 2.487253e-01 23 1 NO -
1udb_UFG_1_site.fgp.u 2.486261e-01 24 1 NO -
2gj3_FAD_2_site.fgp.u 2.481561e-01 25 1 NO -
1r6t_TYM_1_site.fgp.u 2.479084e-01 26 1 NO -
1udc_UFM_1_site.fgp.u 2.475344e-01 27 1 NO -
3qgh_63F_1_site.fgp.u 2.475151e-01 28 1 NO -
4bb5_HD2_3_site.fgp.u 2.468955e-01 29 1 NO -

1xu9_CPS_2_site.fgp.u 2.468544e-01 30 1 NO -
1lqu_FAD_1_site.fgp.u 2.467332e-01 31 1 NO -
2yy5_WSA_2_site.fgp.u 2.465617e-01 32 1 NO -
4o1m_NAD_6_site.fgp.u 2.461197e-01 33 1 NO -
...
1m66_BCP_1_site.fgp.u 2.589000e-04 6564 1 NO -
1m67_BOA_1_site.fgp.u 2.449000e-04 6565 1 NO -

Apêndice D

Artigos

Lopes *et al. J Cheminform (2017) 9:7*
DOI 10.1186/s13321-016-0189-4

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability

Julio Cesar Dias Lopes[1], Fábio Mendes dos Santos[1], Andrelly Martins-José[1], Koen Augustyns[2] and Hans De Winter[2*]

Abstract

A new metric for the evaluation of model performance in the field of virtual screening and quantitative structure–activity relationship applications is described. This metric has been termed the power metric and is defined as the fraction of the true positive rate divided by the sum of the true positive and false positive rates, for a given cutoff threshold. The performance of this metric is compared with alternative metrics such as the enrichment factor, the relative enrichment factor, the receiver operating curve enrichment factor, the correct classification rate, Matthews correlation coefficient and Cohen's kappa coefficient. The performance of this new metric is found to be quite robust with respect to variations in the applied cutoff threshold and ratio of the number of active compounds to the total number of compounds, and at the same time being sensitive to variations in model quality. It possesses the correct characteristics for its application in early-recognition virtual screening problems.

Keywords: Power metric (PM), Virtual screening, Metric, Model performance, Enrichment factor, Area under the curve (AUC), Receiver operating curve enrichment factor (ROCE), Correct classification rate (CCR), Matthews correlation coefficient (MCC), Cohen's kappa coefficient (CKC), Relative enrichment factor (REF)

Background

The field of virtual screening with applications in drug design has become increasingly important in terms of hit finding and lead generation [1–3]. Many different methods and descriptors have emerged over time to help the drug discovery scientist in applying the most optimal techniques for almost any given computational problem [4]. However, still a serious drawback in the domain of virtual screening is the lack of metrics standards to statistically evaluate and compare the performance of different methods and descriptors. Nicholls [5] suggested a few list of desirable characteristics of a good metric:

1. independence to extensive variables,
2. statistical robustness,
3. straightforward assessment of error bounds,
4. no free parameters,
5. easily understandable and interpretable.

In addition to these five characteristics, we believe that a good metric might also benefit from having well-defined lower and upper boundaries as this facilitates quantitative comparison of different models and facilitates optimization of fitness functions based on these metrics.

In this paper a new metric is proposed that adheres to the six desired characteristics of an ideal metric. The metric is based on the principles behind the power of hypothesis test, which is the probability of making the correct decision if the alternative hypothesis is true. Comparison of the new power metric with more established metrics, including the enrichment factor (EF) [6,

*Correspondence: hans.dewinter@uantwerpen.be
[2] Medicinal Chemistry Group, Department of Pharmaceutical Sciences, University of Antwerp, Campus Drie Eiken, Building A, Universiteitsplein 1, 2610 Wilrijk, Antwerp, Belgium
Full list of author information is available at the end of the article

7], the relative enrichment factor (REF) [8], the receiver operating characteristic (ROC) enrichment ROCE [9–11], the correct classification rate (CCR) [12, 13], Matthews correlation coefficient (MCC) [14], Cohen's kappa coefficient (CKC) [15, 16] together with the standard precision (PRE), accuracy (ACC), sensitivity (SEN) and specificity (SPE) metrics, is presented in this paper.

Methods

Definitions

In the field of virtual screening, the quality of a model can be quantified by a number of metrics. The area under the curve (AUC) represents the overall accuracy of a model, with a value approaching 1.0 indicating a high sensitivity and high specificity [17]. A model with an AUC of 0.5 represents a test with zero discrimination. AUC metrics are calculated from typical ROC curves; these are plots of the (1 − SPE) values on the *x*-axis against the SEN values plotted on the *y*-axis for all possible cutoff points. Sensitivity and specificity, and thus the AUC, are good indicators of the validity of a method but are not measuring the predictive value of a method [18].

The AUC is a metric that describes the overall quality of a model. In practical virtual screening experiments however, it is typical to score each molecule according to a value proposed by the model, and rank these molecules in decreasing order based on these calculated values. It is custom to define a cutoff threshold χ that separates predicted actives (all compounds along the 'top' side of this ranked list) from predicted non-actives (all compounds along the 'bottom' side of the ranked list) (see Fig. 1). The cutoff threshold χ is defined as the fraction of compounds selected:

$$\chi = N_s/N \tag{1}$$

with N_s being the number of compounds in the selection set (the predicted actives) and N being the total number of compounds in the entire dataset. The majority of metrics, including all metrics in this paper, are dependent on the value of this cutoff criterion χ since this criterion defines which compounds are predicted to be active and non-active.

Apart from the N_s and N variables, two other definitions are used in the following sections: the number of true active compounds in the selection set that is defined as n_s, and the number of true active compounds in the entire dataset defined as n. Finally, the prevalence of actives R_a in the entire dataset can be defined as:

$$R_a = n/N \tag{2}$$

Definition and calculation of established metrics

The sensitivity of a model is defined as the ability of the model to correctly identify active compounds from all the actives in the screening set (also termed the true positive rate or TPR), while specificity refers to the ability of the model to correctly identify inactives from all inactives in the dataset at a given cutoff threshold χ:

$$SEN(\chi) = TPR(\chi) = \frac{TP}{TP + FN} = \frac{n_s}{n} \tag{3}$$

$$SPE(\chi) = \frac{TN}{FP + TN} = \frac{N - N_s - n + n_s}{N - n} \tag{4}$$

In line with the true positive rate, one can also define a false positive rate FPR as the number of true inactives in the selection set in relation to the total number of inactives in the entire dataset:

$$FPR(\chi) = \frac{FP}{FP + TN} = \frac{N_s - n_s}{N - n} \tag{5}$$

Other well-established metrics include the precision and accuracy:

$$PRE(\chi) = \frac{TP}{TP + FP} = \frac{n_s}{N_s} \tag{6}$$

$$ACC(\chi) = \frac{TP + TN}{TP + TN + FP + FN} = \frac{2n_s + N - N_s - n}{N} \tag{7}$$

The enrichment factor is probably the most used metric in virtual screening and other fields as well. The EF at a given cutoff χ is calculated from the proportion of true active compounds in the selection set in relation to the proportion of true active compounds in the entire dataset:

$$EF(\chi) = \frac{TP/TP + FP}{TP + FN/TP + TN + FP + FN} = \frac{N \times n_s}{n \times N_s} \tag{8}$$

The enrichment factor is very intuitive and easy to understand, but it lacks a strong statistic background and has some drawbacks, including the lack of a well-defined upper boundary [the EF(χ) can vary from 0 in the case that there are no active compounds in the selection set ($n_s = 0$), and up to $1/\chi$ when all active compounds are located in the selection set ($n_s = n$); see Ref. [19] for the derivation], the dependency of the value on the ratio of active to inactive compounds in the dataset, and a pronounced 'saturation effect' when the actives saturate the early positions of the ranking list and the performance metric cannot get any higher, thereby preventing to distinguish between good and excellent models [6].

To avoid the problems associated to EF, a number of other metrics have been proposed. The first of these is the relative enrichment factor [8], a metric in which the problem associated to the saturation effect is fixed by

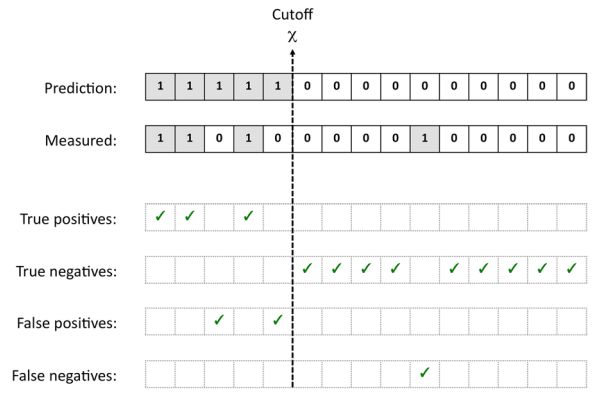


Fig. 1 Illustration of the relation between cutoff χ and number of predicted actives/non-actives. Assuming a list of compounds ranked according to their predicted activity values, all compounds that are located on the *left side* of χ on this ranked list are predicted to be active, while all compounds that are located on the *right-hand side* of χ on this list are predicted to be non-active. All compounds that fall along the *left-hand side* of χ define the 'selection set'; in this example this includes five compounds. The total number of compounds in the selection set is N_s (here: 5), while the total number of compounds in the entire collection is N (here: 15). The number of true actives in the selection set is n_s (here: 3) and the number of true actives in the entire data collection is n (here: 4). Using these abbreviations, one can define the number of true positives TP as being equal to n_s, the number of true negatives TN equal to $(N - N_s - n + n_s)$, the number of false positives FP equal to $(N_s - n_s)$, and the number of false negatives FN being equal to $(n - n_s)$

considering the maximum EF achievable at the cutoff point:

$$REF(\chi) = \frac{100 \times n_s}{\min(N \times \chi, n)} \tag{9}$$

The REF, has well defined boundaries—ranging from 0 to 100—and is less subject to the saturation effect.

The ROC enrichment metric is defined as the fraction of actives found when a given fraction of inactives has been found [9]:

$$ROCE(\chi) = \frac{n_s/n}{(N_s - n_s)/(N - n)} = \frac{n_s \times (N - n)}{n \times (N_s - n_s)} \tag{10}$$

The ROCE metric has been advocated by some researches as a better approach to address early recovery [5, 9]. However, some issues still remain, such as the lack of a well-defined upper boundary [which is equal to $1/\chi$ when TPR(χ) equals 1], a smaller but still noticeable saturation effect, and a statistic robustness which is not as desirable as we will demonstrate later.

Another metric often considered to measure classification performances is the correct classification rate [12], defined as the percentage of instances correctly classified:

$$\begin{aligned} CCR(\chi) &= \frac{1}{2}\left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right] \\ &= \frac{1}{2}\left[\frac{n_s}{n} + \frac{N - N_s - n + n_s}{N - n}\right] \end{aligned} \tag{11}$$

The CCR is sometimes also called the balanced accuracy [20].

Matthews correlation coefficient has been advocated as a balanced measure that can be used on classes of different sizes [14]. The MCC is in essence a correlation coefficient between the measured and predicted classifications; it returns a coefficient of +1 in the case of a perfect prediction, 0 when no better than random prediction and −1 in cases of total disagreement between prediction and observation:

$$\begin{aligned} MCC(\chi) &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ &= \frac{N \times n_s - N_s \times n}{\sqrt{N_s \times n \times (N - n) \times (N - N_s)}} \end{aligned} \tag{12}$$

The last metric that is evaluated with respect to its performance as compared to the here developed power metric is Cohen's kappa coefficient [21–24]:

$$\begin{aligned} CKC(\chi) &= 1 - \frac{1 - \frac{TP+TN}{TP+TN+FP+FN}}{1 - \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+TN+FP+FN^2)}} \\ &= 1 - \frac{N \times n + N \times N_s - 2 \times n_s \times N}{N \times n + N \times N_s - 2 \times n \times N_s} \end{aligned} \tag{13}$$

Derivation of a new metric: the power metric

In virtual screening studies, we can assume all compounds being inactive as the null hypothesis, and the assumption that some compounds are active as the alternative hypothesis. The statistical power, also known as sensitivity or recall, is equal to the true positive rate.

However, the statistical power alone does not include information about the distribution of negative instances or the size effect. Therefore, a metric based on statistical power and suited for applications in the field of virtual screening should incorporate information about the negative instances as well. Ideally, a good virtual screening method must be able to perform a good prediction of true positive instances combined with a small false positive prediction rate. This translates in a metric that combines the TPR with the false positive rate:

$$'net\ power'\ (\chi) = TPR(\chi) - FPR(\chi) \tag{14}$$

Graphically, the 'net power' is the area of the distribution of positive instances or the alternative hypothesis, minus the area of the distribution of negative instances or the null hypothesis (Fig. 2).

The metric is not new; it has been developed independently several times in the past. Its origin can be traced

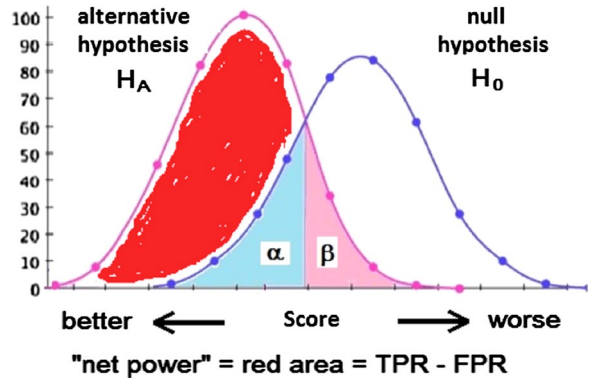


Fig. 2 Distribution curves of null and alternative hypothesis. The *red area* is the area defined by the alternative hypothesis minus the area defined by the null hypothesis, or, put differently, as the true positive rate (TPR, or $1-\beta$) minus the false positive rate (FPR, or α or type I error), hence 'net power' $= 1-\beta-\alpha$. The cutoff point is defined as the crossing point of the two distributions

back to the seminal paper of Peirce [25] with his 'science of the method' [26]. More than 70 years later, it was proposed again by Youden as Youden's index ($Y'J$) [27]. Youden's index is often used in conjunction with the ROC curve as a criterion for selecting the optimum cutoff point [28]. The index has been used to calculate the best cutoff point in the ROC curve. Once more, almost 50 years later in 2003, it was proposed again by Powers who called it 'informedness' [10].

Despite the success of this metric to evaluate the prediction power of a method, it is not entirely appropriate for virtual screening studies due to the lack of early recovery capabilities that are very desirable in any virtual screening application. Consider, for instance, a database of 10,000 compounds of which 1% are active compounds. In this hypothetical thought experiment, we can think of different methods that yield identical Youden's indices calculated from different TPR and FPR values. Thinking of two methods, each produce a Youden's index of 0.5, with the first one characterized by a TPR = 0.9 and a FPR = 0.4, and the second method characterized by a TPR = 0.51 and a FPR = 0.01. In the case of the first method, 4050 compounds will be marked as 'hits' of which only 90 compounds being true active (or 5.7% of the selected compounds). However, for the second method only 150 compounds are flagged as 'hits', of which 51 compounds are true actives (or 34% of the selected compounds). Obviously, for virtual screening applications, the second method provides a more optimal early recovery rate since only 1.5% of the original dataset needs to be tested in order to recover 51% of all active compounds.

Normalization of the 'net power' metric by dividing by the sum of the true positive and false negative rates introduces early recovery capabilities bias into the 'net power' metric. This difference-over-the-sum normalized 'net power' expresses the dominance of the true positive rate over the false positive rate among those instances predict as positive, expressed by its rates:

$$normalized\ 'netpower' = \frac{TPR(\chi) - FPR(\chi)}{TPR(\chi) + FPR(\chi)} \tag{15}$$

The metric ranges from -1 to $+1$ and can easily be modified to range from 0 to $+1$ by adding 1 to the metric and dividing by 2. We call this new metric the power metric (PM) and is defined as follows:

$$\begin{aligned} PM(\chi) &= \frac{\left(\frac{TPR(\chi)-FPR(\chi)}{TPR(\chi)+FPR(\chi)} + 1\right)}{2} \\ &= \frac{TPR(\chi)}{TPR(\chi) + FPR(\chi)} = \frac{n_s \times N - n \times n_s}{n_s \times N - 2 \times n \times n_s + n \times N_s} \end{aligned} \tag{16}$$

Probability distribution function to evaluate the metrics

In order to evaluate the performance of several metrics used in the field of virtual screening, we used the probability distribution function approach as suggested by Truchon and Bayly to build hypothetical models of different qualities [6]. For a typical virtual screening study with N compounds of which n being active compounds, we generated the ranks of these active compounds according to the exponential distribution as proposed by Truchon and Bayly [6]:

$$X_i = \frac{-1}{\lambda} ln\left(1 - U_i\left(1 - e^{-\lambda}\right)\right) \tag{17}$$

The generated real number X_i corresponds to the relative position of active compound i and U_i is a pseudo random number with values between 0 and 1. In this exponential distribution, the λ parameter represents the model quality (lower λ values correspond to poor models and larger λ values correspond to better models). The number X_i is transformed into a rank integer r_i that falls within 1 and N:

$$r_i = int(N \times X_i + 0.5) \tag{18}$$

No ties were allowed and each active compound occupies one unique position. In cases when a clash occurred, a new random number was generated. In our simulations we used values of λ equal to 1, 2, 5, 10, 20 and 40. Visualization of the quality of these models is given in Fig. 3.

To illustrate the model generation process by example, consider a model with quality $\lambda = 20$ and consisting of $n = 100$ active compounds on a total of $N = 10{,}000$

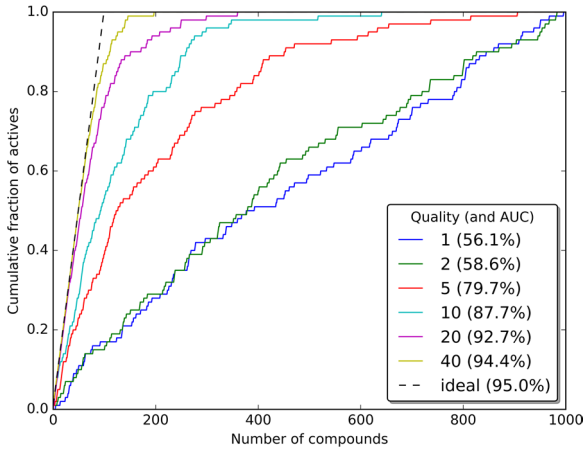


Fig. 3 Visualization of the quality of the test models. In each case, 100 active compounds were distributed in a total set of 1000 compounds according to the distribution as defined by Eq. 17. The term 'quality' corresponds to the λ value of Eq. 17. The 'ideal' case was generated by positioning all 100 actives at the top-100 positions of the dataset. For each model, the AUC was calculated by integration using the composite trapezoidal rule

compounds. To generate the relative rankings of these 100 active compounds, Eq. 17 was called 100 times, each time with a different random number U_i. Using Eq. 18, the 100 generated X_i numbers are then converted into 100 rankings r_i with N set to 10,000. These 100 rankings are the absolute positions of the active compounds; the remaining 9900 ranks ($10{,}000 - 100 = 9900$) are those of the inactive compounds.

In order to evaluate the quality of the PM metric and to compare its behavior to the other metrics, a large number of datasets were generated and analyzed. The total number of compounds N, number of actives n, model quality λ and cutoff parameter χ were varied. Each simulation was repeated 10,000 times and the results were analyzed by inspecting the variations of mean and standard deviation (STD) of the metrics as a function of the number of actives and total compounds. The eleven enrichment-type metrics that were analyzed were the PM, EF, ROCE, CCR, REF, MCC, CKC, together with the standard PRE, ACC, SEN and SPE metrics.

All calculations were performed under Python 2.7 using Numpy and Scipy [29]. The IPython notebook [30] was used as programming environment and figures were generated with Matplotlib [31]. MarvinSketch was used for drawing chemical structures [32].

Results and discussion

Dependency on model quality

One of the key aspects of a suitable metric is that its value is dependent of the model quality. In Table 1, the dependency of the different metrics on the model quality parameter λ was evaluated. All metrics are model quality dependent, but the ROCE, EF, REF, MCC, CKC, SEN and PRE show an approximate tenfold increase when moving from a poor model with quality $\lambda = 2$ to a good model with quality $\lambda = 40$, while in the case of the PM metric a doubling of the parameter value is observed (going from PM = 0.5 for a poor model to a value of 0.98 for a good model; Table 1). Accuracy and specificity metrics are not influenced by the model quality λ or by the cutoff value χ; both metrics fluctuate around a value of 0.97-1.00 irrespective of the underlying model quality or applied threshold cutoff. In the case of the CCR metric, the maximal value of this metric finds it limit at 0.75 ± 0.02 for the case with an extremely good model quality of $\lambda = 40$ in combination with a threshold cutoff χ of 2% (for a model with 100 actives on a total of 10,000 compounds, a model quality of $\lambda = 40$ corresponds to an AUC of 97.25%, as compared to an AUC of 99.5% for the ideal case). This is not what one would like to expect for a metric to separate quality models from poor models. Furthermore, the PM metric seems to be less influenced by the applied cutoff parameter χ, since the PM metric for a good model ($\lambda = 40$) at the different cutoffs of 0.5, 1 and 2% remains largely unchanged (at a constant value of approximately 0.98; see Table 1), while an increase is seen for the CCR metric. It seems that all but the PM, SPE and ACC metrics are more dependent on the applied cutoff threshold χ (indicated by the shifts in the values and by the larger variations on the calculated metrics; Table 1), making it more difficult to define an appropriate metric value for identification proper virtual screening models. Starting with models of reasonable quality, and up to models of higher qualities ($\lambda \geq 10$), the PM is calculated to vary between 0.9 and 1.0 with a relative standard deviation less than 10%. For the other metrics (except the CCR, ACC and SPE metrics), this relative standard deviation is in most instances larger than 10%.

Dependency on the ratio of actives to total number of compounds

The influence of the R_a value, calculated from the ratio of number of actives n to the total number of compounds N, on the different metrics is given in Table 2. For the different model qualities (a poor model with $\lambda = 1$ or a good model with $\lambda = 20$) and different cutoff values ($\chi = 1$ or 10%), there is a significant dependency for the REF, PRE and ACC metrics on the R_a value. The EF, CKC, SEN and ROCE metrics are not very sensitive to the R_a value when applied to poor models ($\lambda = 1$), but show more dependency on the R_a ratio when applied on good models ($\lambda = 20$). In contrast, the REF is very sensitive to the R_a value when used on poor models ($\lambda = 1$), but is not

Table 1 Dependency on the model quality parameter λ using models generated from datasets with 100 actives (*n*) on 10,000 compounds in total (*N*)

Metric	λ					χ (%)
	2	5	10	20	40	
PM	0.51 ± 0.35	0.74 ± 0.24	0.89 ± 0.09	0.95 ± 0.02	0.98 ± 0.01	0.5
ROCE	2.35 ± 2.18	5.13 ± 3.39	10.46 ± 4.99	22.34 ± 7.86	49.96 ± 14.35	
EF	2.28 ± 2.06	4.83 ± 3.03	9.38 ± 4.04	18.08 ± 5.17	32.94 ± 6.22	
REF	2.28 ± 2.06	4.83 ± 3.03	9.38 ± 4.04	18.08 ± 5.17	32.94 ± 6.22	
CCR	0.50 ± 0.01	0.51 ± 0.01	0.52 ± 0.01	0.54 ± 0.01	0.58 ± 0.02	
MCC	0.01 ± 0.01	0.03 ± 0.02	0.06 ± 0.03	0.12 ± 0.04	0.23 ± 0.04	
CKC	0.01 ± 0.01	0.03 ± 0.02	0.06 ± 0.03	0.11 ± 0.03	0.21 ± 0.04	
SEN	0.01 ± 0.01	0.02 ± 0.02	0.05 ± 0.02	0.09 ± 0.03	0.16 ± 0.03	
SPE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
PRE	0.02 ± 0.02	0.05 ± 0.03	0.09 ± 0.04	0.18 ± 0.05	0.33 ± 0.06	
ACC	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	
PM	0.61 ± 0.23	0.80 ± 0.11	0.90 ± 0.04	0.95 ± 0.01	0.98 ± 0.00	1
ROCE	2.32 ± 1.55	5.07 ± 2.31	10.19 ± 3.34	20.97 ± 5.08	44.00 ± 8.22	
EF	2.26 ± 1.48	4.83 ± 2.09	9.25 ± 2.75	17.33 ± 3.46	30.54 ± 3.95	
REF	2.26 ± 1.48	4.83 ± 2.09	9.25 ± 2.75	17.33 ± 3.46	30.54 ± 3.95	
CCR	0.51 ± 0.01	0.52 ± 0.01	0.54 ± 0.01	0.58 ± 0.02	0.65 ± 0.02	
MCC	0.01 ± 0.02	0.04 ± 0.02	0.08 ± 0.03	0.17 ± 0.03	0.30 ± 0.04	
CKC	0.01 ± 0.02	0.04 ± 0.02	0.08 ± 0.03	0.17 ± 0.03	0.30 ± 0.04	
SEN	0.02 ± 0.01	0.05 ± 0.02	0.09 ± 0.03	0.17 ± 0.03	0.31 ± 0.04	
SPE	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	
PRE	0.02 ± 0.01	0.05 ± 0.02	0.09 ± 0.03	0.17 ± 0.03	0.31 ± 0.04	
ACC	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	
PM	0.66 ± 0.13	0.82 ± 0.06	0.90 ± 0.02	0.95 ± 0.01	0.97 ± 0.00	2
ROCE	2.30 ± 1.08	4.91 ± 1.56	9.69 ± 2.18	18.75 ± 3.08	35.21 ± 4.06	
EF	2.26 ± 1.03	4.70 ± 1.43	8.88 ± 1.82	15.87 ± 2.19	26.17 ± 2.23	
REF	4.52 ± 2.07	9.40 ± 2.85	17.76 ± 3.65	31.74 ± 4.38	52.34 ± 4.45	
CCR	0.51 ± 0.01	0.54 ± 0.01	0.58 ± 0.02	0.65 ± 0.02	0.75 ± 0.02	
MCC	0.02 ± 0.01	0.05 ± 0.02	0.11 ± 0.03	0.21 ± 0.03	0.36 ± 0.03	
CKC	0.02 ± 0.01	0.05 ± 0.02	0.11 ± 0.02	0.20 ± 0.03	0.34 ± 0.03	
SEN	0.05 ± 0.02	0.09 ± 0.03	0.18 ± 0.04	0.32 ± 0.04	0.52 ± 0.04	
SPE	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	
PRE	0.02 ± 0.01	0.05 ± 0.01	0.09 ± 0.02	0.16 ± 0.02	0.26 ± 0.02	
ACC	0.97 ± 0.00	0.97 ± 0.00	0.97 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	

Metric abbreviations are given in the Methods section. All metrics are dependent on the model quality, but in case of the ROCE, EF, REF, MCC, CKC, SEN and PRE metrics there is at least a tenfold increase when moving from a bad model ($\lambda = 2$) to a good model ($\lambda = 40$), while for the PM metric there is a doubling of the value. The accuracy ACC and specificity SPE metrics are not dependent on the quality of model, while the correct classification rate metric (CCR) shifts from 0.5 in the case of a bad model to a maximum of 0.75 for the best model. Good models have a PM of >0.9; for good models this value is largely independent on the applied cutoff value χ (see Table 3 as well)

dependent on the R_a value when applied on a good model in combination with a large cutoff value ($\chi = 1\%$; Table 2). In contrast, the PM and CCR metrics remain largely insensitive to the R_a value, unless when the PM metric it is applied to a very poor model ($\lambda = 1$) in combination with a small cutoff threshold value ($\chi = 1\%$). Again, good models all have PM values ≥ 0.9 with small variations, and are independent on the number of actives in relation to the total number of compounds. The combination of a high model quality of $\lambda = 20$ with a cutoff threshold of $\chi = 1\%$, applied to a database with $n = 50$ actives on a total of $N = 5000$ compounds, corresponds to a virtual screening situation characterized by a high true positive and high true negative rate. It is therefore surprising that for the CCR metric a value of 0.58 ± 0.02 is calculated, while for the PM metric a more intuitive value of 0.95 ± 0.02 is found (Table 2). Increasing the cutoff threshold to 10% improves the calculated CCR value to

Table 2 Dependency on the R_a value

Metric	R_a			χ (%)	λ
	0.01 ($n = 50$; $N = 5000$)	**0.05 ($n = 250$; $N = 5000$)**	**0.2 ($n = 1000$; $N = 5000$)**		
PM	0.39 ± 0.36	0.57 ± 0.15	0.62 ± 0.07	1	1
ROCE	1.59 ± 1.83	1.62 ± 0.85	1.73 ± 0.54		
EF	1.55 ± 1.75	1.54 ± 0.74	1.48 ± 0.32		
REF	1.55 ± 1.75	7.69 ± 3.71	29.58 ± 6.38		
CCR	0.50 ± 0.01	0.50 ± 0.00	0.50 ± 0.00		
MCC	0.01 ± 0.02	0.01 ± 0.02	0.02 ± 0.02		
CKC	0.01 ± 0.02	0.01 ± 0.01	0.01 ± 0.01		
SEN	0.02 ± 0.02	0.02 ± 0.01	0.01 ± 0.00		
SPE	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00		
PRE	0.02 ± 0.02	0.08 ± 0.04	0.30 ± 0.06		
ACC	0.98 ± 0.00	0.94 ± 0.00	0.80 ± 0.00		
PM	0.58 ± 0.09	0.60 ± 0.04	0.62 ± 0.02	10	
ROCE	1.50 ± 0.51	1.53 ± 0.24	1.62 ± 0.15		
EF	1.49 ± 0.49	1.49 ± 0.22	1.44 ± 0.09		
REF	14.88 ± 4.95	14.88 ± 2.16	28.73 ± 1.87		
CCR	0.52 ± 0.03	0.53 ± 0.01	0.53 ± 0.01		
MCC	0.02 ± 0.02	0.04 ± 0.02	0.07 ± 0.02		
CKC	0.01 ± 0.01	0.03 ± 0.02	0.07 ± 0.01		
SEN	0.15 ± 0.05	0.15 ± 0.02	0.14 ± 0.01		
SPE	0.90 ± 0.00	0.90 ± 0.00	0.91 ± 0.00		
PRE	0.01 ± 0.00	0.07 ± 0.01	0.29 ± 0.02		
ACC	0.89 ± 0.00	0.86 ± 0.00	0.76 ± 0.00		
PM	0.95 ± 0.02	0.98 ± 0.01	1.00 ± 0.00	1	20
ROCE	21.06 ± 7.29	46.82 ± 15.58	nan[a]		
EF	17.24 ± 4.92	13.94 ± 1.27	5.00 ± 0.00		
REF	17.24 ± 4.92	69.71 ± 6.35	100.00 ± 0.00		
CCR	0.58 ± 0.02	0.57 ± 0.01	0.53 ± 0.00		
MCC	0.16 ± 0.05	0.30 ± 0.03	0.20 ± 0.00		
CKC	0.16 ± 0.05	0.22 ± 0.02	0.08 ± 0.00		
SEN	0.17 ± 0.05	0.14 ± 0.01	0.05 ± 0.00		
SPE	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00		
PRE	0.17 ± 0.05	0.70 ± 0.06	1.00 ± 0.00		
ACC	0.98 ± 0.00	0.95 ± 0.00	0.81 ± 0.00		
PM	0.90 ± 0.01	0.93 ± 0.00	1.00 ± 0.00	10	
ROCE	9.30 ± 0.57	13.38 ± 0.59	1612.74 ± 529.71		
EF	8.58 ± 0.49	8.26 ± 0.22	4.99 ± 0.01		
REF	85.82 ± 4.86	82.60 ± 2.15	99.84 ± 0.18		
CCR	0.88 ± 0.02	0.88 ± 0.01	0.75 ± 0.00		
MCC	0.25 ± 0.02	0.56 ± 0.02	0.67 ± 0.00		
CKC	0.14 ± 0.01	0.52 ± 0.02	0.61 ± 0.00		
SEN	0.86 ± 0.05	0.83 ± 0.02	0.50 ± 0.00		
SPE	0.91 ± 0.00	0.94 ± 0.00	1.00 ± 0.00		
PRE	0.09 ± 0.00	0.41 ± 0.01	1.00 ± 0.00		
ACC	0.91 ± 0.00	0.93 ± 0.00	0.90 ± 0.00		

In the case of bad model quality ($\lambda = 1$), the metrics most sensitive to variations in the R_a value include the REF, PRE and ACC metrics, and also the CKC metric in the case of a large cutoff value of $\chi = 10\%$. This dependency is not so outspoken for the PM metric, except in the case when a very bad model is combined with a low cutoff value ($\chi = 1\%$). In cases with better model quality ($\lambda = 20$), significant dependencies are observed for the ROCE, EF, REF, MCC, CKC, SEN, PRE and ACC metrics, while the PM, CCR and SPE metrics are more stable. The metric that is least sensitive to variations in the R_a value, irrespective of the underlying model quality or cutoff threshold, is the CCR metric

[a] In this case the ROCE metric could not be calculated from Eq. 10 since $(N_s - n_s)$ is equal to 0

0.88 ± 0.02 and decreases the PM case from 0.95 ± 0.02 to 0.90 ± 0.01, again in line what one would expect from considering the true positive and true negative rates in this situation.

Dependency on the cutoff threshold χ

The dependency of the different metrics on the applied cutoff value χ is given in Table 3. This dependency was evaluated using models with $n = 250$ active compounds in a dataset of $N = 10{,}000$ compounds in total, and at five different cutoff values χ (0.5, 1, 2.5, 5 and 10%) for both a poor and high quality model ($\lambda = 1$ and 20, respectively). A significant dependency on the cutoff χ is observed for the REF and SEN metrics, increasing their values with increasing cutoff values. A similar behavior is observed for the CCR, MCC and CKC metrics when applied to the high quality model situation ($\lambda = 20$). Interestingly, the calculated REF metric values remain constant up to a cutoff of 2.5%, but at higher cutoff values this metric increases significantly. It is not surprising that this turning point in metric behavior is observed at a cutoff value of 2.5%, since this corresponds to a selection set of exactly 250 compounds when applied to a dataset of 10,000 compounds with 250 actives mixed into it. In case of a high quality model, this translates to a situation with maximum rates of true positives and true negatives. Focusing on the EF, ROCE, CCR, SPE, ACC and PM metrics, their values are quite constant over the different cutoff values in the case of a bad model quality, but a significant drift is observed for the EF, CCR and ROCE metrics in case of a good model quality. This shift is again observed at a χ cutoff value larger than 2.5%. A similar drift is not observed for the PM metric that, together with the CCR metric, also has the smallest relative standard deviations (Table 3).

Dependency on both model quality λ and cutoff threshold χ

A direct comparison of the variation of the values of the five most commonly used metrics (CCR, ROCE, MCC, REF and CKC) with those of the PM, as a function of both model quality λ and cutoff threshold χ, is provided in Fig. 4. Comparing the results of the PM and CCR metrics, both types of metric values increase with increasing model quality, but the PM metric seems to be less dependent on the applied cutoff threshold as compared

Table 3 Dependency on the χ cutoff value using models generated from datasets with 250 actives (*n*) on 10,000 compounds in total (*N*)

Metric	χ					λ
	0.5%	1%	2.5%	5%	10%	
PM	0.52 ± 0.25	0.57 ± 0.15	0.60 ± 0.08	0.60 ± 0.06	0.60 ± 0.04	1
ROCE	1.60 ± 1.19	1.59 ± 0.81	1.58 ± 0.51	1.55 ± 0.35	1.52 ± 0.23	
EF	1.54 ± 1.10	1.56 ± 0.76	1.55 ± 0.48	1.53 ± 0.33	1.50 ± 0.22	
REF	3.86 ± 2.75	3.89 ± 1.90	3.88 ± 1.20	7.63 ± 1.65	14.97 ± 2.22	
CCR	0.50 ± 0.00	0.50 ± 0.00	0.51 ± 0.01	0.51 ± 0.01	0.53 ± 0.01	
MCC	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	
CKC	0.00 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	
SEN	0.01 ± 0.01	0.02 ± 0.01	0.04 ± 0.01	0.08 ± 0.02	0.15 ± 0.02	
SPE	1.00 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.95 ± 0.00	0.90 ± 0.00	
PRE	0.04 ± 0.03	0.04 ± 0.02	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.01	
ACC	0.97 ± 0.00	0.97 ± 0.00	0.95 ± 0.00	0.93 ± 0.00	0.88 ± 0.00	
PM	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.00	0.94 ± 0.00	0.91 ± 0.00	20
ROCE	28.80 ± 8.24	26.73 ± 5.20	22.13 ± 2.46	16.72 ± 1.11	10.49 ± 0.34	
EF	16.67 ± 2.70	16.12 ± 1.85	14.44 ± 1.03	11.99 ± 0.56	8.48 ± 0.22	
REF	41.68 ± 6.74	40.30 ± 4.62	36.09 ± 2.56	59.97 ± 2.79	84.78 ± 2.18	
CCR	0.54 ± 0.01	0.58 ± 0.01	0.67 ± 0.01	0.78 ± 0.01	0.88 ± 0.01	
MCC	0.18 ± 0.03	0.24 ± 0.03	0.34 ± 0.03	0.40 ± 0.02	0.40 ± 0.01	
CKC	0.13 ± 0.02	0.22 ± 0.03	0.34 ± 0.03	0.38 ± 0.02	0.31 ± 0.01	
SEN	0.08 ± 0.01	0.16 ± 0.02	0.36 ± 0.03	0.60 ± 0.03	0.85 ± 0.02	
SPE	1.00 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.96 ± 0.00	0.92 ± 0.00	
PRE	0.42 ± 0.07	0.40 ± 0.05	0.36 ± 0.03	0.30 ± 0.01	0.21 ± 0.01	
ACC	0.97 ± 0.00	0.97 ± 0.00	0.97 ± 0.00	0.96 ± 0.00	0.92 ± 0.00	

The PM is not so much dependent on the applied cutoff value. For good models the EF and ROCE metrics decrease when the cutoff is increased, while the REF, CCR, MCC and CKC values always increase when the cutoff is increased from 2.5% up to 10%

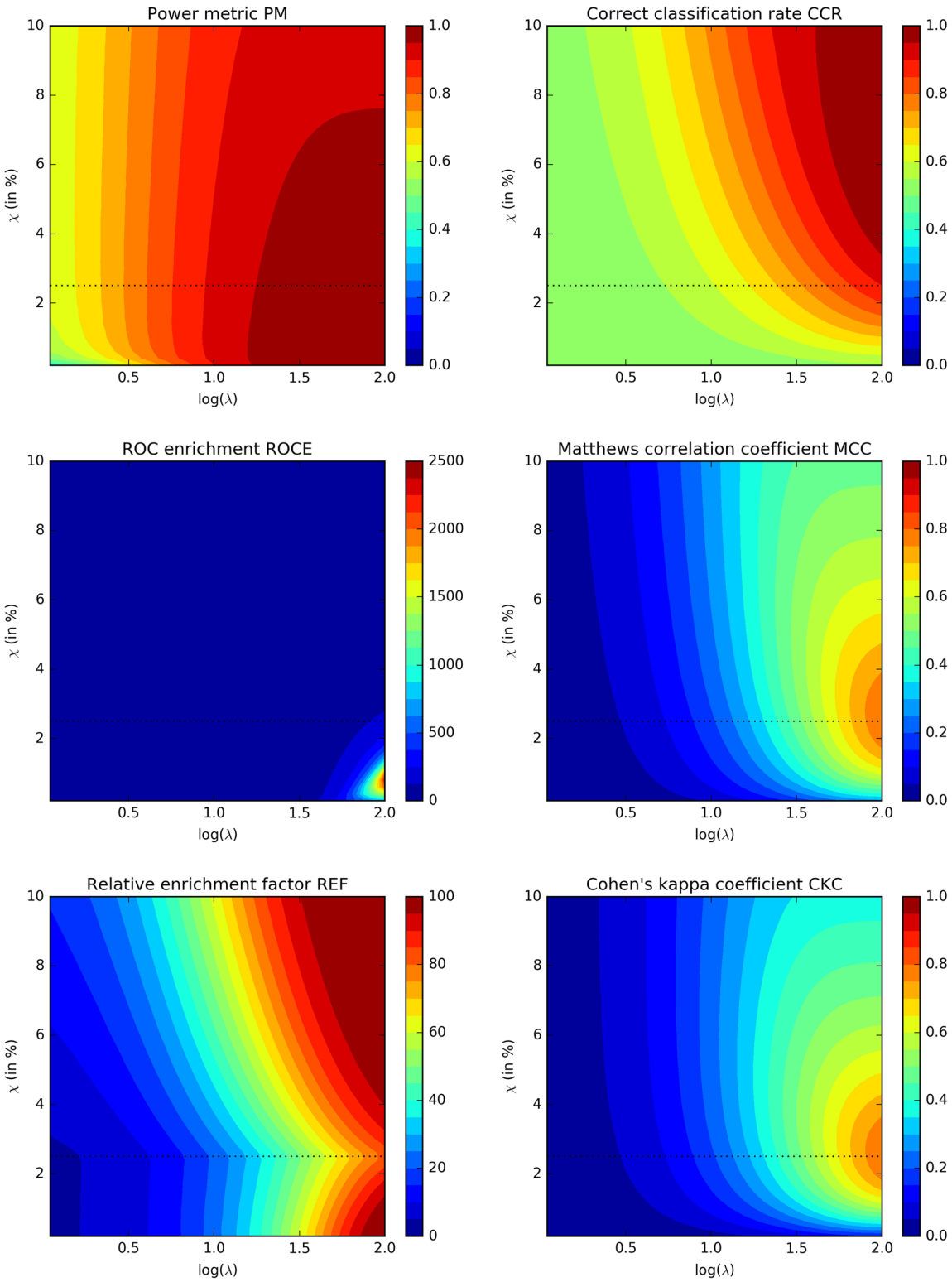


Fig. 4 Comparison of the power metric with the five main other metrics (CCR, ROCE, MCC, REF and CKC) using a model dataset of 250 active compounds on a total number of 10,000. The logarithm of the quality parameter λ is varied along the abscissa [a $\log(\lambda)$ of 2 corresponds to a quality λ of 100] while the applied cutoff threshold χ is varied along the ordinates. The *black dotted line* at a cutoff value χ of 2.5% indicates the boundary of 250 compounds on a total of 10,000. In a perfect model, all 250 active compounds would be located along the topside of this boundary

to the CCR metric (in fact, the CCR metric value is increasing with increasing cutoff thresholds, while the opposite behavior is observed in the case of the PM metric). The CCR metric is finding its highest values at larger cutoff thresholds in combination with high model qualities, making it less suitable for early-recognition problems. A similar conclusion can be drawn for the MCC and CKC metrics, as in both cases maximum values are obtained near a cutoff threshold χ that is equal or close to the fraction of true actives within the entire dataset (in the example of Fig. 4, this is 2.5%). Focusing on the ROCE metric, maximum values are calculated when models of high qualities are combined with cutoff thresholds χ that are smaller than 2.5%, *in casu* the fraction of true actives within the entire dataset of compounds. At very low cutoff thresholds, the ROCE metric decreases again. A main disadvantage of the ROCE metric is the lack of a well-defined upper boundary, hence making it difficult to compare the quality of underlying models and applied cutoff thresholds. Finally, the REF metric is not a continuous function but shows a discontinuity in its metric value along a threshold cutoff value of 2.5%, a value that is equal to the fraction of true actives in the dataset. At this cutoff threshold value and for all model qualities, a minimum in metric value is observed, which makes that for any given model quality under consideration two maxima are found: a first optimum at a cutoff threshold smaller than the 2.5%, and a second optimum that is located at a cutoff threshold χ much larger than the 2.5%.

Based on these observations, it can be concluded that the CCR, MCC and CKC metrics are all less suitable for early-recognition problems; for these problems the PM and ROCE metrics are better suited. The REF metric might also be an option to some extend but some cautions are warranted when used in combination with cutoff thresholds χ that are equal or larger than the fraction of true actives in the entire dataset. In these cases an increase in the REF metric is observed, which makes it less suitable for early-recognition problems. As already mentioned, the main disadvantage of the ROCE metric is the lack of a well-defined upper boundary, and for this reason the PM metric seems to posses powerful early-recognition properties and might be one of the preferred metrics for evaluating virtual screening models.

Conclusions

The power metric PM as described in this paper is a statistically solid metric with little sensitivity to the ratio of actives to the total number of compounds (the R_a value; see Table 2) and little sensitivity to the cutoff threshold parameter χ (Table 3). The metric is dependent on the underlying model quality, in such sense PM values around 0.5 are calculated for poor to random models, and values between 0.9 and 1.0 for high quality models. It is statistically robust in the sense that the calculated standard deviations are small and largely insensitive to the applied threshold cutoff value χ.

Abbreviations

ACC: accuracy; AUC: area under the curve; CCR: correct classification rate; CKC: Cohen's Kappa coefficient; EF: enrichment factor; MCC: Matthews correlation coefficient; PM: power metric; PRE: precision; QSAR: quantitative structure–activity relationship; REF: relative enrichment factor; ROC: receiver operating characteristic; ROCE: ROC enrichment; SEN: sensitivity; SPE: specificity; STD: standard deviation; TNR: true negative rate; TPR: true positive rate.

Authors' contributions

JCDL and FMDS: original idea, manuscript writing and programming; AMJ: original idea; HDW: programming, interpretations and writing of the manuscript. KA: general supervision. All authors have read and approved the final manuscript.

Author details

[1] NEQUIM - Chemoinformatics Group, Departamento de Quimica, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2] Medicinal Chemistry Group, Department of Pharmaceutical Sciences, University of Antwerp, Campus Drie Eiken, Building A, Universiteitsplein 1, 2610 Wilrijk, Antwerp, Belgium.

Competing interests

The authors declare that they have no competing interests.

Funding

Julio Cesar Dias Lopes has received a fellowship from the Brazilian research agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) within the Science Without Border program.

Received: 3 October 2016 Accepted: 30 December 2016
Published online: 02 February 2017

References

1. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49:1455–1474
2. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? J Comput Aided Mol Des 22:213–228
3. Taminau J, Thijs G, De Winter H (2008) Pharao: pharmacophore alignment and optimization. J Mol Graph Model 27:161–169
4. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216
5. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:239–255
6. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem. J Chem Inf Model 47:488–508
7. Fecher U, Schneider G (2004) Evaluation of distance metrics for ligand-based similarity searching. ChemBioChem 5:538–540
8. von Korff M, Freyss J, Sander T (2009) Comparison of ligand- and structure-based virtual screening on the DUD data set. J Chem Inf Model 49:209–231
9. Nicholls A (2014) Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. J Comput Aided Mol Des 28:887–918
10. Powers DMW (2011) Evaluation: from precision, recall and F-score to ROC, informedness, markedness & correlation. J Mach Learn Technol 2:37–63

11. Fawcelt T (2006) An introduction to ROC analysis. Pattern Recogn Lett 2006(27):861–874
12. Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York
13. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. (2010) The balanced accuracy and its posterior distribution. In: Proceedings of the 20th international conference on pattern recognition, pp 3121–3124
14. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA Protein Struct 405:442–451
15. Smeeton NC (1985) Early history of the kappa statistic. Biometrics 41:795
16. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37:360–363
17. Hawkins PCD, Warren GL, Skillman AG, Nicholls A (2008) How to do an evaluation: pitfalls and traps. J Comput Aided Mol Des 22:179–190
18. Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. Brit. Med. J. 309:102
19. Inserting equation 1 into equation 8 gives $EF(\chi) = \frac{1}{\chi}\frac{n_s}{n}$; hence *EF(χ)* can vary from 0 in the case that n_s equals 0, to $1/\chi$ in the case that n_s equals n
20. Hardison NE, Fanelli TJ, Dudek SM, Reif DM, Ritchie MD, Motsinger-Reif AA (2008) A balanced accuracy fitness function leads to robust analysis using grammatical evolution neural networks in the case of class imbalance. Genet Evol Comput Conf 2008:353–354
21. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
22. Ben-David A (2008) About the relationship between ROC curves and Cohen's kappa. Eng Appl Artif Intell 21:874–882
23. Ben-David A (2008) Comparison of classification accuracy using Cohen's weighted kappa. Expert Syst Appl 34:825–832
24. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22:249–254
25. Peirce CS (1884) The numerical measure of the success of predictions. Science 4:453–454
26. Baker SG, Kramer BS (2007) Peirce, Youden, and receiver operating characteristic curves. Am Stat 61:343–346
27. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3:32–35
28. Schisterman EF, Perkins NJ, Liu A, Bondell H (2005) Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 16:73–81
29. van der Walt S, Colbert SC, Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. Comput Sci Eng 13:22–30
30. Pérez F, Granger BE (2007) IPython: a System for interactive scientific computing. Comput Sci Eng 9:21–29
31. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95
32. MarvinSketch (version 15.10.26), calculation module developed by ChemAxon. http://www.chemaxon.com/products/marvin/marvinsketch/