ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR
DOUTORADO EM BIOLOGIA CELULAR E MOLECULAR

MAURÍCIO BOFF DE ÁVILA

MODELOS COMPUTACIONAIS PARA PREVISÃO DE AFINIDADE ENTRE LIGANTES E PROTEÍNAS ALVOS PARA O DESENVOLVIMENTO DE FÁRMACOS

Porto Alegre
2020

MAURÍCIO BOFF DE ÁVILA

MODELOS COMPUTACIONAIS PARA PREVISÃO DE AFINIDADE ENTRE LIGANTES E PROTEÍNAS ALVOS PARA O DESENVOLVIMENTO DE FÁRMACOS

Tese de Doutorado apresentada como requisito final para a obtenção do grau de Doutor em Biologia Celular e Molecular pelo Programa de Pós-Graduação em Biologia Celular e Molecular da Escola de Ciências da Saúde e da Vida da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Walter Filgueira de Azevedo Jr.

PORTO ALEGRE
2020

Ficha Catalográfica

Z99m Ávila, Maurício Boff de

Modelos computacionais para previsão de afinidade entre ligantes e proteínas alvos para o desenvolvimento de fármacos / Maurício Boff de Ávila. – 2020.

116 p.

Tese (Doutorado) – Programa de Pós-Graduação em Biologia Celular e Molecular, PUCRS.

Orientador: Prof. Dr. Walter Filgueira de Azevedo Jr.

1. Docking. 2. Aprendizado de Máquina. 3. InhA. 4. CDK2. 5. Desenho de Drogas. I. Azevedo Jr, Walter Filgueira de. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

MAURÍCIO BOFF DE ÁVILA

MODELOS COMPUTACIONAIS PARA PREVISÃO DE AFINIDADE ENTRE LIGANTES E PROTEÍNAS ALVOS PARA O DESENVOLVIMENTO DE FÁRMACOS

Tese de Doutorado apresentada como requisito final para a obtenção do grau de Doutor em Biologia Celular e Molecular pelo Programa de Pós-Graduação em Biologia Celular e Molecular da Escola de Ciências da Saúde e da Vida da Pontifícia Universidade Católica do Rio Grande do Sul.

Área de concentração: Bioquímica e Bioinformática.

Aprovada em: 15 de Outubro de 2020

BANCA EXAMINADORA:

_____

Profa. Dra. Fernanda Bueno Morrone - PUCRS

_____

Profa. Dra. Alexandra Martins dos Santos Soares - UFMA

_____

Prof. Dr. Geraldo Francisco Donegá Zafalon - UNESP

PORTO ALEGRE
2020

“Existem muitas hipóteses em ciência que estão erradas. Isso é perfeitamente aceitável, elas são uma abertura para achar as que estão corretas”.

Carl Sagan

AGRADECIMENTOS

À minha esposa Nathália Albuquerque, pelo companheirismo e amor incondicional. O teu incentivo e as longas conversas, em todos os momentos, foram cruciais e indispensáveis. A tua vontade de viver, a paz que lhe permeia e alegria sempre tornam os dias mais fáceis e, com certeza, a tua contribuição para a construção dessa tese foi enorme. Também preciso agradecer apoio e compreensão nos momentos em que precisei dedicar a minha atenção ao desenvolvimento do trabalho. Enfim, mesmo não sendo capaz de expressar toda a minha gratidão, te agradeço por tudo.

Ao meu orientador, Prof. Dr. Walter Filgueira de Azevedo Jr., por todo o acompanhamento na minha trajetória acadêmica, que se iniciou a nove anos atrás, desde a iniciação científica, até aqui. Agradeço por todos os ensinamentos sobre como tratar a ciência com seriedade e, também, por todo o conhecimento técnico proporcionado, além da paciência ao longo desses anos. A tua contribuição para o meu desenvolvimento como cientista foi enorme.

Aos meus pais, Saleti e Renato, por toda criação e ensinamento de valores que vocês buscam me proporcionar até hoje. Sem dúvida, é por causa de todo o sacrifício e dedicação de vocês que me foi permitido chegar até onde me encontro hoje. Agradeço por terem me ensina e mostrado e com tanta dedicação o quão importante e bonito é o ato de aprender e estudar. Sem palavras para expressar toda a minha gratidão.

Aos meus colegas e amigos, por estarem ao meu lado e proporcionar muitos momentos de descontração e reflexão. Obrigado por todo apoio na trajetória e por estarem ao meu lado, também, nos momentos difíceis.

Aos colegas de laboratório pela convivência e parceira nos projetos realizados.

À CAPES pela bolsa concedida, à PUCRS e ao PPG-BCM pela oportunidade de realização do projeto.

“O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal Nível Superior – Brasil (CAPES) – Código de Financiamento 001”

RESUMO

Antibióticos são os medicamentos de maior sucesso do século XX e, provavelmente, de toda a história da medicina. Porém, conforme os anos foram passando, as descobertas de novos compostos antimicrobianos tornaram-se cada vez mais escassas e a resistência bacteriana está em evidência. A partir disso, selecionou-se a enzima *Trans-2*-Enoil (ACP) Redutase (InhA) (E.C. 1.3.1.9) como um dos focos desse trabalho, pois apresenta papel crucial no tratamento antituberculose. Da mesma forma que as novas cepas de bactérias resistentes podem trazer complicações para os próximos anos da saúde pública, as neoplasias são doenças conhecidas há muitos anos, mas que ainda carecem de uma resolução rápida e sem efeitos colaterais graves. Câncer pode ser definido, simplificadamente, como um conjunto de células com crescimento descontrolado e com capacidade de invadir novos tecidos. Com um olhar direcionado para essa necessidade de novos fármacos quimioterápicos contra neoplasias, escolheu-se a enzima Quinase dependente de Ciclina tipo 2 (CDK2) (E.C. 2.7.11.22) como outro alvo do trabalho, em virtude, principalmente, da sua atividade controladora do ciclo celular eucariótico. Em consonância as necessidades atuais expostas anteriormente, o objetivo geral do presente trabalho se designa a determinar as bases estruturais para a inibição das enzimas InhA e CDK2 com enfoque nas interações ocorridas no sistema proteína-ligante. A execução do trabalho foi realizada a partir dos métodos de Computação Bioinspirada, campo da Computação Natural, que baseia sua abordagem em processos observados na natureza. Como base metodológica do estudo seguiu-se as etapas de realização do *docking* molecular na tentativa de encontrar termos energéticos que melhor descrevessem as interações de cada uma das enzimas com possíveis ligantes não-covalentes. Com o auxílio do *software SAnDReS* foram utilizados Métodos de Aprendizado de Máquina para que, a partir dos termos energéticos clássicos presentes nos programas MVD, *AD4* e *Vina*, fossem construídas funções escore polinomiais na tentativa de predizer o grau de afinidade entre os dois sistemas biológicos anteriormente citados e possíveis candidatos a inibidores. Para InhA, as duas funções polinomiais, *Polscore231* (*Vina*) (ρ = 0,709; *p-value1* < 0,03) e *Polscore345* (*AD4*) (ρ = 0,717; *p-value1* < 0,03) obtiveram valores estatísticos satisfatórios colocando-se como boas opções em estudos de seleção de inibidores. Para CDK2, a função polinomial *Polscore60* (MVD) (ρ = 0,328; *p-value1* < 0,02) foi a

melhor opção tanto na predição de afinidade de um conjunto de estruturas com resolução menor de 1,5Å ($HRIC_{50}$), quanto para o conjunto de estruturas contendo apenas CDK's2. A partir dos valores de correlação obtidos para cada uma das funções é sugerido que em estudos posteriores as funções polinomiais sejam utilizadas na seleção de candidatos a possíveis novas drogas com ação inibitória sobre o sítio catalítico dessas duas enzimas.
Palavras-chave: *Docking*. Aprendizado de Máquina. InhA. CDK2. Desenho de drogas.

ABSTRACT

Antibiotics are the most successful drugs of the 20th century and, probably, of the entire history of medicine. However, as the years went by, discoveries of new antimicrobial compounds became increasingly scarce and bacterial resistance is in evidence. From this, we selected the enzyme *Trans-2*-Enoyl (ACP) Reductase (InhA) (E.C. 1.3.1.9) as one of the focuses of this work, because it plays a crucial role in the anti-tuberculosis treatment. In the same way that new strains of resistant bacteria can bring complications for the coming years of public health, neoplasms are diseases that have been known for many years, but which still need to be resolved quickly and without serious side effects. Cancer can be defined as a set of cells with uncontrolled growth and the ability to invade new tissues. A directed look to this need for new chemotherapy drugs against neoplasms, we chose the enzyme Cyclin-dependent Kinase type 2 (CDK2) (E.C. 2.7.11.22) as another target of the work, mainly due to its controlling activity of the eukaryotic cell cycle. In line with the current needs exposed before, the general objective of the present work is to determine the structural bases for the inhibition of the enzymes InhA and CDK2 with a focus on the interactions that occur in the protein-ligand system. The work was carried out using the methods of Bioinspired Computing, a field of Natural Computing, which bases its approach on processes observed in nature. The methodological basis of the study followed the steps of performing molecular docking to find energetic terms that best described the interactions of each of the enzymes with possible non-covalent ligands. With the aid of the *SAnDReS* software, Machine Learning Methods were used, based on the classic energy terms present in the *MVD*, *AD4* and *Vina* programs, polynomial score functions were constructed in an attempt to predict the degree of affinity between the two biological systems before cited and possible candidates for inhibitors. For InhA, the two polynomial functions, *Polscore231* (*Vina*) (ρ = 0.709; p-value1 <0.03) and *Polscore345* (*AD4*) (ρ = 0.717; p-value1 <0.03) obtained satisfactory statistical values, placing themselves as good options in inhibitor selection studies. For CDK2, the *Polscore60* (MVD) polynomial function (ρ = 0.328; p-value1 <0.02) was the best option both in predicting the affinity of a set of structures with a resolution less than 1.5Å (HRIC50), and for the set of structures containing only CDK's2.

From the correlation values obtained for each of the functions, is suggested that in later studies the polynomial functions are used in the selection of candidates for possible new drugs with inhibitory action on the catalytic site of these two enzymes.
Keywords: *Docking*. Machine Learning. InhA. CDK2. Drug design.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AD4: *AutoDock4*

AE: Algortimo Evolucionário

AED: Algoritmo de Evolução Diferencial

AG: Algoritmo Genético

AGL: Algoritmo *Genético Lamarckiano*

ANVISA: Agência Nacional de Vigilância Sanitária

APC: Algortimo de Predição de Cavidades

AUC: *area under the curve*

BIA: Algoritmo Bioinspirado

CDK: quinases dependentes de ciclinas

CDK2: quinase dependente de ciclina tipo 2

DA: Acurácia do *docking*

DHFR: Diidrofolato Redutase

DNA: Ácido Desoxirribonucleico

dTTP: Ácido timidílico

EF: *Enrichment Factor*

Elastic Net CV: *Elastic Net with cross-validation*

ETH: etionamida

FE: Função Escore

IC_{50}: Constante Inibitória a 50%

INH: Isoniazida

InhA: *Trans-2*-Enoil (ACP) Redutase

KatG: catalase-peroxidase

K_i: Constante Inibitória

Lasso CV: *Lasso with cross-validation*

Lasso: *Least Absolute Shrinkage and Selection Operator*

LS: Algoritmo *Local Search*

MOAD: *Mother of All Databases*

MVD: *Molegro Virtual Docker*

NAD^+: nicotinamida adenina dinucleotídeo oxidada

NADH: Nicotinamida adenina dinucleotídeo oxidada

NC: Computação Natural

NMR: Ressonância Magnética Nuclear

OMS: Organização Mundial da Saúde

PDB: Protein Data Bank

PZA: pirazinamida

Ridge CV: *Ridge regression with cross-validation*

Ridge: *Ridge regression*

RMP: rifampicina

RMSD: *Root-Mean Square Deviation*

ROC: *receiver operating characteristics*

SA: *Simulated Annealing*

SAnDReS: *Statistical Analysis of Docking Results and Scoring Functions*

SML: Métodos de Aprendizado de Máquina

STM: estreptomicina

Vina: *AutoDock Vina*

SUMÁRIO

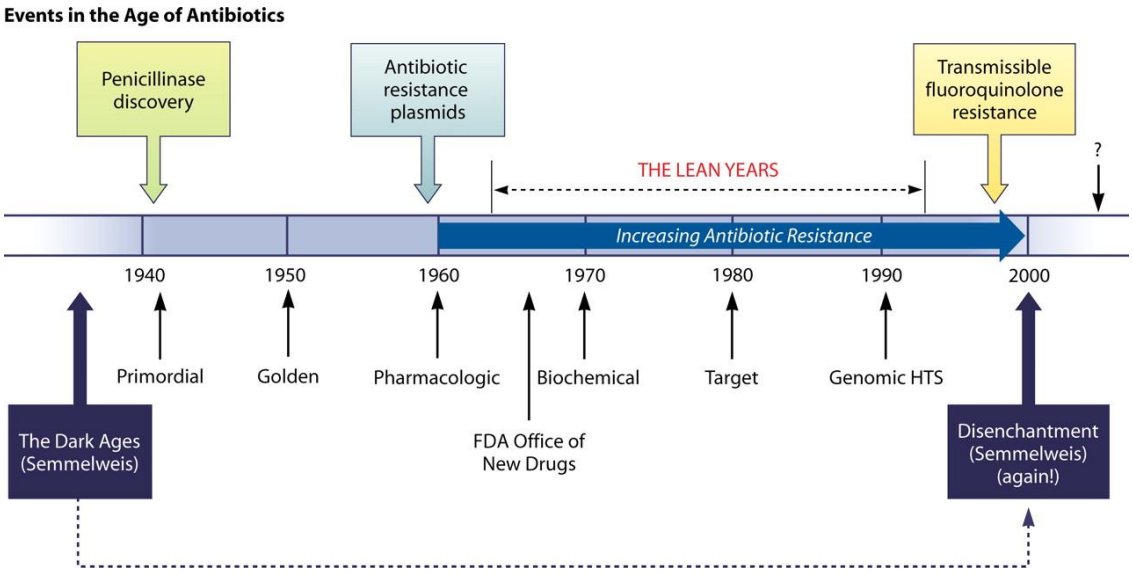
CAPÍTULO 1

1. INTRODUÇÃO

1.1 ANTIMICROBIANOS: UM BREVE HISTÓRICO E SITUAÇÃO ATUAL

Antibióticos são os medicamentos de maior sucesso do século XX e, provavelmente, de toda a história da medicina (Wright, 2007). Representam a primeira, e única, linha de tratamento contra doenças infecciosas ocasionadas por bactérias. Porém, a utilização desenfreada desses compostos a partir da década de 1940 (início de utilização da penicilina) ocasionou um grave problema de saúde pública, que é a seleção de cepas resistentes a antimicrobianos. No conceito básico de antibacterianos descrevem-se moléculas produzidas tanto por micro-organismos (por exemplo a penicilina produzida pelo fungo *Penicillium notatum*), quanto compostos sintetizados em laboratório (isoniazida) que possuem a capacidade de inibir o crescimento (bacteriostático), ou eliminar bactérias (bactericidas) e outros microrganismos (Scholar & Pratt, 2000). Também se relaciona ao conceito de antibiótico a *toxicidade seletiva* (Kaufmann, 2008), em que “somente” as bactérias parasitas sofrem a ação do medicamento (Scholar & Pratt, 2000). Após a descoberta da *penicilina* por Fleming em 1928, houve um considerável aumento no desenvolvimento de novos grupos de antibióticos favorecendo novas formas e possibilidades de tratamentos para doenças infecciosas bacterianas, como pode ser visto na Figura 1. Porém, conforme os anos foram passando, as descobertas de novos compostos antimicrobianos foram ficando cada vez mais escassas dificultando a terapia para novas cepas resistentes que surgem na atualidade (Maamar *et al*. 2020).

Figura 1. Esquema ilustrativo da evolução do desenvolvimento dos antibióticos ao longo dos anos.



Fonte: Extraído de Davies & Davies, 2010

Mesmo com uma escassez na produção de novos quimioterápicos bacteriostáticos e bactericidas, podemos encontrar treze classes de compostos que agem com este efeito. Abaixo estão listadas as seis classes mais importantes na clínica médica com um breve resumo do seu mecanismo de ação (ANVISA, MINISTÉRIO DA SAÚDE, BRASIL, 2007):

1. **β-Lactâmicos:** apresentam como característica em comum a presença de um anel β-lactâmico em seu núcleo estrutural. Atuam dificultando a realização das rotas que sintetizam as moléculas componentes da parede celular bacteriana. A partir da cadeia lateral complementar ao anel pode-se definir o espectro e o grupo de composto. Ex.: penicilinas, cefalosporinas, carbapenemas e monobactamicos.
2. **Quinolonas:** são moléculas derivadas a partir do ácido nalidíxico, sendo mais utilizadas em tratamentos de infecções do sistema gastrintestinal e genito-urinário. Atuam inibindo a ação da DNA-girase (*Enzyme Classsification*: 5.6.2.2), enzima responsável pela compactação do Ácido Desoxirribonucleico (DNA) no nucleoide bacteriano. A inibição da enzima expande o material genético induzindo a taxas exageradas de transcrição, o que leva a morte bacteriana. As fluoroquinolonas apresentam um átomo de flúor a mais que

as quinolonas em geral. Ex.: Ciprofloxacina, Norfloxacina, Gmifloxacina, Levofloxacina e etc.

3. **Glicopeptídeos:** constituídos por grandes estruturas cíclicas complexas, onde são encontradas moléculas de açúcares e peptídeos. Em virtude de sua composição molecular, não são destruídas por enzimas de resistência, como as beta-lactamases. Agem inibindo as reações de síntese da parede polipeptídica. Ex.: Vancomicina e Teicoplanina.
4. **Aminoglicosídeos:** apresentam em sua composição dois ou mais açúcares aminados unidos a um anel por meio de ligações glicosídicas. Representam um dos grupos de antibióticos mais antigos e atuam diretamente sobre a porção 30S do ribossomo bacteriano causando a inibição irreversível da síntese proteica bacteriana. Ex.: estreptomicina, gentamicina, netilmicina etc.
5. **Macrolídeos:** estruturalmente formados por um anel macrocíclico de lactona em que estão ligados um ou mais açúcares. A ação desse grupo é semelhante aos aminoglicosídeos, pois o alvo é o ribossomo, porém o sítio de ligação é na porção 50S o que impede as reações de translocação e transpeptidação. Ex.: azitromicina, claritromicina e eritromicina.
6. **Tetraciclinas:** estrutura química constituída por um núcleo com quatro anéis aromáticos apresentando um baixo nível de toxicidade e pouco custo para produção, características que ajudaram na sua disseminação. Agem diretamente sobre a síntese proteica bacteriana inibindo reversivelmente a porção 30S do ribossomo, também apresenta um largo espectro de ação podendo ser utilizada contra diversos grupos bacterianos. Ex.: Tetraciclina e Doxiciclina.

Apesar de todos os esforços desenvolvidos pelo meio científico no combate contra organismos microbianos, a resistência bacteriana ainda é uma das grandes preocupações da medicina atual. Wright, 2007 descreve a resistência bacteriana como uma rede molecular interligada que confere proteção contra os diversos compostos presentes nos antimicrobianos. Os mecanismos mais comuns de resistência a esses compostos são: efluxo dos compostos antimicrobianos por porinas - presentes na membrana plasmática e parede celular - e mutações gênicas que alteram o sítio de ligação de algumas enzimas específicas, diminuindo a afinidade pelos quimioterápicos. Certos grupos bacterianos podem apresentar casos mais

específicos de resistência, como por exemplo: 1. Bactérias Gram-negativas: apresentam outra membrana externa (lipopolissacarídeo-fosfolipídica) formando uma barreira física contra a entrada dos compostos; 2. Produção de enzimas que destroem os compostos químicos fazendo com que percam a sua atividade antimicrobiana, caso das β-lactamases (destroem o anel β-lactâmico presente na penicilina e em seus derivados) e aminoglicosídeo quinase (grupo de enzimas que inativam compostos antibióticos por meio da fosforilação em região específica). Hughes & Anderson, 2017 entendem que há uma possibilidade de predição da capacidade de resistência desenvolvida por algum organismo contra agentes antimicrobianos caso sejam considerados os seguintes pontos: a) taxa de mutação efetiva na população; b) o nível de resistência gerado pelo mecanismo de resistência; c) a sobrevivência das bactérias mutantes perante diversas taxas de concentração do antibiótico; d) a força das pressões seletivas geradas pelos compostos, além, claro, de diversas outras forças seletivas, como “pescoço de garrafa”, interações epistáticas, evolução compensatória, coevolução e interferências clonais.

Para a Organização Mundial da Saúde (OMS, 2018) a emergência da resistência microbiana é um fenômeno natural acelerado pelo mau uso, ou uso exacerbado de compostos antimicrobianos. Seguindo o *WHO Report on Surveillance of Antibiotic Consumption* (OMS, 2016) micro-organismos responsáveis por doenças graves, como *Pseudomonas spp*., *Klebsiella pneumoniae*, *Escherichia coli*, *Salmonella enterica*, *Staphylococcus aureus*, *Streptococcus pneumoniae* e *Acinetobacter spp*. têm seu tratamento prejudicado em virtude do surgimento de novas cepas resistentes. Além disso, foi reportada a falha da terceira geração de cefalosporinas no tratamento de algumas cepas de *Neisseria gonorrhoeae*, gerando uma série de debates e elaboração de novas estratégias para combater esses micro-organismos (Schrijver *et al.* 2017). Além das doenças destacadas anteriormente, existem outras que assolam a comunidade médica há mais tempo, como a tuberculose.

Na busca por medidas que barrem o surgimento de novas cepas resistentes, algumas ações começam a ser tomadas – um exemplo é o controle do uso indiscriminado de medicamentos, como a Resolução de Diretoria Colegiada 44 (RDC44 – 26/10/2010) (ANVISA, MNISTÉRIO DA SAÚDE, BRASIL, 2010) que proíbe a venda/utilização de 119 antibióticos sem prescrição médica - e a elaboração de novos fármacos. Porém, a segunda proposta é a mais desafiante para o meio

científico. Considerando a linha evolucionária desses fármacos (Figura 1), percebe-se que no período 1940-1960 houve considerável avanço na descoberta de antimicrobianos, gerando uma diversidade de tratamentos no combate às bactérias nocivas ao organismo humano. Todavia, a partir desse período, as únicas novidades encontradas nesse campo foram fármacos considerados de 2ª linha que apresentam baixa eficácia de tratamento, efeitos colaterais mais severos ao paciente e um custo mais alto aos hospitais (Fonseca *et al.* 2015). A partir do panorama atual observado, há necessidade da otimização das técnicas utilizadas na formação de novos fármacos, levando a uma maior agilidade na disponibilidade desses compostos que poderão ser utilizados no tratamento de agentes infecciosos resistentes.

Assim, na tentativa de contemplar os incentivos da OMS em formar novos antimicrobianos, o estudo contido nessa tese foi elaborado. Determinação de Funções Escore otimizam e direcionam a busca por fármacos, diminuindo os custos e o tempo de dedicação para testes *in vitro*. Além disso, o entendimento das bases moleculares ocorrentes na interação entre proteína e ligante permitem a elaboração de fármacos focados no sítio ativo e, também, alostérico, aumentando a possibilidade de sucesso da terapia (de Azevedo & Dias, 2008).

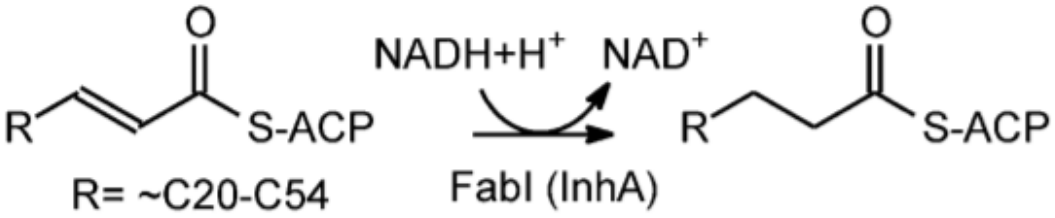
Pensando no panorama montado anteriormente, focou-se o estudo desta tese na enzima *Trans-2-Enoil* (ACP) Redutase (InhA) (E.C. 1.3.1.9), enzima importante na produção de ácidos graxos formadores da parede celular bacteriana, diretamente ligada à proteção contra ataques externos (Rozman *et al*. 2017), e alvo de uma das principais drogas utilizadas no combate ao *M. tuberculosis*, a isoniazida.

A seguir está descrito o sistema biológico ao qual a enzima citada faz parte, as reações catalisadas por ela, sua estrutura tridimensional e caracterização do sítio ativo.

1.1.1 *Trans-2*-Enoil (ACP) Redutase (InhA) (EC 1.3.1.9)

InhA participa da rota metabólica do ácido micólico (importante componente da parede celular de micobactérias), um dos caminhos metabólicos que fazem parte da biossíntese de ácidos graxos, presentes em todos os organismos. Para *M. tuberculosis*, o ácido micólico apresenta importância ainda maior, pois tem papel crucial na reprodução, crescimento do micro-organismo no interior de macrófagos (Pan & Tonge, 2012), além de formar uma barreira de resistência contra compostos antimicrobianos, principalmente os que apresentam características hidrofóbicas (Bhatt *et al*. 2007). Assim, é possível ligar a molécula ao mecanismo de resistência da micobactéria contra o ataques imunológicos e de compostos quimioterápicos. Visto a essencialidade do ácido micólico para a sobrevivência de micobactérias, as enzimas participantes da formação desse composto se colocam como importantes alvos no desenvolvimento de novas drogas (Bhatt *et al*. 2007). InhA é uma enzima nicotinamida adenina dinucleotídeo (NADH) dependente que catalisa a redução estéreo-específica de uma ligação insaturada da cadeia de ácido graxo, realizando a conversão de 2,3-trans-enoil para uma cadeia acila-saturada (Kim *et al*. 2010) (Figura 2).

Figura 2. Reação catalisada pela enzima InhA em que ocorre a redução mediada por $NADH_2$.



Fonte: Extraído de Li *et al*. 2014.

Atualmente, uma grande variedade de compostos é utilizada como inibidora de InhA, com enfoque, principalmente, em micobactérias. Isoniazida (INH), etionamida (ETH), pirazinamida (PZA), rifampicina (RMP) e estreptomicina (STM) são os principais fármacos utilizados na terapia contra a tuberculose (Inturi *et al*. 2016). INH é conhecida como uma droga de primeira linha no tratamento contra o *M*. *tuberculosis*, além de ser considerada o primeiro inibidor conhecido de InhA, ou o pilar na quimioterapia contra a tuberculose (Lee *et al*. 2000). INH é um pró-fármaco, ou seja,

é ingerido na sua forma inativa e depende de uma biotransformação *in vivo* para passar a uma forma ativa e exercer sua função inibitória.

A enzima *KatG* (catalase-peroxidase) (E.C. 1.11.1.21) apresenta papel central nessa necessidade de transformação do inibidor. A enzima é naturalmente encontrada em diversas espécies bacterianas (*E*. *coli*, *Bacillus stearothermophilus*, *Mycobacterium intracellulare* e *M*. *tuberculosis*) realizando as mais variadas funções. Em *M*. *tuberculosis*, a enzima está atrelada a proteção bacteriana, uma vez que inibe os ataques realizados pelo macrófago hospedeiro por meio de compostos reativos de oxigênio (DeVito & Morris, 2003). A relação entre a atividade da *KatG* e a virulência em *M*. *tuberculosis* ainda não está bem estabelecida, existindo estudos que indicam essa relação (Morse *et al*. 1954; Mitchison *et al.* 1963) e outros que trazem uma baixa conexão entre os fatores (Goulding *et al.* 1952; Jackett *et al*. 1978). Além das funções naturais indicadas, a enzima é a responsável pela ativação da INH (Rawat *et al.* 2003), que forma um radical acila ao tornar-se ativada, possibilitando a ligação e, consequente inibição, da InhA. Em estudo realizado por Johnsson & Shultz, 1994 foi demonstrado que a inibição por INH só possível na presença de $NADH^+$, ou NAD^+, sugerindo que o cofator induza uma modificação conformacional na enzima, facilitando o acesso do inibidor ao sítio de ligação.

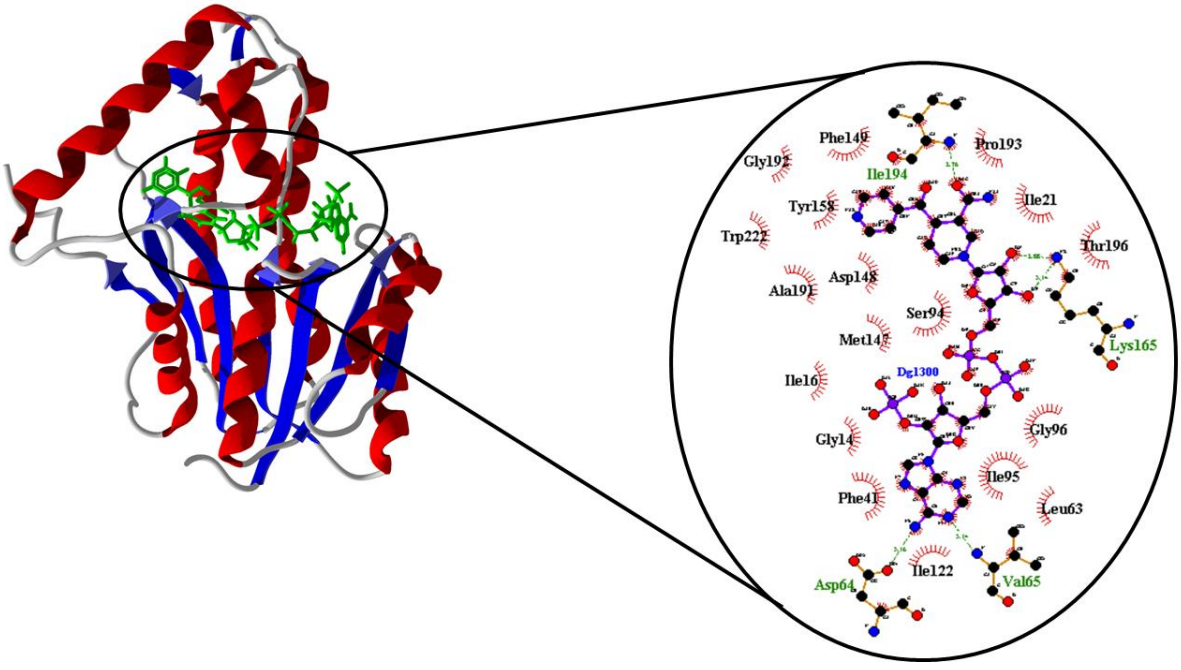
Nos últimos anos foram identificadas várias cepas resistentes, tanto a INH, quanto a outras drogas antituberculose. O principal mecanismo de resistência, no caso de InhA está relacionado com mutações presentes em *KatG* (Vilchèze *et al*. 2017) o que impede a ativação da INH, diminuindo seu potencial de inibição. Bactérias com mutações em *KatG* poderiam ser mais suscetíveis aos ataques realizados pelos macrófagos, já que a função protetora da enzima estaria diminuída, ou perdida. Porém, como relatado por Sherman *et al*. 1996 e Manca *et al*. 1999, a perda de funcionalidade da *KatG* resulta no aumento da expressão de outra peroxidase que acaba por cumprir o papel protetor contra ataques através de H_2O_2. A partir de todo esse panorama, um dos focos desse projeto é facilitar a busca por novos inibidores de InhA, no intuito de contornar esses fenômenos de resistência (Laborde *et al*. 2017).

InhA apresenta 268 resíduos de aminoácidos, com uma massa molecular de 29kDa. A enzima é um homotetrâmero com sítios de ligação ao substrato independentes um do outro. Cada uma das subunidades possui uma estrutura que

lembra uma cadeira (com porções semelhantes à “pernas” e ao “assento”) (Dessen *et al*. 1995), sendo composta por uma folha-β (7 fitas-β paralelas) flanqueada por oito hélices-α (Figura 3) (Scior *et al*. 2002). O bolsão de ligação ao NADH está posicionado entre a porção posterior e o “assento” da “cadeira”. O NADH, quando ligado ao sítio, encontra-se de forma estendida, com a região COOH-terminal próxima ao topo da folha-β e o anel de adenina paralela ao “assento”. A porção de nicotinamida posiciona-se voltada para a região posterior entre as fitas-β e hélices-α (Dessen *et al*. 1995).

Os resíduos de aminoácidos descritos a seguir fazem parte do sítio de ligação da enzima ao inibidor formado pela interação entre INH e $NADH^+$ (derivado da ativação da INH pela *KatG*) (Argyrou *et al*. 2007). Os resíduos que formam ligações de hidrogênio com o inibidor são: Asp64, Val65, Lys165, Ile194. Os contatos hidrofóbicos são formados entre: Phe41, Gly14, Ile16, Met147, Ser194, Asp148, Ala191, Trp222, Tyr158, Gly192, Phe149, Pro193, Ile21, Thr196, Gly96, Ile95 e Leu63 (Figura 3).

Figura 3. Estrutura tridimensional (esquerda) da enzima *trans*-2-enoil (ACP) redutase de *Mycobacterium tuberculosis* **(Difração de Raios X (2,5 Å)) e detalhamento do sítio ativo (direita)**. **Ligante: Complexo formado entre INH-NADP.** Em vermelho estão representadas as dez hélices-α e em azul as dez fitas-β. Em cinza estão representados os *loops*. Em verde o ligante. Pontilhado verde: ligações de hidrogênio; Meios círculos com riscos vermelhos: contatos hidrofóbicos. Código de acesso PDB: 2PR2 (Argyrou *et al*. 2007). *Softwares* utilizados: MVD (Thomsen & Christensen, 2006) e $LigPlot^+$ (Laskowski & Swindells, 2011).



Fonte: próprio autor.

1.2 CÂNCER: ANÁLISE DAS CARACTERÍSTICAS DAS CÉLULAS TUMORAIS

Da mesma forma que as novas cepas de bactérias resistentes podem oferecer complicações nos próximos anos da saúde pública, as neoplasias são doenças conhecidas há muitos anos, mas que ainda carecem de uma resolução rápida e sem efeitos colaterais graves. No ano de 2018, estimativas da OMS indicam que ocorreram 9,6 milhões de mortes no mundo todo em decorrência de complicações ocasionadas por diferentes tipos de cânceres (WHO, 2020). Para o mesmo ano, as estimativas do Instituto Nacional de Câncer (INCA) indicavam 625.370 novos casos no Brasil, sendo os primeiros da lista, o câncer de próstata (65.840 novos casos) e câncer de mama feminino (66.280 novos casos) (www.inca.gov.br). Quanto ao número de mortes ocasionadas pela doença no Brasil, a estimativa é de 218.640 mortes no ano de 2019.

Segundo Alberts (2004, p. 1314) “um tumor é considerado um câncer apenas se for maligno, isto é, somente se suas células tiverem adquirido a capacidade de invadir tecidos adjacentes”. No processo de carcinogênese, há a necessidade da ocorrência de mutações somáticas como substituições, inserções ou deleções de bases, rearranjos de partes recém quebradas de DNA e mudanças na sequência de bases (Stratton, 2011). Essas modificações aleatórias geralmente ocorrem em genes que controlam os mecanismos de proliferação celular e apoptose. Essas alterações no DNA não são herdadas, mas podem ser facilitadas por certos compostos/fatores que “permitem” essas mutações, ou prejudicam a ação de moléculas que regulam os mecanismos de reconhecimento/reparo de alterações no material genético (Bertram, 2000).

Esses fatores estão diretamente relacionados ao modo de vida do indivíduo e trabalham ao longo dos anos a favor das mutações e formação de novas células cancerosas. Em estudo realizado por Zhang e colaboradores, 2017 relacionam uma maior chance de desenvolvimento de câncer com idades mais avançadas, em virtude do acúmulo de mutações e seleção de células mais adaptadas ao ambiente inóspito do tecido saudável.

Em trabalho publicado por Hanahan & Weinberg, 2011 foram listadas seis características essenciais de células cancerígenas, o que levou a uma melhor compreensão e possibilidades de tratamento ampliadas para essa patologia. As seis características são:

1. **Autossuficiência em sinais de crescimento:** esse processo envolve o crescimento descontrolado desses grupos celulares. As células tumorais são capazes de se auto estimularem ao crescimento desordenado (autócrinas), ou receberem sinais de outros tumores (parácrinas). Acredita-se que esse fator esteja relacionado a alta expressão de certos grupos de oncogenes.
2. **Insensibilidade a fatores inibidores de crescimento:** grande maioria das células saudáveis permanece um bom período em quiescência, isso deve-se, geralmente, a fatores internos (genes supressores de crescimento), ou externos (contato célula-célula). As células tumorais não reconhecem esses sinais, ou impedem a ação de genes supressores aumentando, assim, sua capacidade de crescimento e multiplicação.
3. **Evasão da Morte Celular Programada:** a morte celular é um mecanismo de impedimento do desenvolvimento de células danificadas. Esse tipo de mecanismo pode ser ativado a partir da liberação da proteína mitocondrial *citococromo c*, o que estimula a ativação de caspases que evoluem para a apoptose. Células tumorais perdem sensibilidade para esses sinais de morte celular o que as leva a crescerem e dividirem-se de forma indevida em um grande espaço de tempo.
4. **Limite de replicações aumentado:** toda célula apresenta um certo limite de divisões antes de sua morte. Esse número, geralmente, é condicionado pela redução dos telômeros a cada ciclo de divisão celular. Células tumorais apresentam um maior limite de replicações intimamente relacionado a presença de telomerases ativadas que impedem a redução dos telômeros e permitem um maior número de divisões.
5. **Angiogênese:** consiste na formação de novos vasos, ou indução de novas capilaridades a partir de vasos existentes podendo estar associado a inflamação ou cura de ferimentos. Esse mecanismo é realizado pelas células tumorais com o objetivo de nutrição e ganho energético, uma vez que a taxa metabólica desse tipo de célula é alta.
6. **Metástase:** invasão de novos tecidos e de órgãos distantes do tumor primário. É favorecido pelo processo de angiogênese, já que esse mecanismo facilita o acesso de células neoplásicas a corrente sanguínea. A metástase está

associada a 90% das mortes por câncer e o seu controle é um dos principais alvos na busca por tratamentos.

Além das seis características clássicas descritas anteriormente, os autores ainda indicam outros dois *hallmarks* emergentes (Hanahan & Weinberg, 2011). São classificados dessa forma, já que ainda não foram totalmente validados, ou observados de forma geral em todos os tumores. São eles:

1. **Desregulação energética celular:** refere-se à capacidade que algumas células tumorais apresentam de reprogramar seu metabolismo energético com objetivo de suportar a alta proliferação tumoral. Ocorre uma substituição do metabolismo comum à outras células e tecidos pelo reprogramado, como por exemplo a “glicólise aeróbia”, relatado por Warburg, 1930, 1956a, 1956b (*apud* Hanahan & Weinberg, 2011), em que as células tumorais, mesmo na presença de oxigênio, não realizam as fases mitocondriais da respiração celular.
2. **Evitação da destruição imunológica:** envolve vários mecanismos utilizados pelas células tumorais para escapar do ataque e eliminação realizado por células imunológicas, principalmente de linfócitos T e B, macrófagos e células *natural killers*.

Um olhar mais atento para essas características das células tumorais, principalmente as seis clássicas primeiramente citadas, indica uma tentativa desses grupos celulares em aumentar a taxa de sobrevivência junto ao desenvolvimento acelerado, podendo levar à invasão de novos tecidos e a imortalidade celular. A partir do exposto, pode-se entender que algumas dessas características podem servir como alvo na tentativa de barrar o desenvolvimento de células carcinogêneas e impedir a progressão do ciclo de vida desses grupos.

1.2.1 Quinase dependente de ciclina 2 (CDK2) (E.C. 2.7.11.22)

Um dos focos do desenvolvimento de novos fármacos no combate ao câncer são as Quinases dependentes de Ciclina (CDK's) que controlam os *checkpoints* presentes no ciclo celular de células eucarióticas (Russo *et al*. 1996). Para a realização desse projeto foi utilizada a enzima Quinase dependente de Ciclina 2, uma das proteínas reguladoras do ciclo celular e importante alvo no desenvolvimento de fármacos antitumorais (Azevedo *et al*. 1997).

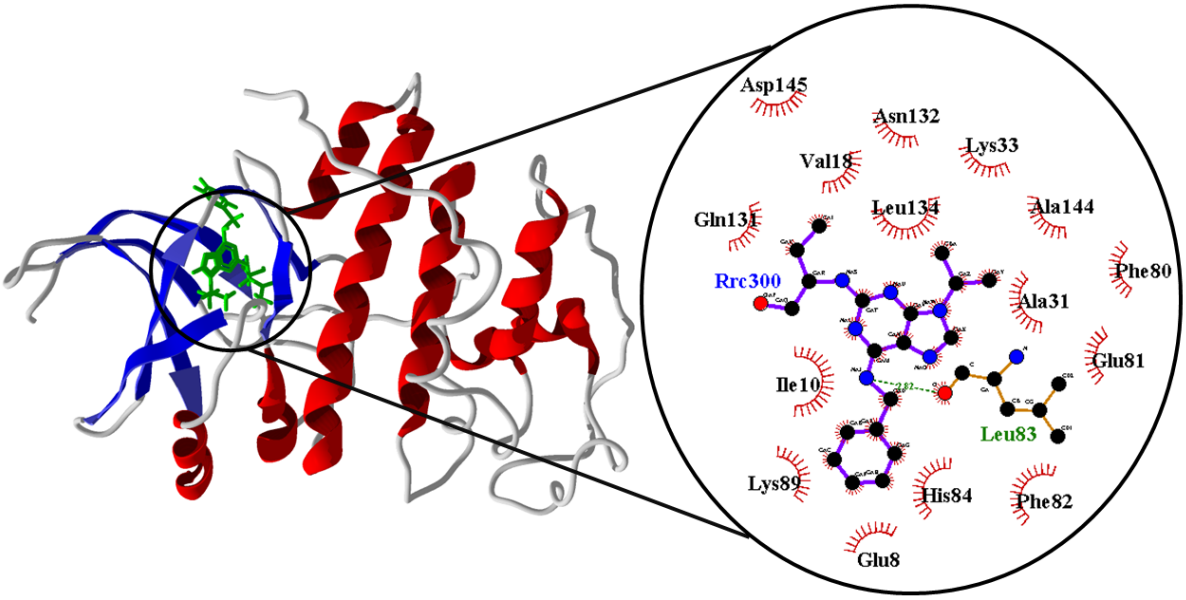
CDK2 é considerada a chave na regulação do ciclo celular, inativando a fosforilação de proteínas da família supressora de tumor RB1 e controlando a transição entre as fases G1/S e G2/M do ciclo celular (Bačević *et al*. 2017). A enzima pode ser encontrada complexada a dois tipos de ciclinas diferentes, A e E, indicando qual evento do ciclo celular está sendo regulado. Complexada a ciclina do tipo A, a enzima promove os eventos de replicação do DNA e a progressão da fase S (Murray, 2004). O complexo CDK2/ciclinaE regula a transição do *checkpoint* final da G1 e os próximos eventos ocorridos na fase S (Li *et al*. 2017).

Em revisão publicada por Malumbres & Barbacid, 2009 é apresentado um panorama do envolvimento de todas as CDK's com a carcinogênese. Os autores não encontraram estudos relacionando mutações em CDK2 presentes em células tumorais. Porém, van den Heuvel & Harlow, 1993; Hochegger *et al*. 2008 indicaram células presas na fase G1 quando apresentavam mutações nas CDK's2. Um estudo mais recente, Yin *et al.* 2018, utilizando diversas bases de dados genéticas demonstraram uma queda na metástase a partir de tumores de próstata quando houve inibição das CDK's2 presentes nesses grupos celulares. Além desses, outros dois estudos relatam a desregulação das CDK's2 como um fator presente e importante em cânceres primários (Cordon-Cardo, 1995; Karp & Broder, 1995), indicando que a enzima é um forte alvo para o desenvolvimento de fármacos contra os mais diversos tipos de câncer. Em recente revisão, Tadesse e colaboradores, 2020 trazem a indicação de novas gerações de inibidores comuns a CDK2 e outras CDK's como possíveis vias de tratamento para cânceres específicos. Caso de tumores endócrino-resistentes, altamente dependentes do eixo G1/S, e, consequentemente, dependentes da CDK2.

A CDK2 é formada por 298 resíduos de aminoácidos, com uma massa molecular 34kDa. A estrutura secundária é chamada bilobar, uma vez que pode ser dividida em dois domínios distintos, um formado por uma folha-β antiparalela (6 fitas-β), localizado próximo a porção N-terminal e outro formado por 7 α-hélices, na região C-terminal. O sítio de ligação ao ATP encontra-se entre os dois domínios (Azevedo *et al*. 1997) (Figura 4). Dois resíduos participantes do sítio de ligação ao ATP (Glu81 e Leu83) são altamente conservados na interação com possíveis inibidores. Por meio desses dois resíduos, são possíveis três ligações de hidrogênio com os grupos C=O (Glu81), N-H e C=O (Leu83) formando um padrão denominado “Garfo Molecular” (Azevedo *et al*. 2002).

Desde a descoberta do *roscovitine*, um inibidor altamente seletivo para CDK’s, há uma grande busca por inibidores análogos a molécula e, também, por novos inibidores (Nekardová, 2017). Azevedo e colaboradores, 1997, descreveram a estrutura cristalográfica da CDK2 complexada ao inibidor *roscovitine* (Código PDB: 2A4L). O inibidor compete pelo sítio de ligação ao ATP, impedindo, assim, a fosforilação de ligantes naturais da enzima (Cicenas *et al*. 2015). O resíduo presente no sítio ativo da enzima que forma ligações de hidrogênio é a Leu83. Já interações de van der Waals, há um número maior de resíduos, sendo eles: Glu8, Ile10, Leu13, Val18, Lys33, Ala31, Phe82, His84, Lys89, Glu81, Phe80, Ala144, Gln131, Asn132 e Asp145 (Figura 4). A alta densidade de resíduos realizando interações com o inibidor demonstra uma grande estabilidade do sistema e um alto grau de afinidade pelo sítio ativo da enzima.

Figura 4. Estrutura tridimensional (esquerda) da enzima quinase dependente de ciclina 2 de *Homo sapiens* **(Difração de Raios X (2,4Å)) e detalhamento do sítio ativo (direita)**. Em vermelho estão representadas as dez hélices-α e em azul as dez fitas-β. Em cinza estão representados os *loops*. Em verde o ligante. Pontilhado verde: ligações de hidrogênio; Meios círculos com riscos vermelhos: contatos hidrofóbicos. Código de acesso PDB: 2A4L (Azevedo *et al*. 1997). *Softwares* utilizados: MVD (Thomsen & Christensen, 2006) e LigPlot+ (Laskowski & Swindells, 2011).



Fonte: próprio autor.

1.3 *DOCKING* MOLECULAR

O posicionamento espacial de duas moléculas que interajam de maneira satisfatória e estável é um problema que desperta interesse em diversas áreas da pesquisa científica (Kunz, *et al.* 1982). A simulação de *docking* de pequenos ligantes contra a estrutura tridimensional de uma proteína alvo é um processo em que se visa buscar, entre as possíveis orientações/conformações de um ligante no sítio ativo de uma proteína, aquela que apresenta a menor energia de ligação (Mitrasinovic, 2012, 2013, Huang *et al*. 2010). Como base para as simulações do *docking*, são utilizadas proteínas/enzimas com estruturas tridimensionais determinadas, geralmente, por técnicas como a Difração de Raios X e a Ressonância Magnética Nuclear (NMR) (Teague, 2003). A informação experimental gerada por meio da difração de Raios X e da NMR é de fundamental importância, uma vez que será a referência para a validação dos protocolos utilizados nas simulações.

A partir disso, é possível dividir o *docking* em duas etapas principais. Na primeira, por meio de diversos algoritmos de busca complexos, são geradas posições possíveis para o ligante, chamadas *poses* (Kitchen, 2006). O sucesso das *poses* é

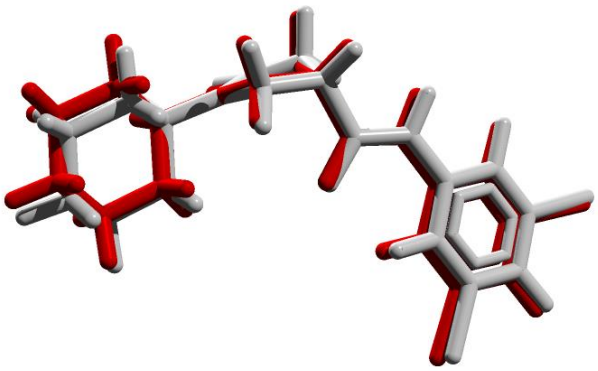
normalmente medido através do valor calculado para o Desvio Médio Quadrático (*Root-Mean Square Deviation* (RMSD)) entre a posição experimental dos átomos do ligante e a posição, para esses mesmos átomos, prevista pelos algoritmos de busca (Sousa, *et al.* 2006) (Figura 5). O cálculo de RMSD está indicado na equação 1,

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)} \qquad \text{(Equação 1)}$$

em que, *vix*, *viy* e *viz*: posição cristalográfica do ligante, *wix*, *wiy* e *wiz*: posição do ligante obtida por meio da simulação de d*ocking*. Um RMSD aceitável para o posicionamento de pequenos ligantes por meio das simulações de *docking* deve variar entre 0 e 2Å (Novic *et al*. 2016). A segunda etapa do processo consiste no cálculo da energia de ligação gerada entre o sistema proteína-ligante simulado (Kitchen, 2006). A energia de ligação pode ser calculada por meio de um conjunto de equações denominadas Funções Escore (Heberlé & de Azevedo, 2011). O cálculo da energia de ligação por meio das Funções Escore tem por objetivo avaliar e classificar as milhares de *poses* geradas por cada algoritmo de busca (Li, *et al.* 2019). Ao final do processo, o algoritmo de busca, gerador da melhor *pose*, e a função escore que calculou a menor energia de ligação, são indicados como o melhor protocolo para predição de afinidade entre a proteína-alvo e o ligante (Eldridge, *et al.* 1997; Jorgensen, 2008).

Identificando o melhor protocolo de *docking*, este pode ser utilizado para vasculhar uma base de dados de pequenos ligantes, como as disponíveis no ZINC (Irwin & Shoichet, 2005). Os valores de RMSD e de energia de ligação indicam que aquele protocolo terá melhor performance para encontrar os ligantes com maior afinidade pelo sítio ativo da enzima alvo. Tal processo é chamado *Virtual Screening* (VS) (Schneider, 2010; Westermaier, et al. 2015).

Figura 5. Posicionamento da *pose* em relação a posição do ligante no sítio ativo da enzima InhA de *Mycobacterium tuberculosis* (Difração de Raios X (1,62Å). Vermelho: *pose*. Cinza: posição cristalográfica do ligante. Código de acesso PDB: 4TZK (He *et al*. 2006). *Softwares* utilizados: MVD (Thomsen & Christensen, 2006).



Fonte: próprio autor.

CAPÍTULO 2

2 OBJETIVOS

2.1 OBJETIVO GERAL

Determinar as bases estruturais para a inibição das enzimas InhA e CDK2 com enfoque nas interações ocorridas no sistema proteína-ligante.

2.2 OBJETIVOS ESPECÍFICOS

1. Realizar simulações de *docking* (*re-docking* e *ensemble-docking*) com as duas enzimas;

2. Testar diferentes *softwares* de *docking* (*Molego Virtual Docker*; *AutoDock4*; *AutoDockVina*) para as duas enzimas;

3. Identificar a melhor Função Escore para cada uma das duas enzimas;

4. Elaborar Funções Escores Polinomiais que possam predizer a afinidade de ligação entre possíveis ligantes e cada uma das duas enzimas.

CAPÍTULO 3

3. MATERIAL E MÉTODOS

3.1 ANÁLISE DE BASES DE DADOS

Para a realização do trabalho proposto foram utilizadas diversas bases de dados para consulta de informações, *BindingDB* (Liu *et al*. 2007), MOAD (*Mother of All Databases*) (Hu *et al*. 2005) e *PDBbind* (Wang *et al*. 2004) que armazenam dados experimentais de inibidores para diversos alvos metabólicos visando o desenvolvimento de novos fármacos. Além desses, também foi consultado o *Protein DataBank* (PDB) (Berman *et al*. 2000) que contém dados estruturais da maioria das proteínas com estruturas resolvidas por Raios X, ou Ressonância Magnética Nuclear (NMR). Na realização do estudo foram selecionadas, usando o próprio filtro do PDB, apenas as estruturas complexadas com inibidores que apresentassem valores de Constante Inibitória a 50% (IC_{50}). Foram utilizadas todas as estruturas de proteínas que apresentavam esses dois valores sem considerar a partir de qual espécie elas foram extraídas.

Para a enzima InhA foram encontradas, após uma série de critérios (presença de água próximo ao sítio ativo, repetição de ligantes, entre outros), trinta e duas estruturas cristalográficas com valores de IC_{50}. Para CDK2, após os mesmos critérios de seleção seguidos para InhA, onze estruturas com valores para IC_{50} (Tabela 1). Esses dados foram encontrados no início da execução do projeto, 25/07/2017 (de Ávila *et al*. 2017).

Tabela 1. Estruturas cristalográficas utilizadas em cada conjunto nas simulações de *docking*.

Conjunto Códigos PDB	**Códigos PDB**
InhA-IC_{50}[1]	1I2Z, 1LXC, 1MFP, 1P44, 1QSG, 2B37, 2FOI, 2NSD, 2NV6, 2OL4, 2OOS, 2X23, 3AM4, 3FNE, 3FNF, 3FNG, 3OJF, 3ZU4, 3ZU5, 4BII, 4BQP, 4BQR, 4COD, 4D44, 4IGE, 4IGF, 4Q9N, 4TRJ, 4TZK, 4TZT, 4U0J, 4U0K
$HRIC_{50}$[2]	2GG3, 2GG7, 2GG9, 2HU6, 2I5F, 2IKG, 2NMZ, 2NNG, 2OW6, 2PDG, 2PIY, 2PZN, 2QCF, 2R3I, 2W14, 2W3B, 2W9H, 2WUU, 2WZX, 2X5O, 2XPC, 2XU3, 2XU4, 2Y1O, 2Y68, 2YC3, 2YEX, 2YJ2, 2YJ8, 2YJ9, 2YJ, 2YK9, 2YKE, 2YKJ, 3B28, 3B7E, 3B8Z, 3BCJ, 3BLB, 3CBP, 3DCR, 3DD0, 3DN5,3EJS, 3EJT, 3EJU, 3ESS, 3EWZ, 3EX3, 3F66, 3FCI, 3FS6, 3GHV, 3GHW, 3H5B, 3HHA, 3HJ0, 3HNB, 3HS4, 3HYG, 3I06, 3I33, 3I6C, 3I6O, 3IOG, 3IU7, 3KFA, 3KIG, 3KKU, 3KL6, 3KWZ, 3L14, 3M0I, 3M4H, 3NKK, 3NTZ, 3NU0, 3NU3, 3NWB, 3NXO, 3NXX, 3NZB, 3OND, 3OT3, 3OVX, 3OZS, 3OZT, 3PA3, 3PKA, 3PKB, 3PX8, 3R6T, 3RL4, 3S1Y, 3S71, 3SP, 3TEM, 3U2C, 3UHM, 3VF3, 3VHV, 3VW9, 3WFG, 3ZSJ, 3ZXH, 4A6V, 4A6W, 4BW1, 4DHR, 4DRI, 4DRN, 4DRO, 4DRQ, 4E4A, 4F3I, 4FH2, 4FLK, 4FYO, 4GCJ, 4GQR, 4GV1, 4HCT, 4HCU, 4HCV, 4HWW, 4HXQ, 4HXS, 4HY4, 4HYI, 4IGH, 4IKU, 4JHT, 4KEB, 4L7G, 4M5R
CDK2-IC_{50}[3]	1GII, 1OIR, 2B53, 2B54, 2R3H, 3IGG, 3LE6, 3PXZ, 3PY0, 3RZB, 4RJ3

[1] 32 estruturas de InhA.
[2]173 estruturas de alta resolução (<1,5Å)
[3] 11 estruturas de CDK2.

3.2 SIMULAÇÕES DE *DOCKING*

Como pode ser observado na Tabela 1, as duas enzimas alvo deste trabalho apresentam mais de uma estrutura cristalográfica, complexada a um ligante, depositada no PDB. A estrutura escolhida para a realização de *re-docking*, validação dos protocolos de *docking* utilizados pelos *softwares*, é aquela que apresenta a melhor resolução entre todas do conjunto montado. Para InhA a estrutura escolhida foi a de código PDB 4TZK e para o conjunto de alta resolução ($HRIC_{50}$) foi a de código 1US0 (Tabela 2). Dessa forma pode-se estender a validação do protocolo estabelecido pelo *re-docking* para todas as estruturas componentes dos conjuntos montados. Esse processo de validação utilizando outras estruturas é chamado de *ensemble-docking*. O objetivo do *ensemble-docking* é confirmar se o protocolo estabelecido pelo *re-*

docking é capaz de reconhecer outros ligantes no sítio ativo de estruturas diferentes da mesma enzima seguindo os mesmos valores de RMSD (0-2Å). Para todas as simulações de *docking* presentes no projeto utilizamos o programa *Molegro Virtual Docker* (MVD) (Thomsen & Christensen 2006), *AutoDock4* (AD4) (Morris *et al.* 1998) e *AutoDock Vina* (Vina) (Trott & Olson, 2010).

Tabela 2. Estruturas cristalográficas utilizadas nas simulações de *re-docking*.

Código Acesso PDB	Resolução (Å)	Nome da enzima (organismo)	Código Acesso Ligante	Coordenadas do ligante no sítio da enzima
4TZK	1,62	Enoil redutase (InhA) *M. tuberculosis*	641 [A]	x: 8,79 y: 32,47 z:60,42
1US0	0,66	Aldose redutase *Homo sapiens*	LDT [A]	x:16,57 y: -7,30 z: 15,60

3.3 COMPUTAÇÃO NATURAL

A computação natural (NC) pode ser definida como a versão computacional do processo de extração de ideias da natureza na resolução de problemas computacionais (de Castro, 2007). Segundo de Castro, 2007 existem três subáreas para a NC, sendo elas:

1. Computação Inspirada na Natureza: representada pelas estratégias desenvolvidas a partir de algum processo existente na natureza. Pode-se citar como exemplo dessa área, as redes neurais artificiais.
2. Estudos da Natureza através da Computação: relacionado com a utilização de procedimentos computacionais para o entendimento de padrões biológicos. Os estudos sobre organismos artificiais podem ser citados como exemplos.
3. Computação com Mecanismos Naturais: nessa área, os mecanismos naturais são utilizados como dados para o desenvolvimento de "computadores naturais".

Partindo das três subáreas citadas, o projeto realizado foi baseado na primeira abordagem ("Computação Inspirada na Natureza), uma vez que os algoritmos utilizados pelos programas de *docking* podem ser classificados como Algoritmos Bioinspirados (BIA's), ou seja, são formulados a partir de processos já existentes na

natureza. Nos últimos anos, os BIA's têm emergido como uma solução viável e efetiva para muitos problemas computacionais. Esses algoritmos mimetizam as ideias evolutivas de Darwin, 1859, ou a inteligência visível em animais sociais (colônias de formigas, bandos de aves, enxames de abelhas etc.) (Folino & Mastroianni, 2011), formando uma classe especial de algoritmos desenvolvidos para a solução de problemas computacionais complexos (Mukhopadhyay, 2014). Assim, os BIA's podem ser definidos como um conjunto de técnicas computacionais que utilizam os fenômenos biológicos como inspiração na sua formação (Heberlé e de Azevedo, 2011).

Dentre os diversos grupos de BIA's presentes atualmente, os que mais se destacam para o desenho de drogas são os algoritmos evolucionários, pois se apresentam, frequentemente, com sucesso na resolução de problemas complexos (Parril, 1996). Os algoritmos evolucionários são considerados como um subcampo do BIA, tendo emergido em meados da década de 1990 a partir da ideia de utilizar os conceitos descritos em "A Origem das Espécies" (Darwin, 2014) com o objetivo de resolver problemas de busca e otimização (Lameijer *et al*. 2005).

A seguir estão as definições e métodos de funcionamento de cada um dos algoritmos utilizados na abordagem computacional do projeto.

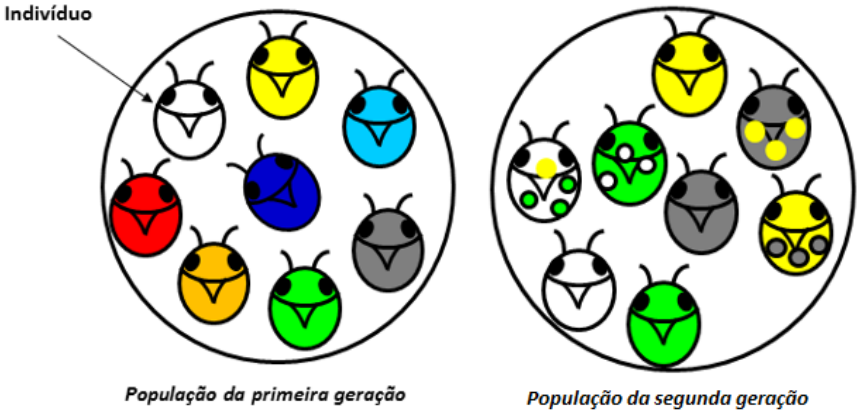
3.4 ALGORITMOS EVOLUCIONÁRIOS

Algoritmos evolucionários (AE's) usam as ideias da evolução com o objetivo de gerar abordagens computacionais para tratar problemas de otimização e biológicos. De forma geral, AE's podem ser classificados como algoritmos estocásticos, classe em que se enquadram também Monte Carlo, *Tabu Search* e *Swarm Optimization* (Floreano & Mattiussi, 2008).

Pode-se utilizar uma analogia biológica para o melhor entendimento desse tipo de algoritmo. Por exemplo, um ambiente que sustenta somente um determinado número de joaninhas (Figura 6) irá selecionar os animais que apresentarem características que os favoreçam na disputa por recursos do ambiente, ou seja, a seleção dos indivíduos mais adaptados aquele determinado local. Utilizamos, também, outro conceito básico da teoria de Darwin, a variação do fenótipo entre os indivíduos da população. Traços do fenótipo são aqueles aspectos físicos e de

comportamento de um indivíduo, que estão relacionados com sua reação ao ambiente. Os traços do fenótipo determinam seu ajuste (*fitness*) ao ambiente (Heberlé & Azevedo, 2011).

Figura 6. Indivíduos criados aleatoriamente na primeira geração são submetidos a operadores genéticos. Após seleção usando-se uma função ajuste, passam a compor a população da segunda geração.



Fonte: Extraído de Heberlé & Azevedo, 2011.

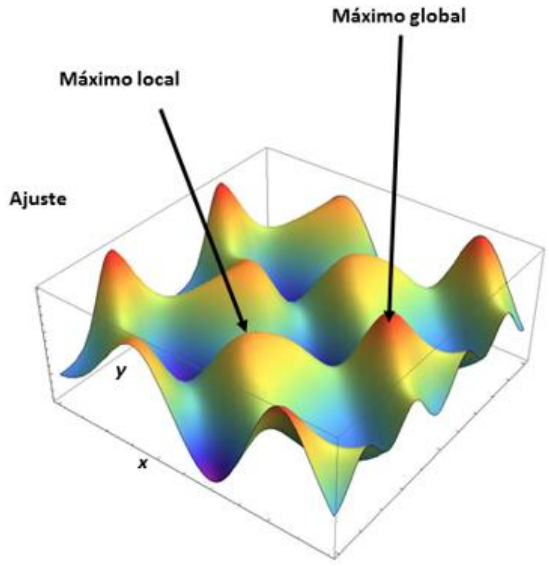
Numa população, cada indivíduo mostra um conjunto de traços do fenótipo exclusivos, que estão continuamente sendo testados pelo ambiente (Holland, 1975). O ajuste de cada indivíduo é quantificado numa análise matemática do sistema biológico, os indivíduos com maior sucesso apresentam um ajuste mais alto, ou seja, uma função ajuste (*fitness function*) mais alta. Assim, pode-se perceber que uma população é um conjunto de indivíduos, onde cada um é considerado uma "unidade de seleção", sendo que seu sucesso depende do quão bem adaptado ele está ao seu ambiente. Indivíduos mais bem adaptados apresentam probabilidade mais alta de gerar descendentes, e, mutações ocasionais, dão origem a novos indivíduos que serão testados pelo ambiente (Herberlé & Azevedo, 2011).

Assim, de uma geração para outra, o ambiente modula a variação no genótipo da população, uma vez que os genes selecionados serão passados para as próximas gerações. Na figura 6 pode-se perceber diferenças nos padrões de cores, mesmo que alguns indivíduos tenham sobrevivido de uma geração para outra (Herberlé & Azevedo, 2011).

O processo de evolução também pode ser visualizado numa superfície adaptativa (*adaptive surface*), que utiliza uma representação multidimensional. Nossa

função ajuste levará em consideração dois parâmetros baseados em traços do fenótipo (Figura 6). No gráfico da figura 7 considera-se que o eixo x representa a capacidade de mimetismo (camuflagem) das joaninhas, e o eixo y a capacidade de escalar obstáculos. A combinação entre x e y nos dará a função ajuste. Os picos mais altos na figura 7 representam os indivíduos mais bem adaptados. Os vales indicam indivíduos menos adaptados ao ambiente. Cada pico na superfície indica um ponto ótimo local (ou simplesmente máximo local). O pico mais alto representa o ponto de ótimo global. Com esta analogia podemos visualizar a utilização da evolução em problemas de otimização, onde procuramos a melhor solução para um determinado problema. O conjunto de soluções possíveis é chamado de espaço de busca, a superfície adaptativa, da figura 7, é uma representação gráfica do espaço de busca.

Figura 7. Superfície adaptativa com indicação de máximo local. Os eixos x-y representam traços do fenótipo. A altura z representa a função ajuste. Cada ponto na superfície representa um indivíduo com a combinação de traços do fenótipo.



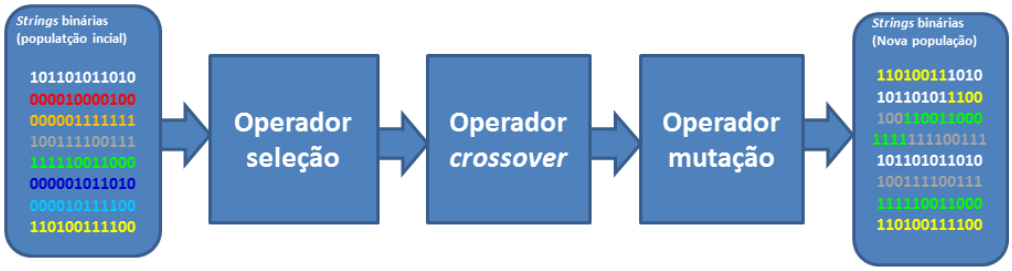
Fonte: Extraído de azevedolab.net

3.4.1 Algoritmo genético

Entre os AEs, o algoritmo genético (AG) é o mais estudado. Um AG usa, frequentemente, três operadores para manipular uma população de indivíduos gerados aleatoriamente, são eles: seleção, *crossover* (cruzamento) e mutação. O diagrama da figura 8 nos fornece uma visão geral destes passos para uma iteração (geração) (Goldberg, 1989). A população inicial pode ser formada de *strings* binárias

geradas aleatoriamente, os números indicados à esquerda. O operador seleção é uma implementação computacional da “seleção natural”, sendo uma tentativa de aplicar a pressão evolucionária sobre indivíduos de uma população. Indivíduos não adaptados ao “ambiente”, ou com função ajuste baixa serão descartados. Os indivíduos mais adaptados, ou seja, os que possuem uma maior função ajuste apresentam maior probabilidade de sobreviver. Os sobreviventes farão parte da população que começará a nova geração (iteração).

Figura 8. Operadores de um algoritmo genético.



Fonte: Extraído de azevedolab.net

O operador *crossover*, também conhecido como recombinação, faz com que os indivíduos *(strings* binárias) troquem sua informação genética, de forma análoga à reprodução sexuada.

O operador *crossover* é aplicado a um par de pais, para gerar um par de novas *strings* binárias (descendentes). Um par de pais será submetido ao operador *crossover,* se um número aleatório (Rn) for menor que a probabilidade de *crossover* (Pc). Rn está no intervalo [0,1], e um típico valor de Pc está no intervalo [0,4, 0,9] (Coley, 1999). O valor de Pc escolhido irá determinar a parcela da nova população formada por seleção e *crossing-over*. Por exemplo, se Pc = 0,7, 70% da nova população será formada por seleção e permutação, os outros 30% será formada, somente, por seleção. Além de usarmos *strings* binárias, poderíamos, também, utilizarmos números reais decimais para representação das *strings*. Historicamente os AGs foram inicialmente desenvolvidos usando-se *strings* binárias, ou cromossomos usando a analogia biológica (Goldberg, 1989; Holland, 1975).

3.4.2 Algoritmo de Evolução Diferencial (AED)

Na evolução diferencial tem-se os operadores clássicos dos algoritmos genéticos: seleção, *crossover* e mutação, como mostrados na figura 8. O algoritmo de evolução diferencial executa a parte inicial de geração aleatória de uma população com N cromossomos (ou *strings*), em seguida avalia a função ajuste de cada cromossomo (Storn & Price, 1997). A diferença está na formação dos cromossomos filhos. Eles são gerados da seguinte forma, para cada cromossomo pai na população, chamado aqui de cromossomo j, são escolhidos aleatoriamente três outros cromossomos pais distintos, cromossomos k, l e m. Estando selecionados esses três cromossomos, calcula-se um novo cromossomo, chamado aqui de n, da seguinte forma:

$$\text{cromossomo}(n) = \text{cromossomo}(m) + \text{peso}.[\text{cromossomo}(k) - \text{cromossomo}(l)] \quad \text{(Equação 2)}$$

sendo que o peso pode variar entre 0 e 2, esse valor é utilizado para controlar a amplificação da variação diferencial (Storn & Price, 1997). O cromossomo novo (n) será incorporado à população se o resultado da equação estiver no intervalo [0,1] e este número for menor que a probabilidade de crossover. Caso não seja, o novo cromossomo filho não é considerado. Um último teste é realizado no cromossomo filho n, se este apresenta função ajuste maior que o cromossomo pai j. Caso seja, o cromossomo j é deletado e substituído pelo cromossomo n (Price, 1996). As etapas seguintes são idênticas ao algoritmo genético original.

3.5 FUNÇÕES ESCORES EMPÍRICAS

A utilização dos AEs destaca a função ajuste como critério de seleção de indivíduos mais bem adaptados, sendo considerado que quanto maior o seu valor melhor adaptado está o indivíduo. Na análise de resultados da simulação de *docking* molecular, considera-se que os melhores resultados são os de menor energia, assim na aplicação de AEs ao problema de *docking* molecular, o tipo de função a ser utilizada na seleção de melhores resultados será a função escore. As funções escores são divididas em três grandes famílias. A primeira família é formada por funções baseadas

em campos de força, como a função escore do programa DOCK (Meng *et al.,* 1992), que se baseia no campo de força AMBER (Weiner *et al.,* 1984). A segunda família é formada por funções escores empíricas, originalmente propostas por Böhm (Böhm, 1994; 1998), onde os termos de uma função recebem peso, de forma a concordar com afinidades determinadas experimentalmente. A terceira família são funções baseadas em conhecimento, que usam funções de energia potencial derivadas de estruturas obtidas experimentalmente (Tanaka & Scheraga, 1976; Sippl, 1990).

Nosso foco inicial será a implementação de funções escores empíricas, que são relativamente mais rápidas de calcular que as outras citadas acima (Guedes *et al*. 2018). Testaremos a adequação de três tipos de funções escores empíricas, aqui chamadas de funções escores lineares, não-lineares e não-lineares mistas. A função escore empírica linear (ΔG1) tem a seguinte forma geral,

$$\Delta G1 = \sum_j w_j \Delta g_j \qquad \text{(Equação 3)}$$

onde w_i é o peso de cada termo energético Δg_j. Os pesos relativos de cada termo da somatória, serão obtidos a partir de ajuste contra um conjunto de estruturas, para as quais há informação estrutural e de afinidade disponíveis nas bases de dados MOAD (*Mother Of All Databases*) (Hu *et al.,* 2005), *BindingDB* (Liu *et al*., 2007) e *PDBbind* (Wang *et al.,* 2004). Foram testados diferentes termos relevantes para interações intermoleculares, tais como ligações de hidrogênio, área de contato, área de contato polar e hidrofóbica, área acessível ao solvente, interação eletrostática entre outros possíveis (de Azevedo & Dias, 2008).

A função escore empírica não-linear (ΔG2) admite termos de potenciais mais altas, como $(\Delta g_j)^2$ ou outros expoentes. Os termos serão os mesmos descritos da ΔG1, com a diferença do expoente de cada termo. A função escore empírica não-linear mista (ΔG3) adiciona termos mistos. Por exemplo, as interações de ligação de hidrogênio são representadas pelo termo Δg_1 e as interações de van der Waals por Δg_2, teremos um termo misto (Δg_1 Δg_2) na somatória da equação 3, além dos termos lineares da ΔG1 e não-lineares, já citados para a ΔG2.

Os três tipos de famílias de funções escores empíricas serão ajustadas por classe enzimática, sendo selecionada aquela que apresenta melhor concordância

com dados experimentais, usando-se critérios estatísticos, como o coeficiente de correlação de Spearman (Zar, 1972), entre a afinidade prevista pela função escore empírica e a experimental. Formamos dois tipos de bases de dados de afinidade, o conjunto treino (*training set*) com 70% dos dados experimentais disponíveis sobre afinidade, usados para obter os pesos da equação 3. A segunda base de dados foi formada por 30% dos dados experimentais disponíveis, que não foram usados para obtermos os pesos da equação 3, mas sim para testar se a função estabelecida realmente é capaz de prever a afinidade entre enzima e ligante. Estes dados são chamados de conjunto teste (*test set*). Tal abordagem normalmente é usada na calibragem das funções escores empíricas (de Azevedo & Dias, 2008).

Os pesos relativos das variáveis independentes das funções escores empíricas são determinadas a partir de técnicas de aprendizado de máquina supervisionado implementados no programa *Statistical Analysis of Docking Results and Scoring functions* (SAnDReS) (Xavier *et al.* 2016; Heck *et al*. 2017). As técnicas utilizadas pelo *software* são: *Least Absolute Shrinkage and Selection Operator* (Lasso), *Lasso with cross-validation* (Lasso CV), *Ridge regression* (Ridge), *Ridge regression with Cross-Validation* (Ridge CV), *Elastic Net*, *and Elastic Net with Cross-Validation* (Elastic Net CV)

O uso da metodologia descrita acima junto a descrição de um modelo de aprendizado de máquina utilizado no presente estudo para a previsão computacional do log(IC_{50}) foi publicado na revista *Biochemical and Biophysical Research Communications* (de Ávila *et al*. 2017).

3.6 FERRAMENTAS UTILIZADAS

Como explicado nas sessões anteriores, na execução do projeto de pesquisa foram utilizadas quatro ferramentas para as simulações de *docking* e, posterior, análise estatística dos resultados. Nas simulações, os *softwares Molegro Virtual Docker*, *AutoDock4* e *AutoDock Vina* foram usados para as tarefas de *docking*. A seleção dos softwares de *docking* do foi baseada nos valores de acurácia e capacidade de posicionamento das poses no sítio da enzima (Thomsen & Christensen 2006; Morris *et al.* 1998; Trott & Olson, 2010; Fraczek *et al*. 2013; Gaillard, 2018). Por meio da ferramenta *SAnDReS* realizou-se as análises estatísticas e a construção de

novas funções escores polinomiais que possibilitaram predições de afinidade do sistema proteína-ligante mais significantes estatisticamente.

A seguir encontram-se as descrições de funcionamento de cada uma das ferramentas.

3.6.1 Molegro Virtual Docker (MVD)

O *software* MVD é um simulador de *docking* focado em interações não-covalentes realizadas entre proteínas e ligantes. Do mesmo modo que outros diversos *softwares*, o MVD trabalha a partir de funções escore baseadas no cálculo de energia gerada pelo sistema proteína-ligante para a determinação da afinidade estabelecida entre as duas moléculas (Thomsen & Christensen, 2006).

O programa utiliza, como base para as buscas e cálculos, o algoritmo de evolução diferencial (AED), uma variação do algoritmo evolucionário que consiste gerar novos "indivíduos na população" de *poses* a partir de outras já existentes, conforme explicado anteriormente. Aliado ao AED o *software* utiliza um segundo tipo de algoritmo que auxilia na restrição de predição das poses que serão geradas, o Algoritmo de Predição de Cavidades (APC) (Thomsen & Christensen, 2006). O APC utiliza uma série de passos para identificar/predizer as possíveis cavidades presentes na proteína-alvo utilizando como referência a conformação espacial e o volume ocupado pelos átomos dos resíduos presentes na proteína. Em comparação a outros tipos de algoritmo de busca de cavidades, o APC tem um desempenho melhor, uma vez que busca de forma aleatória e em maior número de vezes mensurar o espaço ocupado por cada átomo.

As Funções Escores (FE's) utilizadas pelo software no cálculo da energia gerada entre a molécula de proteína e as *poses* formadas, são as seguintes, *MolDock Score* e *Plants Score*, sendo a escolha do usuário qual tipo de função será utilizada em conjunto aos algoritmos de busca (*MolDock Optimizer*, *MolDock SE* e *Interated Simplex*).

A função *MolDock Score* atua calculando a energia gerada pelas ligações de hidrogênio estabelecidas entre os átomos da proteína alvo e do ligante utilizado na simulação. São consideradas ligações de hidrogênio os pares de átomos que podem atuar como doadores, ou receptores de átomos de hidrogênio durante a interação

(Muller, 1994). Também é considerado como fator relevante para os cálculos de energia a possibilidade de interações de Van der Waals entre os átomos das duas moléculas (proteína e ligante). A equação utilizada pela função para os devidos cálculos está descrita na Equação 4,

$$E_{score} = E_{inter} + E_{intra} \quad \text{(Equação 4)}$$

onde, E_{score}: função *MolDock Score*; E_{inter}: energia de interação intermolecular (proteína-ligante); E_{intra}: energia intramolecular (ligante). Além dos termos utilizados pela função, o *software* permite a inclusão de diversos outros grupos que possibilitam uma análise mais complexa do sistema, os termos energéticos utilizados e suas descrições constam na Tabela 3:

Tabela 3. Descrição dos termos energéticos utilizados no cálculo da função *MolDock Score* e *PLANTS Score*.

Termos	Descrição
Interaction	Energia total de interação entre proteína e *pose*
Cofactor	Energia de interação entre *pose* e cofatores
Protein	Energia de interação entre *pose* e proteína
Water	Energia de interação entre *pose* e moléculas de água
Internal	Energia interna da *pose*
Electro	Interações eletrostáticas entre proteína-ligante de curto alcance (r<4.5Å)
Electro Long	Interações eletrostáticas entre proteína-ligante de longo alcance (r>4.5Å)
H-Bond	Energia entre ligações de hidrogênio
LE1	Eficiência do Ligante 1: *MolDock Score* dividido pela contagem de átomos pesados
LE3	Eficiência do Ligante 2: *Rerank Score* dividido pela contagem de átomos pesados

A função *PLANTS Score* é derivada da função *PLANTS* originalmente descrita e formulada por Korb *et al*. 2009. A equação utilizada para estabelecer o valor de *PLANTS Score* é descrita abaixo na equação 5 abaixo,

$$E_{plantsscore} = F_{plp} + F_{clash} + F_{tors} + C_{site} - 20 \quad \text{(Equação 5)}$$

onde, $E_{plantsscore}$: função *PLANTS Score*; F_{plp}: *piecewise linear potential*; F_{clash}: "conflitos" entre as ligações internas no ligante; F_{tors}: contribuições das torções para as ligações flexíveis no ligante; C_{site}: penalidade para caso a pose localize-se fora do sítio ativo da enzima). O *piecewise linear potential* (potencial em partes) é utilizado de forma semelhante ao presente na *MolDock Score*, porém apresenta um número maior de interações (repulsão, interações não polares; ligações de hidrogênio e metálicas) favorecendo a uma aproximação entre o modelo montado e a realidade dos ligantes reportada nos dados experimentais. O termo F_{clash} considera o conjunto de átomos do ligante que apresentam, no mínimo, três ligações um do outro e estão em fragmentos rígidos separados por ligações flexíveis, ou com capacidade de rotação. F_{tors} é estabelecido a partir de todas as ligações rotáveis presentes no ligante, exceto as ligações formadas pelos grupos terminais que são doadores nas ligações de hidrogênio (Korb *et al*. 2009). O desconto de 20 ao final da equação tem o objetivo de deixar a função *PLANTS Score* mais próxima a função *PLANTS* original. As diferenças importantes presentes entre as duas funções é a não utilização da função *"Tripos Torsional"*, que visa buscar a energia de torção total presente no ligante, pela função *PLANTS* original; a segunda diferença encontrada é a penalização de 10000 caso a *pose* seja encontrada fora do sítio ativo, já que não é adequada a utilização do termo C_{site} para os algoritmos de busca *MolDock Optimizer* e *MolDock SE*. Os termos citados na Tabela 3 também podem ser utilizados durante o cálculo da função *PLANTS Score.*

3.6.2 AutoDock4 (AD4)

O programa AD4 é um exemplo de ferramenta que busca um acoplamento mais flexível e fisicamente detalhado entre proteína e ligante (Morris *et al*. 1998). O *software* combina o campo de força de energia livre com o Algoritmo Genético Lamarckiano (AGL) fornecendo maior predição das interações presentes no sistema, além dos cálculos das energias de associações (Morris *et al*. 2009).

São utilizados três tipos de algoritmos de busca no posicionamento das poses pela ferramenta: *Simulated Annealing* (SA), Algoritmo Genético (AG) e o Algoritmo Genético Lamarckiano (AGL).

O SA (Goodsell & Olson, 1990) é uma variante do algoritmo genético que atua com as variantes energéticas do sistema. Ela atua tanto de uma forma global, no momento em que busca todas os valores energéticos do sistema estabelecido, quanto com uma busca local, uma vez que busca, dentro dos valores energéticos calculados, aquelas posições atômicas que apresentem menor valor energético em relação aos átomos participantes do sítio ativo da enzima (Morris *et al*. 1998).

O AG utilizado aqui segue o mesmo princípio de funcionamento explicitado no tópico sobre Algoritmos Evolucionários explicados anteriormente.

O AGL utiliza o raciocínio inverso ao utilizado pelo AG, seguidor das ideias de Darwin, ou seja, o genótipo é quem define o fenótipo. Já no AGL, são utilizadas as ideias evolutivas de Lamarck que partiam do pressuposto de que modificações fenotípicas ocorridas ao longo da vida do indivíduo poderiam ser herdadas pelas próximas gerações, ou seja, o fenótipo seria capaz de modificar o genótipo (Purves *et al*. 2005). Seguindo Morris e colaboradores, 1998 nas simulações de *docking* o genótipo representa o “estado” do ligante na simulação, enquanto o fenótipo é dado pela posição/coordenadas que o ligante ocupa no espaço, geralmente o sítio ativo da enzima. Para a implementação dessa ideia nas simulações, o AGL é considerado um híbrido entre o AG e Algoritmo *Local Search* (LS) (Hart, 1994). Esse último algoritmo trabalha no sítio ativo da enzima, pré-determinado durante a preparação para o *docking*, calculando os valores energéticos a partir das coordenadas de cada átomo. A partir disso, pode-se entender que a grande diferença do Algoritmo Darwiniano (Evolucionário) para o Lamarckiano é uma inversão dos passos para chegar até a simulação da melhor *pose* representante do ligante (Morris *et al*. 1998). Enquanto os

Algoritmos Evolucionários passam pela seguinte sequência, *seleção* → *crossover* → *mutação* para chegar nas coordenadas atômicas do ligante e calcular a energia de ligação, o AGL, com auxílio do LS, inicia pela busca das coordenadas de cada átomo (fenótipo) seguindo, posteriormente, para os mesmos passos dos AE's.

As FE's utilizadas por esse *software* são as seguintes: baseadas fisicamente; semi-empíricas; Lennard-Jones; termo direcional de ligações de hidrogênio; potencial de Coulomb; potencial de desolvatação; termo entrópico proporcional ao número de ligações rotáveis (Gaillard, 2018).

3.6.3 AutoDock Vina (Vina)

A ferramenta *Vina* (Trott & Olson, 2010) é muito semelhante ao programa AD4, porém apresenta algumas diferenças em relação as FE's utilizadas.

O princípio de funcionamento do programa para posicionamento das *poses* é basicamente o mesmo utilizado pelo AD4, trabalhando com os mesmos algoritmos de busca analisados anteriormente, Genético e SA. A diferença entre os dois *softwares* é encontrada nas funções escore utilizadas por cada um. No *Vina*, a função escore é totalmente empírica contendo os seguintes termos: interação estérica Gaussiana; repulsão; interações hidrofóbicas e ligações de hidrogênio; termo entrópico proporcional ao número de ligações rotáveis (Trott & Olson, 2010). Porém, apesar de apresentar menos termos, quando comparado ao AD4, a acurácia e performance do *Vina* é superior. Em estudo comparativo realizado por Gaillard, 2018 entre as duas ferramentas, foi constatado que das 93% das *poses* geradas pelo *Vina* estavam presentes entre as três melhores *poses* formadas, além de outros estudos que demonstram boa capacidade do programa em resolver problemas biológicos diferentes (Masters *et al.* 2020; Russo & de Azevedo, 2019; de Ávila & de Azevedo, 2018).

3.6.4 Statistical Analysis of Docking Results and Scoring functions (SAnDReS)

O *software* SAnDReS é uma ferramenta que visa a análise dos resultados gerados pelos programas de *docking* (MVD, *AD4* e *Vina*) com o objetivo final de desenvolvimento de funções escore polinomiais (Xavier *et al.* 2016). O programa atua a partir de uma série de etapas em que os dados experimentais da proteína-alvo são utilizados como base para as análises estatísticas e a formação de novas funções escore a partir das já existentes (de Ávila *et al.* 2017).

As etapas utilizadas pelo *software* para o alcance do objetivo final, são as seguintes:

1. Pré-*Docking*: essa etapa objetiva a formação do *dataset* de trabalho com a(s) proteína(s)-alvo do estudo. Para o *download* das estruturas resolvidas é utilizado o banco de dados PDB, onde também é possível acessar dados experimentais das proteínas. Porém, nessa etapa também podem ser acessados bancos de dados como o *BindingDB*, MOAD e *PDBbind* na busca por dados de inibição como K_i e IC_{50}. A partir de uma série de filtros, definidos pelo usuário, chega-se no *dataset* final, em que encontramos as estruturas cristalográficas com dados experimentais de inibição e, entre essas estruturas presentes, aquela que apresenta melhor resolução. A estrutura que apresentar melhor resolução, será utilizada nos próximos estágios do processo.

2. *Docking Hub*: nesse momento do processo as simulações de *docking* são realizadas e os dados são armazenados para as análises estatísticas seguintes. Nessa etapa do processo através das análises estatísticas é possível encontrar quais protocolos de *docking* (específicos de cada *software*) apresentam maior acurácia no cálculo de RMSD e no posicionamento do ligante (*pose*). O protocolo selecionado a partir desses valores, será o utilizado nas etapas a seguir.

3. *Ensemble docking*: nessa fase do estudo, o protocolo selecionado na etapa anterior será utilizado para replicar a simulação de *docking* nas estruturas restantes do *dataset* formado na fase de pré-*docking*. Durante essa fase é possível perceber se o protocolo selecionado realmente é capaz de simular as interações existentes entre a proteína-alvo e seus possíveis ligantes.

4. *Structural Parameters*: esse recurso visa identificar se há alguma relação entre os dados obtidos no *ensemble* com as informações estruturais presentes nas estruturas participantes do *dataset*. O programa tem a capacidade de analisar 100

parâmetros estruturais de cada uma das estruturas trabalhadas no projeto (Xavier *et al*. 2016).

5. *Scoring functions*: o objetivo dessa fase é testar a performance que as funções escores apresentam em predizer a afinidade de ligação existente entre a proteína e o ligante utilizado na resolução da estrutura cristalográfica. As funções que apresentarem os melhores resultados de correlação são selecionadas para participar da formação das funções escore polinomiais na próxima etapa do estudo.

6. *Polynomial Scoring Functions*: na formação de novas funções escore polinomiais são utilizados "Métodos de Aprendizado de Máquina" (SML). O *SAnDReS* emprega as seguinte técnicas de SML: Lasso, Lasso CV, Ridge, Ridge CV, Elastic Net CV. Os polinômios formados possuem como termos as funções escore com os maiores valores de correlação nas fases anteriores, sendo construídos por três variáveis independentes. A função básica do *software* é encontrar os pesos para os termos da equação 6 (Xavier *et al*. 2016),

$$\begin{aligned} \mathrm{score} = {} & \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \\ & \omega_4 x_1 x_2 + \omega_5 x_1 x_3 + \omega_6 x_2 x_3 + \\ & \omega_7 x_1^2 + \omega_8 x_2^2 + \omega_9 x_3^2 \end{aligned}$$

Equação 6

onde, *score* é o valor da função escore, ω_0 é a constante de regressão e os outros ω's são os pesos para cada variável independente da equação. Desde que apresente os nove termos, teremos 511 possibilidades de polinômios gerados. Na validação das funções construídas, elas são expostas a dois conjuntos de moléculas, o *conjunto treino* e o *conjunto teste*. O *conjunto treino* contém 70% das moléculas do *dataset* original visando a formação de um modelo de análise. Já o *conjunto teste* contém os 30% restantes das estruturas e tem por objetivo validar o modelo estabelecido pelo *conjunto treino* (Cichero *et al*. 2010). A avaliação da performance das funções polinomiais é dada a partir dos coeficientes de correlação de Sperman (Zar, 1972).

7. *Decoys and Actives*: essa é a última etapa de análise que o *software* realiza sobre o *dataset*. É aqui que as funções escores, tanto as utilizadas nos simuladores de *docking*, quanto as polinomiais formadas pelo próprio *SAnDReS* tem sua capacidade de reconhecimento dos reais ligantes da proteína-alvo testada. Para a realização desse objetivo as funções escores são utilizadas contra um conjunto de moléculas, em que estão presentes "falsos ligantes" (aqui chamados de "*decoys*", ou

seja, moléculas que reconhecidamente não possuem alto valor de afinidade pelo sítio da enzima) e “ligantes ativos” (chamados “*actives*”, ligantes que sabidamente apresentam afinidade pela proteína-alvo, retirados das estruturas cristalográficas utilizadas no estudo). O arquivo de *decoys* é montado de forma randômica pelo *software* DUD-E (Mysinger *et al*. 2012). Com o arquivo pronto, é rodado um pequeno VS, utilizando-se a estrutura com melhor resolução e o protocolo com os melhores valores de RMSD (ambos utilizados nas simulações de *docking*). As funções com capacidade de predição da afinidade entre proteína e ligante devem ser capazes de encontrar o maior número de *actives* em meio aos *decoys* presentes no arquivo. Para avaliar as funções que apresentaram melhor performance, são utilizados dois parâmetros, os valores de a*rea under the curve* (AUC) associados a curva *receiver operating characteristics* (ROC) e o *Enrichment Factor* (EF) (Brooijmans & Kuntz, 2003). O AUC e a curva ROC medem a capacidade de reconhecimento dos ligantes positivos e o EF a porcentagem de ligantes verdadeiros presente em 1% (EF1), 2% (EF2), 5% (EF5), 10% (EF10) e 20% (EF20) da amostra total do conjunto de *decoys and actives* (Heck *et al*. 2017; de Ávila *et al.* 2017; de Ávila & de Azevedo, 2018).

CAPÍTULO 4

4. RESULTADOS E DISCUSSÕES

4.1 *TRANS-2*-ENOIL (ACP) REDUTASE (InhA)

4.1.1 Pré-*Docking*

Como já descrito anteriormente, para essa fase do desenvolvimento do projeto foi construído um *dataset* contendo estruturas da enzima com dados de inibição para IC_{50}. Dentre as estruturas presentes no conjunto de dados, a que apresentou melhor resolução e, consequentemente, foi escolhida para a validação dos protocolos de *docking* foi a estrutura 4TZK (He *et al*. 2006). A estrutura tem uma resolução de 1,62Å, sendo a menor entre todas as componentes do *dataset* e, por consequência, a que mais se aproxima da conformação *in vivo* da enzima aqui estudada.

4.1.2 Re-Docking

Para identificar a capacidade de simulação do sistema proteína-ligante pelos softwares são considerados dois parâmetros, Acurácia do *docking* 1 (DA1) e Acurácia do *docking* 2 (DA2). DA1 representa a porcentagem de *poses* geradas com RMSD até 2Å e DA2 a porcentagem de *poses* até 3Å. Os valores de acurácia tiveram uma grande variação entre os *softwares* utilizados na tentativa de simular o complexo proteína-ligante. Enquanto para o *Vina* foi obtido um DA1 de 11,905% e DA2 igual a 14,286%, os valores indicados para o *AD4* foram 95,000% tanto para DA1, quanto para DA2. Os valores de RMSD calculados por cada uma das funções escore/termos energéticos deixa ainda mais clara a diferença encontrada entre os dois softwares. Para o *Vina* o RMSD variou de 2,20Å, até 7,82Å. Já para o *AD4* foram encontrados valores entre 1,16Å e 2,88Å. Na tabela 4 estão descritos os valores para cada uma das funções escore/termos energéticos apresentados pelos programas envolvidos. Entre as funções do *Vina*, o termo *Affinity* apontou um RMSD de 2,20Å, porém com baixos valores de correlação (ρ = 0,184; *p-value* < 0,5). Um aspecto interessante dos observados para o *Vina* é o alto valor de RMSD para Repulsion (6,625Å), mas índices

significativos de correlação para essa função (ρ = - 0,587; *p-value* < 0,01) (Figura 9). Apesar dos valores elevados de RMSD a correlação alta para essa função e um ρ negativo pode indicar a importância de poucas interações repulsivas entre enzima e ligante na simulação do sistema.

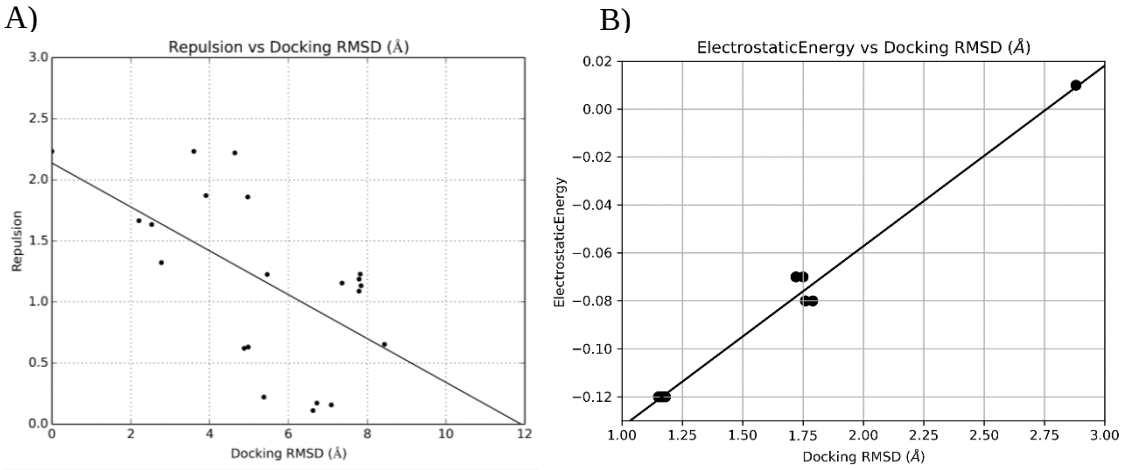
Nas simulações com o *AD4*, a função/termo que mais se sobressaiu em resultados foi o *Electrostatic Energy* (RMSD = 1,16 Å) junto de coeficientes de correlação significativos estatisticamente, tanto para ρ (0,765 e *p-value* < 0,01) e para R^2 (0,986 e *p-value* < 0,001) (Tabela 4), indicando uma grande importância do termo energético, e desse tipo de interação, na simulação das interações entre enzima e ligante.

Tabela 4. Resultados de *re-docking* para a estrutura 4TKZ [He *et al*. 2006] utilizando o protocolo do *software AutoDock Vina* e *AutoDock 4.*

Funções Escore	**RMSD (Å)**	**ρ**	**p-value1**	**R^2**	**p-value2**
Affinity[a]	2,208	0,184	$4{,}236.10^{-01}$	0,045	$3{,}544.10^{-01}$
Gauss1[a]	5,463	-0,329	$1{,}459.10^{-01}$	0,139	$9{,}662.10^{-02}$
Gauss2[a]	5,378	-0,461	$3{,}452.10^{-02}$	0,120	$1{,}236.10^{-01}$
Repulsion[a]	6,625	-0,587	$5{,}149.10^{-03}$	0,343	$5{,}254.10^{-03}$
Hydrophobic[a]	6,724	-0,210	$3{,}600.10^{-01}$	0,062	$2{,}755.10^{-01}$
Hydrogen[a]	7,822	-0,382	$8{,}748.10^{-02}$	0,140	$9{,}416.10^{-02}$
Free Energy[b]	2,880	0,427	$2{,}186.10^{-01}$	0,069	$4{,}636.10^{-01}$
Final Intermolecular[b]	2,880	0,427	$2{,}186.10^{-01}$	0,074	$4{,}476.10^{-01}$
vdW+Hbond+desenvolv Energy[b]	2,880	0,432	$2{,}129.10^{-01}$	0,143	$2{,}807.10^{-01}$
Electrostatic Energy[b]	1,160	0,765	$9{,}922.10^{-03}$	0,986	$1{,}052.10^{-08}$
Final Total Internal Energy[b]	1,760	- 0,330	$3{,}513.10^{-01}$	0,000	$9{,}858.10^{-01}$

“a” representa as funções escore, ou termos energéticos do *Vina*.
“b” representa as funções escore, ou termos energéticos do *AutoDock4*.

Figura 9. Representação do resultado de *re*-docking para o *Vina* e *AD4*. A) 21 *poses* geradas com os resultados do protocolo de *docking* do *Vina*. B) 10 *poses* geradas com os resultados do protocolo de *docking* do *AD4*. Cada círculo representa uma *pose*. (au: unidades arbitrárias).



Fonte: próprio autor.

Seguindo os próximos passos após o *re-docking*, o protocolo vigente foi utilizado contra todas as outras estruturas presentes no *dataset* com objetivo de validação com ligantes distintos. Em relação aos valores de acurácia anteriormente encontrados para a estrutura 4TZK, houve uma significativa melhora com DA1 = 28,125% e DA2 = 29,688%. Olhando com mais atenção para os RMSD's de cada um dos ligantes, a elevação de valores, muito provavelmente, deve-se a presença das estruturas 1L2Z (DA1 = 33,333% e DA2 = 34,524) e 2NSD (DA1 = 30,952% e DA2 = 35,714%). Esses valores, junto ao RMSD de 1,744Å calculado pela função *Hydrogen*, indicam uma certa habilidade das funções escore em compreender e tentar simular as interações realizadas entre proteína e ligantes distintos. Quanto as correlações entre as funções escore nenhuma apresentou resultado elevados como observados anteriormente (Tabela 5).

Nas simulações realizadas pelo *AD4*, foram obtidos valores mais significativos para as estruturas presentes no conjunto. Tanto DA1, quanto DA2 obtidos ficaram acima de 50%, sendo esses valores mais elevados que os gerados pelo *Vina*. No entanto, nenhuma das funções escore/termos energéticos foi capaz de gerar valores de RMSD médio abaixo do limite de 2Å, sendo observados valores relativamente baixos para ρ, R^2, *p-value1* e *p-value2* (Tabela 5).

Tabela 5. Resultados de *re-docking* para as 32 estruturas* presentes no *dataset*.

Funções Escore	RMSD (Å)	ρ	p-value1	R^2	p-value2
Affinity[a]	2,006	0,088	$6{,}321.10^{-01}$	0,019	$4{,}484.10^{-01}$
Gauss1[a]	5,993	-0,088	$6{,}307.10^{-01}$	0,007	$6{,}418.10^{-01}$
Gauss2[a]	5,993	0,033	$8{,}593.10^{-01}$	0,019	$4{,}462.10^{-01}$
Repulsion[a]	5,993	0,216	$2{,}345.10^{-01}$	0,001	$8{,}796.10^{-01}$
Hydrophobic[a]	5,993	-0,208	$2{,}528.10^{-01}$	0,037	$2{,}923.10^{-01}$
Hydrogen[a]	1,744	-0,021	$9{,}113.10^{-01}$	0,004	$7{,}313.10^{-01}$
Free Energy[b]	3,150	0,285	$1{,}265.10^{-01}$	0,083	$1{,}233.10^{-01}$
Final Intermolecular[b]	2,390	0,281	$1{,}322.10^{-01}$	0,080	$1{,}303.10^{-01}$
vdW+Hbond+desenvolv Energy[b]	2,390	0,287	$1{,}247.10^{-01}$	0,081	$1{,}265.10^{-01}$
Electrostatic Energy[b]	3,150	0,108	$5{,}706.10^{-01}$	0,000	$9{,}665.10^{-01}$
Final Total Internal Energy[b]	3,830	0,060	$7{,}520.10^{-01}$	0,000	$9{,}466.10^{-01}$
Torsional Free Energy[b]	4,950	-0,125	$5{,}114.10^{-01}$	0,002	$7{,}947.10^{-01}$

* Nas análises de *ensembledock* para o *AD4*, foram retiradas duas estruturas (2FOI (Aksyuk *et al*. 2009), 3FNG (Freundlich *et al*. 2009)) por apresentarem valores de RMSD extremamente elevados e potencializarem uma visão geral errônea dos dados.
"a" representa as funções escore, ou termos energéticos do *Vina*.
"b" representa as funções escore, ou termos energéticos do *AutoDock4*.

4.1.3 Funções Escore

A partir dos dados obtidos com as análises anteriores, a próxima etapa do estudo objetiva esclarecer quais funções escore do *Vina* são mais indicadas na elaboração de uma função polinomial tendo como base os valores de correlação explicitados a seguir. Nesse momento, é testada a capacidade de cada função escore na predição de afinidade entre cada uma das trinta e duas estruturas e seus respectivos ligantes, levando em consideração os valores energéticos calculados. A função/termo que apresentou melhor valor de correlação foi a *Affinity* (ρ = 0,466; *p-value1* < 0,01) em relação as demais funções utilizadas na predição da afinidade. Entre as demais funções escore/termos energéticos, Gauss1 apresentou os melhores valores, porém com um *p-value* mais elevado que o exigido para modelos biológicos (ρ = - 0,301; *p-value1* = 0,09) (Tabela 6).

Tabela 6. Análise estatística do poder de predição de afinidade das funções escore presentes no *Vina* e *AD4*.

Funções Escore	ρ	p-value1	R^2	p-value2
Affinity[a]	0,466	$7{,}155.10^{-03}$	0,012	$5{,}484.10^{-01}$
Gauss1[a]	-0,301	$9{,}429.10^{-02}$	0,048	$2{,}286.10^{-01}$
Gauss2[a]	-0,284	$1{,}147.10^{-01}$	0,025	$3{,}911.10^{-01}$
Repulsion[a]	-0,159	$3{,}860.10^{-01}$	0,005	$6{,}876.10^{-01}$
Hydrophobic[a]	-0,154	$4{,}012.10^{-01}$	0,013	$5{,}416.10^{-01}$
Hydrogen[a]	0,030	$8{,}700.10^{-01}$	0,000	$9{,}552.10^{-01}$
Free Energy[b]	0,271	$1{,}478.10^{-01}$	0,050	$2{,}331.10^{-01}$
Final Intermolecular[b]	0,338	$6{,}772.10^{-02}$	0,079	$1{,}334.10^{-01}$
vdW+Hbond+desenvolv Energy[b]	0,336	$6{,}967.10^{-02}$	0,081	$1{,}284.10^{-01}$
Electrostatic Energy[b]	-0,071	$7{,}098.10^{-01}$	0,000	$9{,}243.10^{-01}$
Final Total Internal Energy[b]	0,229	$2{,}238.10^{-01}$	0,024	$4{,}185.10^{-01}$
Torsional Free Energy[b]	-0,300	$1{,}078.10^{-01}$	0,063	$1{,}802.10^{-01}$

"a" representa as funções escore, ou termos energéticos do *Vina*.
"b" representa as funções escore, ou termos energéticos do *AutoDock4*.

Seguindo como critério de escolha os maiores fatores de correlação, como já citado, as funções selecionadas para compor as diversas combinações de funções escore polinomiais foram as seguintes: *Affinity*, *Gauss1* e *Gauss2*. Após a elaboração das 511 combinações possíveis estabelecidas pelo *SAnDReS*, as funções foram testadas perante os dois conjuntos de estruturas organizados aleatoriamente, *training set* (vinte e duas estruturas) e *test set* (dez estruturas). Para o conjunto treino, os polinômios elaborados apresentaram valores de correlação entre 0,400 e 0,588, enquanto as funções escore clássicas variaram de – 0,233 até 0,479 (Tabela 7). No conjunto treino, a função escore polinomial que apresentou a melhor performance foi a *Polscore231* que é conformada pela seguinte equação:

$$Polscore231 = -5.819376 + -0.098942*(x) + 0.006167*(y) + 0.001186*(z.x) + -0.001946*(z^2) + 0.000455*(x^2) + -0.000002*(y^2)$$

onde, x representa *Gauss1*, y é *Gauss2* e z substitui o termo *Affinity*. A *Polscore231* apresentou valores satisfatórios e mais elevados ($\rho = 0{,}709$; *p-value1* < 0,03) (Tabela 7) (Figura 10) na predição de afinidade da enzima com possíveis ligantes que as demais funções clássicas.

Tabela 7. Valores de correlação para as funções escore clássicas do *Vina* e da função polinomial *Polscore231* na predição de afinidade das estruturas presentes no *training* e *test set*.

Funções Escore	**ρ (*training set)***	**p-value**	**ρ**	**p-value (*test set)***
Affinity	0,479	$2{,}407.10^{-02}$	0,382	$2{,}763.10^{-01}$
Gauss1	-0,190	$3{,}975.10^{-01}$	- 0,588	$7{,}388.10^{-02}$
Gauss2	-0,233	$2{,}960.10^{-01}$	- 0,345	$3{,}282.10^{-01}$
Repulsion	-0,034	$8{,}810.10^{-01}$	- 0,636	$4{,}791.10^{-02}$
Hydrophobic	-0,051	$8{,}203.10^{-01}$	- 0,406	$2{,}443.10^{-01}$
Hydrogen	0,122	$5{,}872.10^{-01}$	- 0,239	$5{,}061.10^{-01}$
Polscore231	0,519	$1{,}328.10^{-02}$	0,709	$2{,}167.10^{-02}$

Entre as funções escore/termos energéticos apresentados pelo *AD4* para a tentativa de simulação do sistema biológico, os coeficientes de correlação encontrados ficaram abaixo dos calculados pelo *SAnDRes* para o *Vina*, com os valores girando entre -0,300 e 0,338. Porém, os *p-value's* obtidos para ρ em relação as duas funções/termos ficaram próximos de valores significativos para sistema biológicos (*vdW+Hbond+desenvolv Energy* (*p-value1* < 0,07) e *Final Intermolecular Energy* (*p-value1* < 0,07)) (Tabela 6). Seguindo esses valores, os dois termos energéticos se candidataram, junto ao *Total Free Energy* (ρ = -0,300; *p-value1* = 0,1), a participar da elaboração de funções escore polinomiais construídas por meio do *SAnDRes* a partir dos métodos de aprendizado de máquina presentes no *software*. Para os termos energético apresentados pelo *AD4*, o método que apresentou as melhores performances foi o *Ridge*. Quando confrontadas com o conjunto treino (21 estruturas), as funções polinomiais indicaram maiores coeficientes de correlação (0,270 até 0,336) que os termos clássicos do *AD4* (-0,119 até 0,203). Na predição de afinidade contra um conjunto menor de estruturas (teste (9 estruturas)), a função polinomial *Polscore345* apresentou melhores resultados (ρ = 0,717; *p-value1* < 0,03) em comparação aos termos clássicos (Tabela 8 e Figura 10) e à função elaborada com base nos termos energéticos do *Vina* (*Polscore231*) (Tabela 7).

Tabela 8. Valores de correlação para as funções escore clássicas do *AD4* e da função polinomial *Polscore345* na predição de afinidade das estruturas presentes no *training* e *test set.*

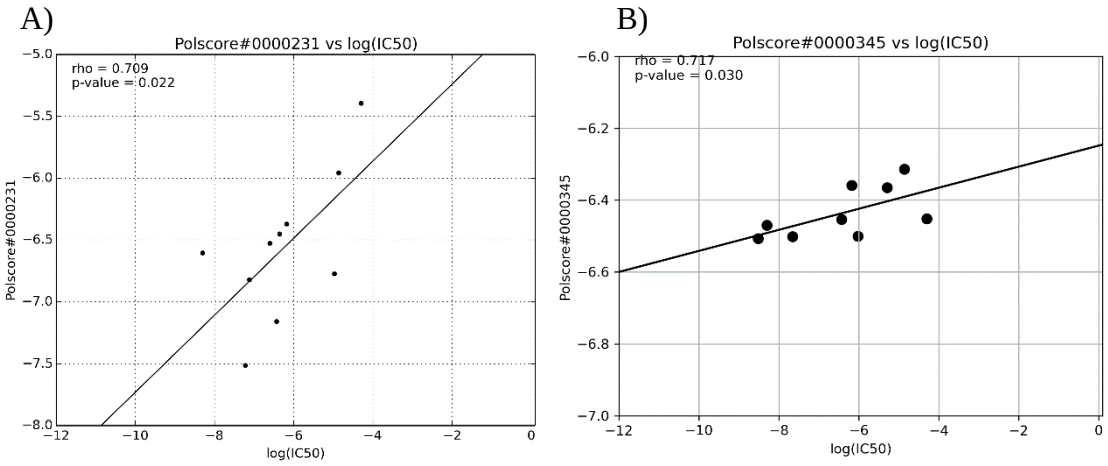
Funções Escore	**ρ (*training set*)**	**p-value (*training set*)**	**ρ (*teste set*)**	**p-value (*test set*)**
Free Energy	0,194	$4{,}003.10^{-01}$	0,600	$8{,}762.10^{-02}$
Final Intermolecular	0,203	$3{,}781.10^{-01}$	0,483	$1{,}875.10^{-01}$
vdW+Hbond+desenvolv Energy	0,198	$3{,}907.10^{-01}$	0,483	$1{,}875.10^{-01}$
Electrostatic Energy	-0,119	$6{,}070.10^{-01}$	- 0,033	$9{,}319.10^{-01}$
Final Total Internal Energy	-0,103	$6{,}577.10^{-01}$	0,711	$3{,}166.10^{-02}$
Torsional Free Energy	-0,104	$6{,}539.10^{-01}$	- 0,487	$1{,}832.10^{-01}$
Polscore345	0,288	$2{,}058.10^{-01}$	0,717	$2{,}982.10^{-02}$

A equação polinomial *Polscore345* segue a seguinte fórmula

$$Polscore345 = -6.413391 + -0.028074*(x) + -0.065376*(y) + - 0.149268*(x.y) + 0.188724*(z.y) + 0.138307*(y^2)$$

onde, *x* indica *Final Intermolecular Energy*, *y Torsional Free Energy* e *z* corresponde a *vdW+Hbond+desolv Energy*.

Figura 10. Gráficos de dispersão das três equações que apresentaram os melhores resultados de *test set* em relação ao log(IC_{50}). Cada ponto no gráfico representa o log(Ki) calculado pela equação para cada estrutura da enzima. A) *Polscore231* (*Vina*); B) *Polscore345* (*AD4*).



Fonte: próprio autor.

A tabela 9 demonstra os valores de predição de logIC_{50} das funções polinomiais para cada uma das estruturas e ligantes participantes do conjunto de InhA. É possível confirmar o poder de predição de afinidade, já observado nos valores de correlação, das equações *Polscore#231* e *Polscore#345*.

Analisando-se os polinômios formados a partir das funções/termos clássicos apresentados por cada um dos *softwares* de *docking*, é possível identificar os fatores de maior relevância a serem considerados na predição de afinidade da enzima em relação a diferentes tipos de ligantes. As funções polinomiais que apresentaram melhores resultados serem formadas por termos como *Final Intermolecular Energy*, *Torsional Free Energy* e *vdW+Hbond+desolv Energy* (*AD4*) e *Affinity* (*Vina*) é um indicativo da importância dessas interações no processo inibitório da enzima. As funções Gauss1 e Gauss2 consideram como fator importante a distância entre os átomos do ligante e do sítio da enzima, sendo também indicado como fator relevante na inibição. A INH, importante inibidora da InhA de *M. tuberculosis*, quando associada ao $NADH^+$ e ligada ao sítio enzimático realiza um diverso e grande número de interações com os resíduos participantes do sítio ativo (de Ávila *et al*. 2020), fortalecendo, de maneira prática, ainda mais a hipótese, a partir da análise dos polinômios, a importância desses termos para simular/prever o processo inibitório.

Tabela 9. Valores de logIC_{50} experimental e predito para todas* as estruturas de InhA.

Código Acesso PDB	Código Acesso Ligante	Resolução (Å)	IC_{50} (nM)	log(IC_{50})	log(IC_{50}) predito[1]	log(IC_{50}) predito[2]
1I2Z	654	2,80	13700	-4,863	-5,957	-6,314
1LXC	AYM	2,40	370	-6,431	-7,158	-6,453
1MFP	IDN	2,33	70	-7,154	-7,350	-6,500
1P44	GEQ	2,70	200	-6,698	-6,794	-6,250
1QSG	TCL	1,75	665	-6,177	-6,371	-6,359
2B37	8PS	2,60	5	-8,301	-6,605	-6,470
2FOI	JPA	2,50	440	-6,356	-6,451	ND*
2NSD	4PI	1,90	5160	-5,287	-5,693	-6,365
2NV6	ZID	1,90	323	-6,490	-7,341	-6,250
2OL4	JPN	2,20	440	-6,356	-5,819	-6,087
2OOS	JPJ	2,10	76	-7,119	-6,820	-6,542
2X23	TCU	1,80	22	-7,657	-7,152	-6,501
3AM4	FT1	2,30	97	-7,013	-6,486	-6,389
3FNE	8PC	1,98	29	-7,537	-6,685	-6,466
3FNF	JPM	2,30	51	-7,292	-6,701	-6,475
3FNG	JPL	1,97	110	-6,958	-7,212	ND*
3OJF	IMJ	2,20	60	-7,221	-7,513	-6,492
3ZU4	ZU4	2,01	28500	-4,545	-5,783	-6,464
3ZU5	AEW	2,00	50500	-4,296	-5,394	-6,451
4BII	PYW	1,95	6000	-5,221	-5,859	-6,513
4BQP	VMY	1,89	3	-8,522	-7,301	-6,506
4BQR	IBH	2,05	200	-6-698	-6,028	-6,450
4COD	KV1	2,40	34	-7,468	-7,619	-6,552
4D44	JA3	1,80	13	-7,886	-6,722	-6,394
4IGE	CHJ	2,15	250	-6,602	-6,526	-6,326
4IGF	CHV	2,30	6000	-5,221	-6,011	-6,336
4Q9N	0WE	1,79	950	-6,022	-7,678	-6,500
4TRJ	665	1,73	890	-6,050	-7,222	-6,403
4TZK	641	1,62	390	6,408	-7,067	-6,413
4TZT	468	1,86	23120	-4,636	-6,644	-6,406
4U0J	566	1,62	10660	-4,972	-6,772	-6,404
4U0K	744	1,90	970	-6,013	-6,127	-6,410

[1] Predição de log(IC_{50}) pela equação *Polscore#231 (Vina)*

[2] Predição de log(IC_{50}) pela equação *Polscore#345 (AD4)*

*ND: as estruturas 2NSD e 3FNG foram consideradas *outliers* nas simulações de *re-docking* com o *software AD4*.

4.2 QUINASE DEPENDENTE DE CICLINA 2 (CDK2)

4.2.1 Pré-*Docking*

Nessa fase de preparação para a realização de *docking* para a enzima alvo, foi escolhida a estrutura que servirá de base para as simulações. O critério adotado para todas as enzimas utilizadas neste projeto, foi a que apresentou melhor resolução cristalográfica entre as estruturas presentes no *dataset*. Para os estudos com a CDK2 foi seguido um modelo diferente para desenvolvimento, validação e testagem das funções escore clássicas e das funções escore polinomiais elaboradas por meio do *SAnDReS.* No início foram organizados dois *datasets*, o primeiro ($HRIC_{50}$) foi composto de estruturas com informação de IC_{50} e que apresentam resolução cristalográfica menor que 1,5Å. O segundo ($CDK2IC_{50}$) composto por estruturas de CDK2 com resolução menor que 2,0Å. Nenhuma das estruturas participou dos dois *datasets* (Tabela 1).

As estruturas presentes em $HRIC_{50}$ foram utilizadas no estabelecimento do melhor protocolo e na determinação das melhores funções escore para estruturas com valor de inibição IC_{50}. O conjunto com estruturas de CDK2 foi utilizado para testagem das funções escore clássicas e polinomiais na predição de afinidade do complexo proteína-ligante. Dentre as estruturas participantes do conjunto $HRIC_{50}$, seguindo o critério de melhor resolução cristalográfica, a escolhida foi a de código PDB 1US0 (Howard *et al.* 2004), por apresentar uma resolução de 0,66Å.

4.2.2 Re-Docking

Nas simulações de *re-docking* a estrutura cristalográfica escolhida na fase anterior foi empregada aos 32 protocolos presentes no *software* MVD (Anexo A). Para cada um dos protocolos foram calculadas as posições de 1000 *poses* sendo identificada a melhor performance do protocolo 31 (*Plants Score* e *Iterated Simplex with Ant Colony Optimization*). O valor de RMSD calculado por esse protocolo para a estrutura 1US0 foi de 0,594Å indicando uma ótima performance para essa enzima, sendo que para simulações de *docking* contra ligantes não-covalentes o valor de RMSD deve estar abaixo de 2Å. A partir das análises de correlação entre a função

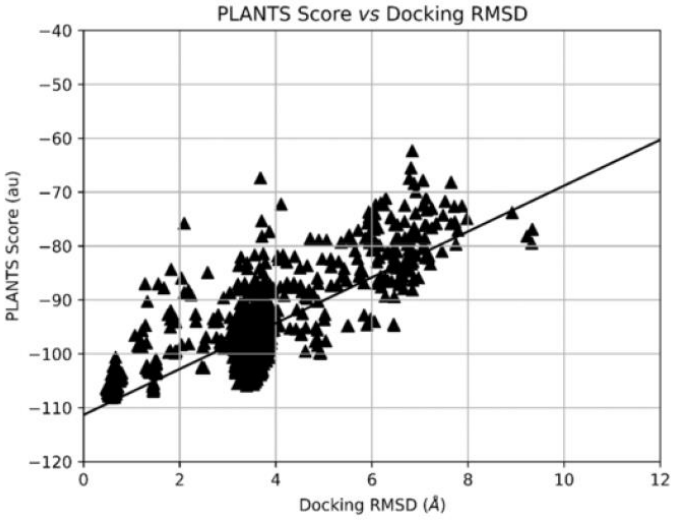
escore (*Plantas Score*) e o valor de RMSD percebeu-se para esse protocolo os maiores valores (ρ = 0.673 e *p-value1* < 0.001; R^2 = 0.563 e *p-value2* < 0.001) (Tabela 10 e Figura 11).

Tabela 10. Resultados de *re-docking* para a estrutura 1US0 [Howard *et al*. 2004] utilizando o protocolo 31. (Termo 1: Interaction; Termo 2: Cofactor; Termo 3: Protein; Termo 4: Water; Termo 5: Internal; Termo 6: Electro; Termo 7: ElectroLong; Termo 8: HBond).

Funções Escore e Termos Energéticos	**RMSD (Å)**	**ρ**	***p-value***	**R^2**	***p-value***
PLANTS Score	0,594	0,673	$9{,}329.10^{-133}$	0,563	$1{,}435.10^{-181}$
MolDock Score	0,663	0,713	$5{,}417.10^{-156}$	0,601	$1{,}350.10^{-201}$
Rerank Score	0,633	0,771	$1{,}449.10^{-197}$	0,594	$1{,}499.10^{-197}$
Termo 1	1,460	0,748	$3{,}624.10^{-180}$	0,653	$1{,}517.10^{-231}$
Termo 2	1,281	0,408	$1{,}942.10^{-41}$	0,155	$2{,}304.10^{-38}$
Termo 3	6,660	0,763	$5{,}141.10^{-191}$	0,673	$1{,}599.10^{-244}$
Termo 4[1]	ND[2]	ND	ND	ND	ND
Termo 5	7,045	-0,017	$5{,}810.10^{-01}$	0,001	$3{,}192.10^{-01}$
Termo 6	1,069	0,597	$1{,}546.10^{-97}$	0,400	$7{,}337.10^{-113}$
Termo 7	1,892	0,649	$2{,}151.10^{-120}$	0,315	$5{,}918.10^{-84}$
Termo 8	1,445	0,529	$4{,}020.10^{-73}$	0,326	$1{,}569.10^{-87}$
Ligand Efficiency 1	0,663	0,713	$5{,}451.10^{-156}$	0,601	$1{,}349.10^{-201}$
Ligand Efficiency 3	0,663	0,771	$1{,}448.10^{-197}$	0,594	$1{,}499.10^{-197}$
Docking Score	0,594	0,673	$9{,}329.10^{-133}$	0,563	$1{,}435.10^{-181}$
Displaced Water Score	7,182	-0,163	$2{,}099.10^{-07}$	0,070	$2{,}223.10^{-17}$

[1]Termo 4 representa a interação enegética entre o ligante e as moléculas de água (Water) (Thomsen & Christensen, 2006), e não foi determinado para as simulações de *re*-docking da estrutura 1US0 [Howard, 2004].
[2]ND: Não determinado.

Figura 11. Representação do resultado de 1000 *poses* geradas com os resultados do melhor protocolo de *docking*. Cada triângulo representa uma *pose*. (au: unidades arbitrárias).



Fonte: próprio autor.

A acurácia do *docking* (DA1) também foi calculada apresentando um valor de 63,9%, indicando a capacidade do protocolo em posicionar as *poses* no sítio de ligação da enzima.

Após a determinação do melhor protocolo com base nos valores de RMSD e correlação estabelecidos, o próximo passo foi a validação do protocolo através do *ensemble docking*. Nesse momento o protocolo 31 foi validado contra as estruturas restantes do *dataset* estabelecido nos passos anteriores. Como cada estrutura utilizada aqui apresenta um tipo de ligante distinto, é possível avaliar se o protocolo utilizado é realmente capaz de compreender e simular a diversidade de interações realizadas entre as enzimas e os seus ligantes. Os resultados obtidos com as simulações de *ensemble docking* indicaram como sendo a melhor função escore *PLANTS Score* com um RMSD de 0,346Å (ρ = 0.556 e *p-value1* < 0.001; R2 = 0.336 e *p-value2* < 0.001) (Tabela 11). Com os valores de RMSD encontrados para essa função escore nas simulações com todas as estruturas participantes do $HRIC_{50}$, é visível e validada a capacidade da *PLANTS Score* de simular os sistemas proteína-ligante de forma aproximada a realidade, além de calcular os valores energéticos gerados a partir dessa interação.

Tabela 11. Resultados de *docking* para o protocolo 31 com todas as estruturas do *dataset* $HRIC_{50}$. (Termo 1: Interaction; Termo 2: Cofactor; Termo 3: Protein; Termo 4: Water; Termo 5: Internal; Termo 6: Electro; Termo 7: ElectroLong; Termo 8: HBond).

Funções Escore e Termos Energéticos	RMSD (Å)	ρ	p-value	R^2	p-value
PLANTS Score	0,346	0,556	$1,898.10^{-15}$	0,336	$6,281.10^{-17}$
MolDock Score	0,346	0,430	$3,585.10^{-09}$	0,126	$1,599.10^{-06}$
Rerank Score	0,346	0,469	$7,316.10^{-11}$	0,160	$4,823.10^{-08}$
Termo 1	0,346	0,382	$2,102.10^{-07}$	0,086	$8,791.10^{-05}$
Termo 2	1,157	-0,071	$3,548.10^{-01}$	0,007	$2,899.10^{-01}$
Termo 3	0,346	0,392	$1,003.10^{-07}$	0,077	$2,158.10^{-04}$
Termo 4[1]	ND[2]	ND	ND	ND	ND
Termo 5	0,346	0,118	$1,229.10^{-01}$	0,008	$2,431.10^{-01}$
Termo 6	5,460	0,052	$4,972.10^{-01}$	0,006	$2,969.10^{-01}$
Termo 7	5,460	0,023	$7,654.10^{-01}$	0,004	$4,099.10^{-01}$
Termo 8	0,183	0,286	$1,403.10^{-04}$	0,057	$1,526.10^{-03}$
Ligand Efficiency 1	11,624	0,389	$1,181.10^{-07}$	0,069	$4,999.10^{-04}$
Ligand Efficiency 3	11,624	0,439	$1,559.10^{-09}$	0,094	$3,929.10^{-05}$
Docking Score	0,346	0,556	$1,898.10^{-15}$	0,336	$6,281.10^{-17}$
Displaced Water Score	2,812	0,080	$2,948.10^{-01}$	0,055	$1,850.10^{-03}$

[1]Termo 4 representa a interação energética entre o ligante e as moléculas de água (Water) (Thomsen & Christensen, 2006), e não foi determinado para as estruturas presentes no *dataset*. [2]ND: Não determinado.

4.2.3 Funções Escore

As análises de correlação entre as funções escore presentes no MVD e o log(IC_{50}) das estruturas constituintes da base de dados encontraram baixa relação entre os dados experimentais e os de predição de afinidade. A maior correlação encontrada foi o termo energético que considera a energia de interação entre o ligante e a proteína-alvo (ρ = 0.312 e *p-value* < 0.001). Também foram obtidos valores *p-value* < 0,05 para *PLANTS*, *MolDock* e *Re-rank Score* (Tabela 12).

Tabela 12. Análise estatística do poder de predição de afinidade das funções escore e termos energéticos para todas as estruturas do *dataset* $HRIC_{50}$. (Termo 1: Interaction; Termo 2: Cofactor; Termo 3: Protein; Termo 4: Water; Termo 5: Internal; Termo 6: Electro; Termo 7: ElectroLong; Termo 8: HBond).

Scoring Functions and Energy Terms[1]	**ρ**	***p-value***	**R^2**	***p-value***
PLANTS Score	0,245	$1,137.10^{-03}$	0,002	$5,407.10^{-01}$
MolDock Score	0,283	$1,621.10^{-04}$	0,002	$5,378.10^{-01}$
Rerank Score	0,198	$9,053.10^{-03}$	0,000	$8,824.10^{-01}$
Termo 1	0,312	$2,979.10^{-05}$	0,010	$1,854.10^{-01}$
Termo 2	0,152	$4,633.10^{-02}$	0,001	$6,477.10^{-01}$
Termo 3	0,290	$1,100.10^{-04}$	0,059	$1,240.10^{-03}$
Termo 4	0,189	$1,255.10^{-02}$	0,009	$2,160.10^{-01}$
Termo 5	-0,024	$7,549.10^{-01}$	0,019	$6,922.10^{-02}$
Termo 6	-0,143	$9,523.10^{-01}$	0,000	$9,400.10^{-01}$
Termo 7	0,124	$6,102.10^{-02}$	0,002	$5,916.10^{-01}$
Termo 8	-0,121	$1,032.10^{-01}$	0,001	$6,659.10^{-01}$
Ligand Efficiency 1	-0,011	$1,119.10^{-01}$	0,027	$3,196.10^{-02}$
Ligand Efficiency 3	-0,079	$8,906.10^{-01}$	0,000	$9,338.10^{-01}$

[1] Os valores dos termos energéticos foram calculados usando a posição cristalográfica dos ligantes.

Na construção de funções escores polinomiais, intermediada pelo *SAnDReS*, foram considerados os termos energéticos presentes no MVD (Tabela 3). Para a elaboração dos polinômios, considerou-se os termos energéticos dos *softwares* MVD, *AD4* e *AutoDockVina*. Na elaboração dos polinômios, são criados grupos de três termos, gerando a possibilidade de 56 combinações distintas. Para cada um dos 56 polinômios formados, o *SAnDReS* é capaz de construir 511 combinações polinomiais. Ao final de todo processo é possível trabalhar com 28.616 modelos de aprendizado de máquina na elaboração dos polinômios (Xavier *et al* 2016).

Para otimizar e agilizar o processo de elaboração dos polinômios, a escolha dos termos participantes é pautada nos valores de correlação entre eles e o log(IC_{50}),

o que já pode trazer um indicativo dos fatores mais importantes a serem considerados para a predição de afinidade do sistema proteína-ligante. Para as estruturas presentes no *dataset* $HRIC_{50}$ os termos que apresentaram maior valor de correlação foram: interação entre a *pose* e a proteína, energia de ligação interna do ligante/*pose* e a energia das ligações de hidrogênio (Tabela 12). Assim, utilizando o método de regressão *ElasticNet CV*, o polinômio que apresentou o maior valor de correlação para predição de afinidade da enzima foi o seguinte,

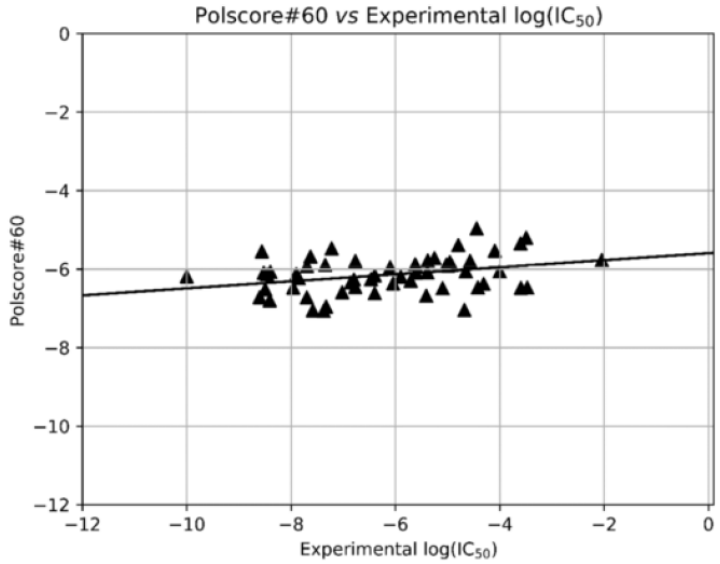
$$Polscore60 = 5{,}763674 + 0{,}000069\,(x.y) + 0{,}000185\,(x.z) - 0{,}001090(y.z) \\ 0{,}000040\,(x^2) \text{ (Equação 7)}$$

onde, *x* representa interação entre a *pose* e a proteína, *y* é a energia de ligação interna do ligante e *z* a energia das ligações de hidrogênio. A partir da elaboração do polinômio, foram realizados testes de predição de afinidade comparando os resultados obtidos com as funções escore e termos energéticos clássicos com a função escore polinomial elaborada. Os valores de correlação para as funções escore clássicas presentes no MVD variaram entre -0,107 e 0,334, já a função polinomial elaborada pelo *SAnDReS* apresentou valor de correlação 0,401 com *p-value* < 0,001(Tabela 13), para valores de *training set*. Considerando o conjunto de dados participantes do *test set*, as funções escore clássicas e os termos energéticos do *software* MVD geraram valores de correlação entre -0,237 e 0,224 com *p-value* > 0,05. Em contraponto, a função polinomial *Polscore60* alcançou valores de correlação de 0,328 e *p-value* < 0,05, indicando estar mais habilitada para os estudos de interação das enzimas participantes do conjunto $HRIC_{50}$ (Tabela 13 e Figura 12).

Tabela 13. Resultados de training e test set obtidos com a utilização de funções escore e termos energéticos para o dataset $HRIC_{50}$. (Termo 1: Interaction; Termo 2: Cofactor; Termo 3: Protein; Termo 4: Water; Termo 5: Internal; Termo 6: Electro; Termo 7: ElectroLong; Termo 8: HBond).

Funções Escore e Termos Energéticos	**ρ (*training set*)**	**p-valeu (*training set*)**	**ρ (*test set*)**	**p-value (*test set*)**
PLANTS Score	0,266	$3,797.10^{-03}$	0,167	$2,185.10^{-01}$
MolDock Score	0,284	$1,939.10^{-03}$	0,224	$9,678.10^{-02}$
Rerank Score	0,227	$1,371.10^{-02}$	0,109	$4,219.10^{-01}$
Termo 1	0,334	$2,305.10^{-04}$	0,215	$1,109.10^{-01}$
Termo 2	0,130	$1,623.10^{-01}$	0,211	$1,192.10^{-01}$
Termo 3	0,340	$1,795.10^{-04}$	0,147	$2,810.10^{-01}$
Termo 4	0,214	$2,032.10^{-02}$	0,083	$5,455.10^{-01}$
Termo 5	-0,077	$4,104.10^{-01}$	0,155	$2,541.10^{-01}$
Termo 6	-0,107	$2,514.10^{-01}$	−0,179	$1,871.10^{-01}$
Termo 7	0,134	$1,511.10^{-01}$	0,101	$4,568.10^{-01}$
Termo 8	-0,067	$4,746.10^{-01}$	−0,237	$7,889.10^{-02}$
Polscore0000060	0,401	$7,243.10^{-06}$	0,328	$1,363.10^{-02}$

Figura 12. Representação de correlação presente entre a predição de $log(IC_{50})$ e o $log(IC_{50})$ experimental. Dados obtidos a partir dos resultados de *test set*.



Fonte: próprio autor.

Após a confirmação dos bons resultados gerados pelo polinômio *Polscore60*, buscamos verificar sua capacidade de predizer a afinidade dos ligantes pelo sítio ativo da CDK2. Para isso, foi utilizado o conjunto de 11 estruturas da enzima complexada a ligantes distintos ($CDK2IC_{50}$). Nesse momento do projeto, além de compararmos os resultados da equação polinomial aos termos clássicos do MVD, também foram

utilizados os termos e funções escore presentes no *AD4* e *Vina*. A análise do poder de predição de afinidade dos termos clássicos indicou um coeficiente de correlação entre -0,773 (Gauss1) e 0,682 (*PLANTS Score*) (*p-value* < 0,05). Os resultados alcançados pela *Polscore60* (ρ = 0,845; *p-value* < 0,001) (Tabela 14) foram melhores em relação as funções escore clássicas indicando um maior poder de predição de afinidade para a CDK2.

Tabela 14. Análise estatística do poder de predição para todas as onze estruturas presente no *dataset* CDK2IC$_{50}$.

Scoring Functions and Energy Terms[a]	**ρ**	**p-value**	**R^2**	**p-value**
Affinity[b]	0,418	$2{,}006.10^{-01}$	0,237	$1{,}289.10^{-01}$
Gauss1[b]	-0,773	$5{,}299.10^{-03}$	0,393	$3{,}889.10^{-02}$
Gauss2[b]	-0,645	$3{,}196.10^{-02}$	0,386	$4{,}125.10^{-02}$
Repulsion[b]	-0,618	$4{,}265.10^{-02}$	0,276	$9{,}715.10^{-02}$
Hydrophobic[b]	-0,391	$2{,}345.10^{-01}$	0,223	$1{,}424.10^{-01}$
Hydrogen[b]	-0,730	$1{,}069.10^{-02}$	0,280	$9{,}386.10^{-02}$
Free Energy[c]	0,445	$1{,}697.10^{-01}$	0,082	$3{,}923.10^{-01}$
Final Intermolecular Energy[c]	0,400	$2{,}229.10^{-01}$	0,082	$3{,}923.10^{-01}$
vdW+Hbond+desolv Energy[c]	0,409	$2{,}115.10^{-01}$	0,082	$3{,}923.10^{-01}$
Electrostatic Energy[c]	-0,209	$5{,}372.10^{-01}$	0,082	$3{,}922.10^{-01}$
Final Total Internal Energy[c]	0,588	$5{,}725.10^{-02}$	0,345	$5{,}730.10^{-02}$
Torsional Free Energy[c]	-0,304	$3{,}637.10^{-01}$	0,106	$3{,}298.10^{-01}$
MolDock Score[d]	0,391	$2{,}345.10^{-01}$	0,173	$2{,}028.10^{-01}$
PLANTS Score[d]	0,682	$2{,}084.10^{-02}$	0,507	$1{,}401.10^{-02}$
Rerank Score[d]	-0,591	$5{,}558.10^{-02}$	0,768	$4{,}044.10^{-04}$
Ligand Efficiency 1[d]	-0,391	$2{,}345.10^{-01}$	0,183	$1{,}888.10^{-01}$
Ligand Efficiency 3[d]	-0,345	$2{,}981.10^{-01}$	0,294	$8{,}516.10^{-02}$
Polscore 60[e]	0,845	$1{,}045.10^{-03}$	0,608	$4{,}650.10^{-03}$

[a]Os valores dos termos energéticos e das funções escore foram calculados usando a posição cristalográfica do ligante
[b] Funções escore e Termos energéticos calculados com o *software Vina*.
[c] Funções escore e Termos energéticos calculados com o *software* AD4.
[d] Funções escore e Termos energéticos calculados com o *software* MVD.
[e] Função escore polinomial gerada usada com o *software* SAnDReS.

Na tabela 15 é possível confirmar o alto poder de predição do polinômio a partir dos dados da predição de log(IC$_{50}$) e os de log(IC$_{50}$) experimental. A proximidade entre os dois valores para todas estruturas componentes do *dataset* indicam que a função escore polinomial *Polscore60* é uma importante aliada em estudos futuros que visem a busca por novos inibidores.

Tabela 15. Log(IC_{50}) experimental e previsto para todas as estruturas do *dataset* CDK2IC_{50}.

Código PDB	Código do ligante	Resolução (Å)	IC_{50}(nM)	log(IC_{50})	log(IC_{50}) Predito
1GII	1PU	2,00	260	-6,585	-6,463
1OIR	HDY	1,91	32	-7,495	-7,495
2B53	D23	2,00	600	-6,222	-6,221
2B54	D05	1,85	20	-7,699	-7,699
2R3H	SCE	1,50	20000	-4,699	-3,839
3IGG	EFQ	1,80	80.75	-7,093	-6,196
3LE6	2BZ	2,00	35	-7,456	-6,277
3PXZ	JWS	1,70	5900	-5,229	-5,277
3PY0	SU9	1,75	79.25	-7,101	-6,678
3RZB	02Z	1,90	100000	-4,000	-5,779
4RJ3	3QS	1,63	93	-7,032	-6,544

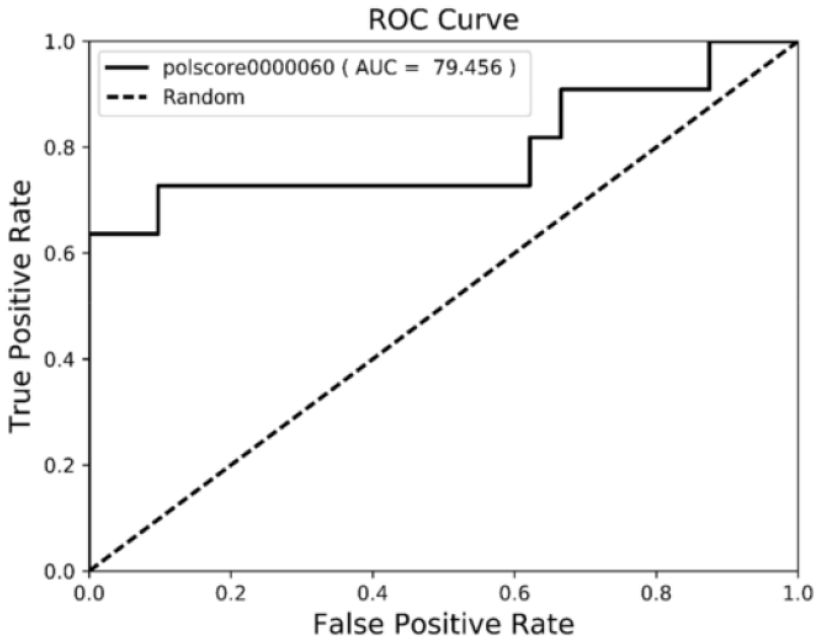
Analisando os resultados obtidos com essa equação polinomial, podemos supor que o seu sucesso se deve à predominância de termos que utilizam as interações intermoleculares (*Internal* (Termo 3) e *H-Bond* (Termo 8)) como base do cálculo de energia. Isso pode ser percebido através de estudos prévios (de Azevedo *et al.* 2002; de Azevedo *et al*. 2002; Canduri *et al*. 2004; Perez *et al*. 2009; Schonbrunn *et al*. 2013) que apontam a alta dependência de ligações de hidrogênio nas interações C=O (Glu81), N-H e C=O (Leu83) do sítio ativo, representativos no “Garfo Molecular”, e o ligante no processo de inibição.

4.2.4 Decoys and Actives

Nessa última fase do estudo de *docking* da CDK2 foi analisada a habilidade do polinômio *Polscore60* de encontrar os ligantes ativos, ou seja, aqueles que realmente apresentam afinidade pela enzima em um universo de ligantes que não apresentam grau de afinidade com a CDK2. Os *decoys* foram retirados da base de dados DUD-E (Mysinger *et al*. 2012) e os ligantes provieram das estruturas cristalográficas de CDK2 utilizadas nesse projeto. Foi elaborada uma base de dados contendo 1100 ligantes, sendo desses 11 ligantes ativos e 1089 sendo *decoys*. A performance foi analisada a partir dos valores de AUC, ROC e de EF. Seguindo o sucesso relatado na sessão anterior, a função polinomial *Polscore60* apresentou um EF de 175,00 e o AUC de 79,450% (Figura 13) indicando a alta capacidade da função em encontrar ligantes verdadeiros e predizer a afinidade de ligação entre CDK2 e moléculas candidatas a

inibidor. Além disso, quando comparada a estudos anteriores, *Polscore60* apresentou melhores valores de AUC e EF (Mysinger *et al*. 2012).

Figura 13. Curva ROC gerada a partir dos resultados obtidos pelo polinômio Polscore#60.



Fonte: próprio autor.

CAPÍTULO 5

5 CONSIDERAÇÕES FINAIS

No atual cenário mundial, a busca rápida e otimizada por novas drogas que possam realizar tratamento efetivo contra diversas enfermidades torna-se extremamente importante. Nesse contexto, as ferramentas computacionais prestam grande auxílio na busca pela compreensão de mecanismos biológicos que possam servir como alvos desses novos medicamentos que nos são tão caros atualmente. A computação natural, aliada aos procedimentos do aprendizado de máquina, ganham ainda mais importância, uma vez que são utilizados processos baseados no funcionamento da natureza aliados ao reconhecimento de padrões e elaboração/utilização de algoritmos. Elas objetivam a descrição e compreensão das diversas possibilidades de inibição de sistemas fundamentais para a sobrevivência de certos micro-organismos, ou conjuntos de células.

A partir dos testes realizados e resultados obtidos durante o andamento desse projeto, foi possível perceber que as técnicas descritas anteriormente, realmente, são capazes de nos prestar auxílio no desenvolvimento de novos fármacos. As funções escore/termos energéticos empregadas/os por cada *software* utilizado, demonstraram-se eficazes na compreensão e simulação das relações intermoleculares entre os sítios de ligação das enzimas-alvo com seus ligantes naturais e possíveis inibidores. Além disso, quando participantes de funções escore polinomiais, elaboradas a partir de métodos de aprendizagem de máquina, as funções e termos clássicos mostraram-se eficientes, já que as funções polinomiais nos trouxeram uma nova possibilidade de compreensão das interações moleculares. Isso pode ser observado a partir da maior eficácia na predição de afinidade entre proteína e possíveis ligantes/inibidores por essas mesmas funções polinomiais.

Ao final do trabalho, é possível afirmar que os métodos utilizados, assim como as funções escore, termos energéticos e funções escore polinomiais, obtiveram resultados satisfatórios na predição de afinidade entre os sítios de ligação das duas enzimas estudadas, InhA e CDK2. A partir dos valores de correlação e dos valores de significância estatística obtidos para cada função polinomial, é sugerido que em estudos posteriores, para identificação de novas drogas para as duas enzimas, as

funções *Polscore#231* (InhA), *Polscores#245* (InhA) e *Polscore#60* (CDK2) sejam utilizadas na seleção de candidatos químicos com potencial inibitório, servindo, assim, como técnica de otimização para as possíveis testagens *in vitro* de novas drogas.

6 REFERÊNCIAS

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). MINISTÉRIO DA SAÚDE, BRASIL. Http://www.anvisa.gov.br/servicosaude/controle/rede_rm/cursos/rm_controle/opas_web/modulo1/conceitos.htm. (Acesso em 17/03/2020).

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). MINISTÉRIO DA SAÚDE, BRASIL. **Resolução-RDC Nº 44, de 26 de outubro de 2010. Dispõe sobre o controle de medicamentos à base de substâncias classificadas como antimicrobianos, de uso sob prescrição médica, isoladas ou em associação e dá outras providências.** Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2010/res0044_26_10_2010.html. Acesso em: 20 janeiro de 2020.

AKSYUK AA *et al*. **The tail sheath structure of bacteriophage T4: a molecular machine for infecting bacteria.** The EMBO Journal. 2009; 28(7): 821 – 829.

ALBERTS B, JOHNSON A, LEWIS J, RAFF M, ROBERTS K, WALTER. **Biologia Molecular da Célula.** 4ª Edição. Porto Alegre (RS): Artmed Editora, 2004.

ARGYROU A, VETTING MW, BLANCHARD JS. **New insight into the mechanism of action and resistance to isoniazid: interaction of *Mycobacterium tuberculosis* enoyl-ACP reductase INH-NADP**. J Am Chem Soc. 2007; 129: 9582-9583.

AZEVEDOLAB. (www.azevedolab.net). (Acesso em: 20/07/2019).

BAČEVIĆ K, LOSSAINT G, ACHOUR TN, GEORGET V, FISHER D, DULIĆ V. **Cdk2 strengthens the intra-S checkpoint and counteracts cell cycle exit induced by DNA damage**. Nature Scientific Reports. 2017; 7(1): 1-14.

BERMAN HM, *et al*. **The Protein Data Bank**. Nucleic Acids Research. 2000; 28:235-242.

BERTRAM JS. **The molecular biology of cancer**. Molecular aspects of Medicine. 2000; 21(6): 167 – 223.

BHATT A, MOLLE V, BESRA GS, JACOBS WR, KREMER L. **The Mycobacterium tuberculosis FAS-II condensing enzymes: their role in mycolic acid biosynthesis, acid-fastness, pathogenesis and in future drug development.** Molecular Microbiology. 2007; 64(6): 1442 – 1454.

BÖHM HJ. **The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure**. J Comput Aided Mol Des. 1994; 8(3): 243-56.

BÖHM HJ. **Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs**. J Comput Aided Mol Des. 1998; 12(4): 309-23.

BROOIJMANS N & KUNTZ D. **Molecular recognition and docking algorithms.** Annual Review of Biophysics and Biomolecular Structure. 2003; 32(1): 335 - 373

CANDURI F, UCHOA HB, de AZEVEDO WF Jr. **Molecular models of cyclin-dependent kinase 1complexed with inhibitors**. Biochem Biophys Res Commun. 2004; 324: 661 – 666.

CICENAS J *et al*. **Roscovitine in cancer and other diseases**. Ann Transl Med. 2015; 3(10): 135

CICHERO E, CESARINI S, MOSTI L, FOSSA P. **CoMFA and CoMSIA anlysis of 1,2,3,4-tetrahydropyrrolo[3,4-]índole and benzimidazole derivatives and selective CB2 receptor agonists**. Journal of Molecular Modeling. 2010; 16(9): 1481 – 1498.

COLEY DA. **An Introduction to Genetic Algorithms for Scientists and Engineers.** 1ª Edição. Londres (UK): World Scientific Publishing, 1999.

CORDON-CARDO C. **Mutations of cell cycle regulators. Biological and clinical implications for human neoplasia**. Am. J. Pathol. 1995. 147: 545-560.

DARWIN C. **A origem das Espécies**. 1ª Edição. São Paulo (SP): Martin Claret, 2014.

DAVIES J & DAVIES D. **Origins and Evolution of Antibiotic Resistance**. Microbiol Mol Bio Rev. 2010; 74: 417 – 433.

DE AVILA MB, XAVIER MM, PINTRO VO, de AZEVEDO WF Jr. **Supervised Machine learning techniques to predicting binding affinity. A study for cyclin-dependent kinase 2**. Biochem Biophys Res Commun. 2017; 9: 305 – 310.

DE AVILA MB & DE AZEVEDO WF Jr. **Development of machine learning models to predict inhibition of 3-dehydroquinate dehydratase**. Chem Bio Drug Design. 2018; 92: 1468 – 1474.

DE AVILA MB, BITTENCOURT-FERREIRA G, DE AZEVEDO WF. **Structural Basis for Inhibition of Enoyl-[Acyl Carrier Protein] Reductase (InhA) from Mycobacterium tuberculosis.** Current Medicinal Chemistry. 2020; 27: 745 – 759.

DE AZEVEDO WF, LECLERC S, MEIJER L, HAVLICEK L, STRNAD M, KIM SH. **Inhibition of cyclin-dependent kinases by purine analogues**. The FEBS Journal. 1997, 243(1-2): 518-526.

DE AZEVEDO WF Jr., CADURI F, da SILVEIRA NJ. **Structural Basis for inhibition of cyclin-dependent kinase 9 by flavoripidol**. Biochem Biophys Res Commun. 2002; 239: 566 – 571.

DE AZEVEDO WF, GASPAR RT, CANDURI F, CAMERA JR JC, FREITAS DA SILVEIRA NJ. **Molecular modeling of cyclin-dependent kinase 5 complexed with roscovitine.** Biochem Biophys Res Commun. 2002; 297: 1154 – 1158.

DE AZEVEDO WF JR, DIAS R. **Evaluation of ligand-binding affinity using polynomial empirical scoring functions**. Bioorganic & Medicinal Chemistry. 2008, 16(20): 9378-9382.

DE CASTRO LN. **Fundamentals of natural computing: an overview**. Physics of Life Reviews. 2007; 4: 1 – 36.

DESSEN A, QUÉRMAD A, BLANCHARD JS, JACOBS Jr WR, SACCHETTINI C. **Crystal Structure and Function of the Isoniazid Target of Mycobacterium tuberculosis.** Science. 1995, 267: 1638 – 1641.

DEVITO JA & MORRIS S. **Exploring the Structure and Function of the Mycobacterium KatG Protein Using *trans*-Dominant Mutant**. Antimicrobial Agents and Chemotherapy. 2003; 47: 188 – 195.

ELDRIDGE MD, MURRAY CW, AUTON TR, PAOLINI GV, MEE RP. **Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.** Journal of Computer-Aided Molecular Design. 1997; 11: 425 – 445.

FLOREANO D & MATTIUSSI C. **Bio-inspired Artificial Intelligence. Theories, Methods, and Technologies**. Cambridge: The MIT Press 2008.

FOLINO G & MASTROIANNI C. **Special Issue: Bio-Inspired Optimization Techniques for High Performance Computing**. New Generation Computing. 2011; 29: 125 – 128.

FONSECA JD, KNIGHT GM, MCHUGH TD. **The complex evolution of antibiotic resistance in *Mycobacterium tuberculosis*. International Journal of Infectious Diseases**. Internation Journal of Infectious Disease. 2015, 32: 94-100.

FRACZEK T, SIWEK A, PANETH P. **Assessing Molecular Docking Tools for Relative Biological Activity Prediction: A Case Study of Triazole HIV-1 NNRTIs**. Chemical Information and Modeling. 2013; 53: 3326 – 3342.

FREUNDLICH JS *et al*. **Triclosan dreivatives: towards potent inhibitors of drug-sensitive and drug-resistance Mycobacterium tuberculosis.** ChemMedChem> Chemistry Enabling Drug Discovery. 2009; 4(2): 241 – 248.

GAILLARD T. **Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark**. J Chem Inf Model. 2018; 58: 1697 – 1706.

GOLDBERG DE. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Crawfordsville: Addison Wesley Longman, Inc. 1989.

GOODSELL DS & OLSON AJ. **Automated docking of substrates to proteins by simulated annealing**. Proteins: Structure, Function and Bioinformatics. 1990; 8: 195 – 202.

GOULDING R, KING MB, KNOX R, ROBSON JM. **rRelation between in-vitro and in-vivo resistance to isoniazid**. Lancet. 1952; II: 69 – 70.

GUEDES IA, PEREIRA FSS, DARDENNE LE. **Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges.** Frontiers in Pharmacology. 2018; 9: 1089.

HANAHAN D & WEINBERG RA. **Hallmarks of cancer: the next generation**. Cell. 2000; 144(5): 646 – 674.

HART WE. **Adaptive global optimization with local search**. Tese de Doutorado. University of California, San Diego, Department of Computer Science and Engineering. 1994

HE X, ALIAN A, STROUD R, ORTIZ DE MONTELLANO PR. **Pyrrolidine Carboxamides as a Novel Class Of Inhibitor of Enoyl Acil Carrier Protein Reductase from Mycobacterium tuberculosis.** J Med Chem. 2006; 49(21): 6308 – 6323.

HEBERLÉ G, DE AZEVEDO WF Jr. **Bio-inspired algorithms applied to molecular docking simulations**. Curr Med Chem. 2011; 18: 1339-52.

HECK GS *et al.* **Supervised machine learning methods applied to predict ligand-binding affinity**. Curr Med Chem. 2017; 24(23): 2459-2470.

HOCHEGGER H, TAKEDA S, HUNT T. **Cyclin-dependent kinases and cell-cycle transitions: does one fit all?** Nature Reviews: Molecular Cell Biology. 2008; 9: 910 – 916.

HOLLAND JH. **Genetic Algorithms and the Optimal Allocation of Trials**. SIAM J Comput. 1975; 2: 88-105.

HOWARD EI *et al*. **Ultrahigh resolution drug design I: details of interaction in Human aldose reductase-inhibitor complex at 0.66A.** Proteins. 2004: 55(4): 792 – 804.

HU L, BENSON ML, SMITH RD, LERNER MG, CARLSON HA. **Binding MOAD (Mother Of All Databases)**. Proteins. 2005; 60(3): 333-40.

HUANG SY, GRINTER SZ, ZOU X. **Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions**. Phys Chem Chem Phys. 2010; 12: 899-908.

HUGHES D & ANDERSON DI**. Evolutionary trajectories to antibiotic resistance**. Annual review of microbiology. 2017; 71: 579 – 596.

INTURI B, PUJAR GV, PUROHIT MN. **Recent advances and structural features of enoyl-ACP reductase inhibitors of *Mycobacterium tuberculosis***. Arch Pharm. 2016; 349(11): 817-826.

IRWIN JJ & SHOICHET BK. **ZINC – a free database of commercially available compounds for virtual screening**. J Chem Inf Model. 2005; 45: 177 – 182.

JACKETT PS, ABER VR, LOWRIE DB. **Virulence and resistance to superoxide, low pH and hydrogen peroxide among strains of Mycobacterium tuberculosis**. Microbiology. 1978; 104(1):37 – 45.

JOHNSSON K & SHULTZ PG. **Mechanistic studies of the oxidation of isoniazid by the catalase peroxidase from Mycobacterium tuberculosis**. J Am Chem Soc. 1994; 116: 7425 – 7426.

JORGENSEN WL. **Efficient Drug Lead Discovery and Optimization.** Accounts of Chemical Research. 2009; 42(6): 724 – 733.

KARP JE & BRODER S. **Molecular foundations of cancer: new targets for intervention**. Nat. Med. 1995. 1: 309-320.

KAUFMANN SHE. **Paul Ehrlich: founder of chemotherapy.** Nature Reviews: Drug Discovery. 2008; 7: 374.

KIM SJ *et al*. **Dimeric and tetrameric forms of enoyl-acyl carrier protein reductase from *Bacillus cereus***. Biochemical and Biophysical Research Communications. 2010; 400: 517-522.

KITCHEN DB, DECORNEZ H, FURR JR, BAJORATH J. **Docking and Scoring in Virtual Screening for Drug Discovery: Method and Applications.** Nature Reviews: Drug Discovery. 2004; 3: 935 – 949.

KORB O, STUTZLE T, EXNER TE. **Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS**. J Chem Inf Model. 2009, 49: 84 – 96.

KUNTZ ID, BLANEY JM, OATLEY SJ, LANGRIDGE R, FERRIN TE. **A Geometric Approach to Macromolecular-Ligand Interactions.** J. Mol. Biol. 1982; 161: 269 – 288.

LABORDE J, DERAEVE C, BERNARDES-GÉNISSON V. **Update of antitubercular prodrugs from a molecular perspective: mechanism of action, bioactivation pathways, and associated resistance**. 2017; ChemMedChem. Epub. DOI: 10.1002/cmdc.201700424.

LAMEIJER EW, BACK T, KOK JN, IJZERMAN AP. **Evolutionary algorithms in drug design**. Natural Computing. 2005; 4: 177 – 243.

LASKOWSKI RA & SWINDELLS MB. **LigPlot^{+}: multiple ligand-protein interaction diagrams for drug discovery**. J Chem. Inf. Model. 2011, 51: 2778-2786.

LEE B, WEI CJ, TU SC. **Action mechanism of antitubercular isoniazid. Activation by *Mycobacterium tuberculosis* KatG, isolation, and characterization of inha inhibitor**. J. Biol. Chem. 2000; 275: 2520-2526.

LI HJ *et al.* **A structural and energetic model for slow-onset inhibition of the *Mycobacterium tuberculosis* Enoyl-ACP reductase InhA**. ACS Chemical Biology. 2014; 9: 986-993.

LI J, VERVOOTS J, CARLONI P, ROSSETI G, LÜSCHER B. **Structural prediction of the interaction of the tumor suppressor p27KIP1 with cyclin A/CDK2 identifies a novel catalytically relevant determinant**. BMC Bioinformatics. 2017; 18(15): 1-9.

LI J, FU A, ZHANG L. **An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking.** Interdisciplinary Sciences: Computational Life Sciences. 2019; 11: 320 – 328.

LIU T, LIN Y, WEN C, JORISSEN RN, GILSON MK. ***BindingDB*: a web-accessible database of experimentally determined protein-ligand binding affinities**. Nucleic Acids Research. 2007; 1:198-201.

MAAMAR B, HU J, HARTMANN EM. **Implications of indoor microbial ecology and Evolution on antibiotic resistance**. J Expo Sci Environ Epidemiol. 2020; 30: 1 – 15.

MALUMBRES M, BARBACID M. **Cell cycle, CDKs and cancer: a changing paradigm.** Nature Reviews Cancer. 2009; 9(3): 153-166.

MANCA C, PAUL S, CARRY CE, FREEDMAN VH, KAPLAN G. **Mycobacterium tuberculosis catalase and peroxidase activities and resistance to oxidative killing in human monocytes in vitro**. Infection and immunity. 1999; 67(1): 74 – 79.

MASTERS L, EAGON S, HEYING M. **Evaluation of consensus scoring methods for AutoDock Vina, simna and idock**. Journal of Molecular Graphics and Modeling. 2020; 96: 1 – 10.

MENG EC, SHOICHET BK, KUNTZ ID. **Automated docking with grid-based energy evaluation**. J Comput Chem. 1992; 13: 505-24.

MINISTÉRIO DA SAÚDE DO BRASIL (MS). **Estimativa de câncer no Brasil em 2020.** (www.inca.gov.org) (Acesso em: 31/8/2017)).

MITCHISON DA, SELKON JB, LLOYD J. **Virulence in the guinea pig, susceptibility to hydrogen peroxide, and catalase activity of isoniazid-sensitive tubercle bacilli form South Indian and British patients**. J Pathol Bacteriol. 1963; 86: 377 – 386.

MITRASINOVIC PM. **Progress in structure-based design of EGFR inhibitors.** Current Drug Targets. 2013; 14(7): 817-829.

MITRASINOVIC PM. **Towards an Experimental and Systems Biology Framework for Cancer Cell Therapeutics**. Curr Bioinformatics. 2012; 7:490-504.

MORRIS G *et al.* **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. Journal of Computational Method. 1998; 19: 1639-1662.

MORRIS GM, *et al*. **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility**. Journal of Computational Chemistry. 2009; 30: 2785 – 2791.

MORRIS GM, GOODSELL DS, HALLIDAY RS, HUEY R, HART WE, BELEW RK, OLSON AJ. **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.** Journal of Computational Chemistry. 1998; 19(14): 1639 – 1662.

MORSE WC, WEISER OL, KUHNS DM, FUSILIO M, DAIL MC, EVANS JR. **Study of the virulence of isoniazid-resistant tubercle bacilli in guinea pigs and mice.** Am Rev Tuberc.1952; 69: 464 – 468.

MUKHOPADHYAY M. **A Brief on Bio-Inspired Optimization Algorithm for Molecular Docking**. International Journal of Advances in Engineering & Technologies. 2014; 7: 868 – 878.

MULLER P. **Glossary of terms used in physical organic chemistry (IUPAC Recommendations)**. Pure and Applied Chemistry. 1994; 66(5): 1077 – 1184.

MURRAY AW. **Recycling the cell cycle: cyclins revisited**. Cell. 2004; 116(2): 221–234.

MYSINGER MM, CARCHIA M, IRWIN JJ, SHOICHET BK. **Directory useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmark.** J Med Chem. 2012; 55: 6582 – 6594.

NEKARDOVÁ M, *et* al. **Structural Basis of the Interaction of Cyclin-Dependent Kinase 2 with Roscovitine and Its Analogues Having Bioisosteric Central Heterocycles**. Chem Phys Chem. 2017; 18: 785 – 795.

NOVIC M, TIBAUT T, ANDERLUCH M, BORISEK J, TOMASIC T. The Comparison of Docking Search Algorithms and Scoring Functions: An Overview and Case Studies. *In*: DASTMALCHI S, HAMZEH-MIVEHROUD M, SOKOUTI B (org). **Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery**. 1. ed. USA: IGI Global. 2016.

PAN P & TONGE PJ. **Targeting InhA, the FASII enoyl-ACP reductase: SAR studies on novel inhibitor scaffolds**. Curr Top Med Chem. 2012; 12(7): 672-693.

PARRIL AL. **Evolutionary and Genetic methods in drug design**. Drug Discovery Today. 1996; 12: 514 – 521.

PEREZ PC, CACERES RA, CANDURI F, de AZEVEDO JR WF. **Molecular modeling and dynamics simulation of human cyclin-dependent kinase 3 complexed with inhibitors**. Comput Biol Med. 2009; 39: 130 - 140.

PRICE K. **Differential evolution: a fast and simple numerical optimizer.** Proceedings of North Amercan Fuzzy Information Processing. IEEE, 1996.

PURVES WK, SADAVA D, ORIANS GH, HELLER HC. **Vida: a ciência da biologia**. 6ª Edição. Porto Alegre (RS): Artmed Editora, 2005.

RAWAT R, WHITTY A, TONGE P. **The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the Mycobacterium tuberculosis enoyl reductase: adduct affinity and drug resistance**. Proceedings of the National Academy of Science. 2003; 100(24): 13881 – 13886.

ROZMAN K, SOSIC I, FERNANDEZ R, YOUNG RJ, MENDOZA A, GOBEC S, ENCINAS L. **A new 'golden age' for the antitubecular target InhA**. Drug Dscovery Today. 2017, 22(3): 492-502.

RUSSO AA, JEFFREY PD, PATTEN AK, MASSAGUÉ J, PAVLETICH NP. **Crystal structure of the p27^{Kip1} cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex**. Nature. 1996, 382: 325-331.

RUSSO S & DE AZEVEDO WF Jr**. Computational Analysis od Dipyrone Metabolite 4-Aminoantipyrine as a Cannabinoid Receptor 1 Agonist**. Curr Med Chem. 2019. doi: 10.2174/0929867326666190906155339 (Epub ahead of print).

SCHNEIDER G. **Virtual Screening: an endless staircase?** Nature Reviews: Drug Discovery. 2010; 9: 273 – 276.

SCHOLAR EM & PRATT WB. **The Antimicrobial Drugs**. Oxford University Press. Second Edition. 2000.

SCHONBRUNN *et al*. **Development of Highly Potent and Selective Diaminothiazole Inhibitors of Cyclin-Dependent Kinases**. J Med Chem. 2013; 56: 3768 – 3782.

SCHRIJVER R, STIJNTJES M, RODRÍGUEZ-BAÑO J, TACCONELLI E, RAJENDRAN NB, VOSS A. **A review of antimicrobial resistance surveillance programmes in livestock and meat thereof in Europe with a focus on antimicrobial resistance patterns in humans**. Clinical Microbiology and Infection. 2017; doi: 10.1016/j.cmi.2017.09.013

SCIOR T, MORALES IM, EISELE SJG, DOMEYER D, LAUFER S. **Antitubercular isoniazid and drug resistance of *Mycibacterium tuberculosis* – a review**. Arch Pharm. 2002; 335: 511-525.

SHERMAN DR *et al.* **Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis**. Science. 1996; 272(5268): 1641 – 1643.

SIPPL MJ. **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins**. J Mol Biol. 1990; 213(4): 859-83.

SOUSA SL, FERNANDES PA, RAMOS MJ. **Protein-Ligand Docking: Current Status and Future Challenges.** Proteins: Structure, Function and Bioinformatics. 2006; 65: 15 – 26.

STORN R & PRICE K. **Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces.** Journal of Global Optimization. 1997; 11(4): 341 – 359.

STRATTON MR. **Exploring the genomes of cancer cells: progress and promise**. Science. 2011; 331(6024): 1553 – 1558.

TADESSE S, ANSHABO AT, PORTMAN N, LIM E, TILLEY W, CALDON EC, WANG S. **Targeting CDK2 in cancer: challenges and opportunities for therapy.** Drug Discovery Today. 2020; 25(2): 406 – 413.

TANAKA S, SCHERAGA HA. **Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins**. Macromolecules. 1976; 9(6): 945-50.

TEAGUE SJ. **Implications of Protein Flexibility for Drug Discovery.** Nature Reviews: Drug Discovery. 2003; 2: 527 – 541.

THOMSEN R & CHRISTENSEN MH. **MolDock: a new technique for high-accuracy molecular docking**. J Med Chem. 2006: 49:3315-3321.

TROTT O, OLSON AJ. **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading**. J Comput Chem. 2010; 31(2): 455-461.

VAN DEN HEUVEL S & HARLOWN E. **Distinct roles of cyclin-dependent kinases in cell cycle control**. Science. 1994; 262(5142): 2050 – 2054.

VILCHÈZE C *et al*. **Enhanced respiration prevents drug tolerance and drug resistance in *Mycobacterium tuberculosis***. PNAS. 2017; 114(17): 4495-4500.

WANG R, FANG, X, LU Y, WANG S. The PDBbind **Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures**. J Med Chem. 2004; 47(12): 2977-80.

WARSHAVIAK DT, GOLAN G, BORRELLI KW, ZHU K, KALID O. **Structure-based virtual screening approach for discovery of covalently bound ligands**. Chemical Information and Modeling. 2014; 54: 1941-1950.

WEINER SJ, KOLLMAN PA, CASE DA, et al. **A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins**. J Am Chem Soc. 1984; 106: 765-784.

WESTERMAIER Y, BARRIL X, SCAPOZZA L. **Virtual screening: an in silico tool for interlacing the chemical universe with the proteome**. Methods. 2015; 71: 44-57.

WORLD HEALTH ORGANIZATION. **WHO Report on Surveillance of Antibiotic Consumption: 2016 – 2018 Early Implementation**. 2018.

WORLD HEALTH ORGANIZATION. **WHO report on Cancer: setting priorities, investing wisely and provide care for all.** Geneva: World Health Organization; 2020.

WRIGHT G. **The antibiotic resistance: the nexus of chemical and genetic diversity.** Nat Rev Microbiol. 2007; 5(3): 175-86.

XAVIER MM, *et al***. SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions**. Combinatorial Chemistry & High Throughput Screening. 2016; 19:1-12.

YIN X *et al*. **Identification of CDK2 as a novel target in treatment of prostate cancer.** Future Oncology. 2018; 14(8): 709 – 718.

ZAR JH. **Significance Testing of the Spearman Rank Correlation Coefficient**. J Amer Statist Assoc. 1972; 67(339): 578-80.

ZHANG *et al*. **A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations**. Cancer Cell. 2017; 31(6): 820 – 832.

ZHU K *et al.* **Docking covalent inhibitors: a parameter free approach to pose predicition and scoring**. Chemical Information and Modeling. 2014; 54: 1932-1940.

ANEXO A – Descrição dos trinta e dois protocolos elaborados com as Funções Escore e Algoritmos de Busca presentes no software MVD (Thomsen & Christensen, 2006).

Protocol	Scoring Functions	Search Algorithm	Displaceable Water?	RMSD (Å)	RMSD (Å)	RMSD (Å)	RMSD (Å)
Sorting Criteria				**MolDock Score**	**Rerank Score**	**HBond**	**RMSD**
1	MolDock Score	MolDock Optimizer	Yes				
2	MolDock Score	MolDock Optimizer	No				
3	MolDock Score	MolDock (SE)	Yes				
4	MolDock Score	MolDock (SE)	No				
5	MolDock Score	Iterated Simplex	Yes				
6	MolDock Score	Iterated Simplex	No				
7	MolDock Score	Iterated Simplex (ANT)	Yes				
8	MolDock Score	Iterated Simplex (ANT)	No				
9	MolDock Score [GRID]	MolDock Optimizer	Yes				
10	MolDock Score [GRID]	MolDock Optimizer	No				
11	MolDock Score [GRID]	MolDock (SE)	Yes				
12	MolDock Score [GRID]	MolDock (SE)	No				
13	MolDock Score [GRID]	Iterated Simplex	Yes				
14	MolDock Score [GRID]	Iterated Simplex	No				
15	MolDock Score [GRID]	Iterated Simplex (ANT)	Yes				
16	MolDock Score [GRID]	Iterated Simplex (ANT)	No				
17	PLANTS Score	MolDock Optimizer	Yes				
18	PLANTS Score	MolDock Optimizer	No				
19	PLANTS Score	MolDock (SE)	Yes				

20	PLANTS Score	MolDock (SE)	No				
21	PLANTS Score	Iterated Simplex	Yes				
22	PLANTS Score	Iterated Simplex	No				
23	PLANTS Score	Iterated Simplex (ANT)	Yes				
24	PLANTS Score	Iterated Simplex (ANT)	No				
25	PLANTS Score [GRID]	MolDock Optimizer	Yes				
26	PLANTS Score [GRID]	MolDock Optimizer	No				
27	PLANTS Score [GRID]	MolDock (SE)	Yes				
28	PLANTS Score [GRID]	MolDock (SE)	No				
29	PLANTS Score [GRID]	Iterated Simplex	Yes				
30	PLANTS Score [GRID]	Iterated Simplex	No				
31	PLANTS Score [GRID]	Iterated Simplex (ANT)	Yes				
32	PLANTS Score [GRID]	Iterated Simplex (ANT)	No				

ANEXO B – Artigo publicado na revista Current Medicinal Chemistry. 2020. Fator de Impacto: 4,184.

Structural Basis for Inhibition of Enoyl-[Acyl Carrier Protein] Reductase (InhA) from *Mycobacterium tuberculosis*

Maurício Boff de Ávila[a,b,#], Gabriela Bitencourt-Ferreira[a,#], Walter Filgueira de Azevedo Jr.[* a,b]

[a]Laboratory of Computational Systems Biology, School of Sciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil; [b]Graduate Program in Cellular and Molecular Biology, School of Sciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil

[#]MBA and GBF contributed equally to this paper, and both can be considered as the first author. *Address correspondence to this author at the School of Sciences, Pontifical Catholic University of Rio Grande do Sul-PUCRS, Porto Alegre-RS, Brazil; Tel/Fax: ++55-51-3353-4529, E-mails: walter@azevedolab.net, walter.junior@pucrs.br.

Abstract: Background. The enzyme *trans*-enoyl-[acyl carrier protein] reductase (InhA) is a vital protein target for the development of antitubercular drugs. This enzyme is the target for the pro-drug isoniazid, which after being metabolized by the enzyme catalase-peroxidase (KatG) can potently inhibit InhA as an isoniazid-NAD adduct.

Objective. Our goal here is to review the studies on InhA, starting with general aspects and focusing on the recent structural studies, with emphasis on the crystallographic structures of complexes involving InhA and inhibitors.

Method. We start with a literature review, and then we describe recent studies on InhA crystallographic structures. We use this structural information to depict protein-ligand interactions. We also analyze the structural basis for inhibition of InhA and the evaluation of computational methods to predict binding affinity based on the crystallographic position of the ligands.

Results. Analysis of the structures in complex with inhibitors revealed the critical residues responsible for the specificity against InhA. Most of the intermolecular interactions involve the hydrophobic residues with two exceptions, the residues Ser 94 and Tyr 158. Examination of the interactions has shown that many of the key residues for inhibitor binding were found in mutations of the *InhA* gene in the isoniazid-resistant *Mycobacterium tuberculosis*. Computational prediction of the binding affinity for InhA has indicated a moderate uphill relationship with experimental values.

Conclusion. Analysis of the structures involving InhA inhibitors shows that small modifications on these molecules could modulate their inhibition, which may be used to design novel antitubercular drugs specific for multidrug-resistant strains.

Keywords: Crystal Structure, Protein-Ligand Interactions, Enoyl-[Acyl Carrier Protein] Reductase, Drug Design

1. INTRODUCTION

There are many metabolic pathways considered as targets for the development of antibacterial drugs, for recent reviews see [1-7]. Amongst those with structural and functional data, we focus here on the fatty acid biosynthesis (FAB). The primary protein target of the FAB is the enzyme *trans*-enoyl-[acyl carrier protein] reductase (InhA) (EC 1.3.1.9), the Isoniazid-NAD adduct is a potent inhibitor of this enzyme. Isoniazid (INH) is one of the most used antitubercular drugs. To bind to the InhA, INH needs activation to form the inhibitory INH-NAD adduct. The enzyme catalase-peroxidase (KatG) (EC 1.11.1.21) is responsible for catalyzing this reaction [8-10].

FAB is an essential metabolic route responsible for the production of important compounds that form cell wall of bacteria, amongst them, mycolic acids. This molecule has a long carbon chain (C_{60} – C_{90}) and, in *Mycobacterium tuberculosis*, helps the bacteria to live and reproduce inside of macrophages [11]. There are two types of fatty acid synthase in different organisms. Type I (FABI) is found in animals and yeasts and represents a homodimeric multifunctional protein that elongates acetyl-CoA to form a palmitic acid [12]. Monofunctional enzymes form the Type II (FABII), which are present in bacteria and plants [13].

InhA takes part in the biosynthesis of mycolic acid, an essential component of the cell wall in prokaryotes. It completes each cycle of elongation by catalyzing the stereospecific reduction of the double bond at the second position on a growing fatty acid chain converting a *trans*-2,3 enoyl moiety into a saturated acyl chain. The reaction occurs, depending on the microorganism, in NADH or NADPH-dependent manner [9, 10]. This protein has two types, each one found in a different group of organisms (FASI and FASII). The two enzymes have both structural and catalytic differences.

The present review has as focus the analysis of the structural basis for inhibition of InhA from Mycobacterium tuberculosis. To achieve this goal, we analyzed the crystallographic structures available for complexes involving InhA and ligands with experimental information for inhibition constant (Ki), dissociation constant (Kd), and half maximal inhibitory concentration (IC50). The richness of structural information, made possible the define the central residues responsible for binding specificity, which can be explored to design more efficient inhibitors.

2. MATERIALS AND METHODS

2.1. Structural Data

The Protein Data Bank (PDB) [14-16] is the main repository of structural information for protein, DNA, RNA, and complexes involving these macromolecules and small-molecule binders. Specifically for InhA from *Mycobacterium tuberculosis*, there are 80 crystallographic structures available (search carried out on January 3, 2018). Among these structures, 79 present at least one small-molecule ligand complexed to the InhA.

The most interesting structural information is related to the complex of InhA and inhibitors, or at least ligands with details about the dissociation constant and bound to the active site. The PDB can link the structural information with experimental binding affinity data available from other databases such as MOAD [17, 18], BindingDB [19, 20], and PDBbind [21, 22]. Such facility opens the possibility to filter out structural data focusing on those structures for which experimental binding affinity is known.

Restricting our search for those structures that present ligand with binding affinity data, we have 14 structures for which Ki is known. For those with IC50 data, we have 24 structures. We found five structures with Kd data. There are no structures of InhA Gibbs free energy of binding (ΔG), binding enthalpy (ΔH), and entropy (ΔS) data. As we see from Table 1, there is one structure for which experimental information for IC50, Kd, and Ki is known. We have a total of 36 unique structures of complexes involving InhA and inhibitors or ligands. From now on, we refer to these structures as the

InhA dataset. Although the application of NMR has solved structures with a molecular weight in the range of InhA [23], surprisingly there are no structures of this protein using this technique. We applied the program SAnDReS to carry out statistical analysis of the quality of the structural information in the InhA dataset [24].

2.2. Study of Protein-Ligand Interactions

Analysis of protein-ligand interactions is of pivotal importance to determine the structural basis for the specificity of a given inhibitor for an enzyme. Once the structural data is available, we can analyze the complex structures and map the residues participating in the intermolecular interactions. Such analysis can be carried out visually, using any protein visualization program such as Visual Molecular Dynamics (VMD) [25] or Molegro Virtual Docker (MVD) [26-29]. In such analysis, we focus on the binding pocket of complexed structure and analyze the hydrogen bonds and the van der Waals contacts. To illustrate, we generated a figure of the complex of InhA- NADH-INH [30] using the program MVD. Fig. (**1**) shows the NAD-binding pocket of the complex structure highlighting the intermolecular hydrogen bonds and the residues involved in van der Waals contacts.

The main advantage of such approaches is the easiness in the analysis of the complexes. The main drawback is the necessity to carry out this visual analysis as many times as the number of structures present in the dataset to be studied. Another point that should be considered is related to the size of the ligand to analyzed. In the Fig. (**1**), we see that due the crowed binding pocket to pinpoint the relevant interaction from one snapshot is a difficult task. One alternative method is to use a program that uses physical-chemical criteria to identify van der Waals contacts and hydrogen bonds. One of the most used programs to perform this analysis is the LigPlot+ [31, 32].

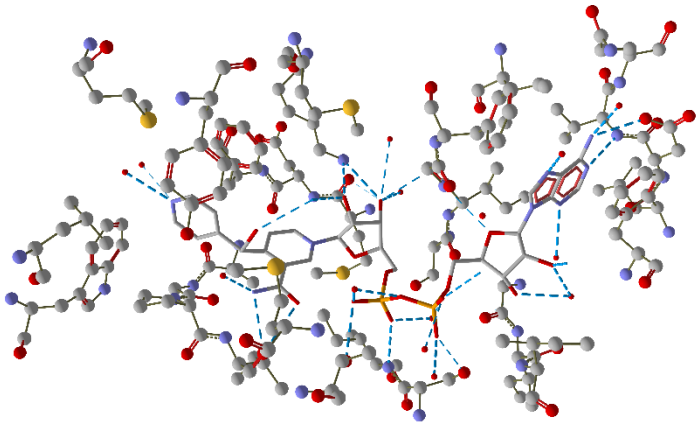


Fig. (1). Analysis of protein-ligand interactions for the crystallographic structure of InhA in complex with NADH-INH (PDB access code: 2IDZ) [30] using the program Molegro Virtual Docker (MVD) [26]. The program MVD draws the ligand (NAD-INH) as thin rods and the protein with the ball-and-stick representation. MVD represents intermolecular hydrogen bonds as dashed lines.

The program LigPlot+ permits determining structural criteria to assess InhA-ligand interactions. This program generates uniformity in the analysis of protein-ligand interactions since it employs the same strong structural evidence to assign a given interaction for a pair of atoms. We used LigPlot+ to analyze the binding of a series of inhibitors to the InhA structure. For comparison reason, we generated the plot of the binding pocket of the same structured portraited using the program MVD. Fig. (**2**) shows the protein-ligand interactions for the structure 2IDZ [30]. As we can see, the program LigPlot+ generates a bidimensional plot highlighting the intermolecular interactions identified in the complex structure.

Table 1. Structural information available in the PDB for complexes involving InhA and ligand with binding data[1]

PDB Code	Ligand Code	Ligand Chain	Ligand Number	IC_{50} (nM)	K_i (nM)	K_d (nM)
1P44	GEQ	A	350	200	ND	ND
1P45	TCL	A	400	1600	210	ND
2B35	TCL	A	300	1600	210	ND
2B36	5PP	A	290	17	11.8	ND
2B37	8PS	C	300	5	1.05	ND
2IDZ	ZID	A	300	ND	ND	0.4
2NSD	4PI	A	400	5160	ND	ND
2NTJ	P1H	A	300	ND	2	ND
2NV6	ZID	A	300	323	ND	ND
2PR2	DG1	A	300	ND	130	ND
2X22	TCU	A	1271	22	7.8	ND
2X23	TCU	A	1271	22	7.8	ND
3FNE	8PC	A	400	29	ND	ND
3FNF	JPM	A	400	51	ND	ND
3FNG	JPL	A	400	110	ND	ND
3OEW	NAD	A	4345	ND	ND	1500
3OEY	NAD	A	4345	ND	ND	3500
3OF2	NAD	A	4345	ND	ND	4700
4BGE	PYW	A	1270	6000	ND	ND
4BII	PYW	A	1270	6000	ND	ND
4BQP	VMY	A	1272	3	13.7	3108.57
4BQR	IBH	A	1271	200	ND	ND
4COD	KV1	B	1270	34	ND	ND
4OHU	2TK	A	301	ND	0.2	ND
4OIM	JUS	A	302	ND	2.14	ND
4OXK	1S5	A	301	ND	40	ND
4OXN	1S5	A	302	ND	40	ND
4OXY	1TN	A	302	ND	129	ND
4OYR	1US	A	302	ND	0.96	ND
4TRJ	665	A	501	890	ND	ND
4TZK	641	A	501	390	ND	ND
4TZT	468	A	501	23120	ND	ND
4U0J	566	A	401	10660	ND	ND
4U0K	744	A	501	970	ND	ND
5COQ	TCU	A	301	22	ND	ND
5CP8	TCU	A	301	22	ND	ND

ND: Not determined

[1]Search performed on January 3, 2018.

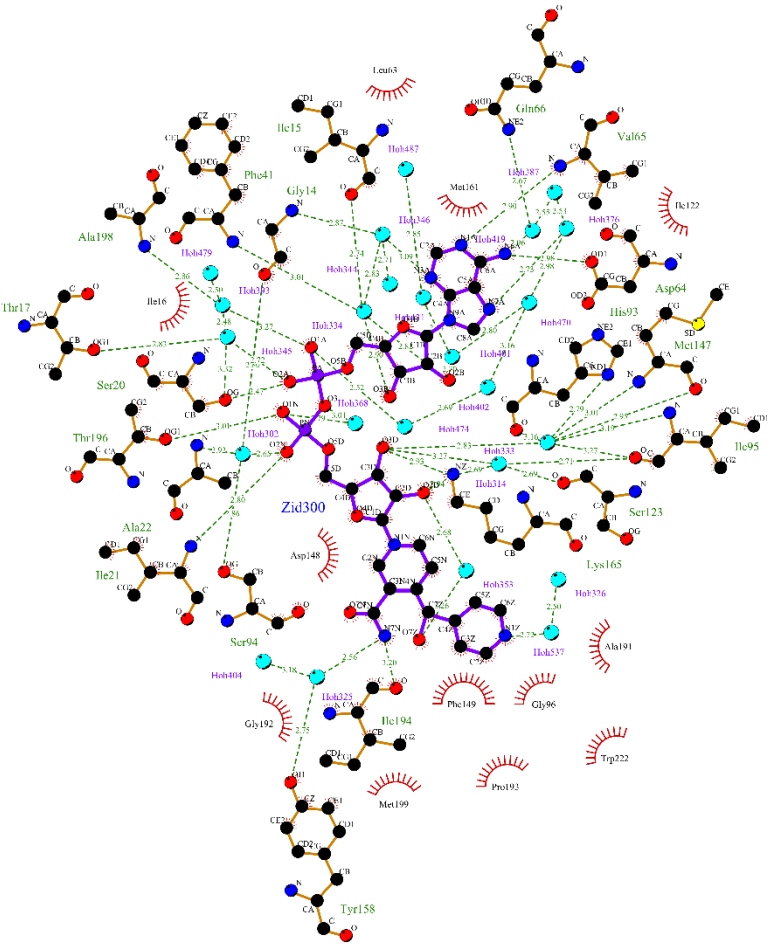


Fig. (2). Analysis of protein-ligand interactions for the crystallographic structure of InhA in complex with NADH-INH (PDB access code: 2IDZ) [30] using the program LigPlot+. Here we represent intermolecular hydrogen bonds as dashed lines. The program LigPlot+ shows the complete structures of the residues involved in the intermolecular hydrogen bonds. The program LigPlot+ depicts other intermolecular interactions indicating the residues as spoked arcs. The isolated spheres in the figure indicate water molecules involved in intermolecular hydrogen bonds and water bridges. The distance of acceptor and donor atoms participating in intermolecular hydrogen bonds are indicated in Å.

One of the most interesting targets for the development of antitubercular drugs are isoniazid-resistant enoyl-ACP(CoA) reductase enzymes due to their presence in the multiresistant and extensively drug-resistant (MDR/XDR) tuberculosis, for a recent review see [33]. The resolution of the crystallographic structures of wild-type, and isoniazid-resistant enoyl-ACP(CoA) reductase enzymes from *Mycobacterium tuberculosis* [34] paved the way to carry out molecular docking simulations focused on relevant isoniazid-resistant InhA. These docking studies had as focus the identification of novel potential inhibitors, many of these computational studies identified true inhibitors of InhA [35-39].

The basic idea behind any docking simulation is to computationally determine the position of the potential binder into the structure of a receptor, for recent reviews see [40-45]. Several docking programs were successfully applied to docking of inhibitors to InhA, for instance: AutoDock and AutoDock Vina [46-49], Glide [50-51], eHiTS [52], and MVD [29]. Our focus here is on protein-ligand interactions, so the binders are small molecules (InhA inhibitors), and the receptor is the structure of InhA. In the next section, we describe the quality of the structural data used to assess protein-ligand interactions and the structural basis for inhibition of InhA.

3. RESULTS AND DISCUSSION

3.1. Structural Data

We applied the program SAnDReS to analyze the InhA dataset, which allows us to have a general view of the quality of the crystallographic structures to be used in the analysis of protein-ligand interactions. We show the overall results in Table **2**. The crystallographic resolution of this dataset ranges from 1.6 to 2.8 Å, being the structure 4OHU [53] the highest resolution one. The histogram plot for the crystallographic resolution indicates that this dataset is right-skewed. Most of the structures present resolution poorer than 2.0 Å, as shown in Fig. (**3A**).

One of the most critical parameter to evaluate the quality of the crystallographic model of a macromolecular structure is the R-factor, which shows the agreement between the experimental crystallographic data and the molecule model. Fig. (**3B**) shows the histogram plot for the R-factor distribution, as seen for the crystallographic resolution data, R-factor is also right-skewed.

A more robust parameter to evaluate the agreement between the model and the X-ray data is the R-free. The program X-Plor introduced the calculation of R-free for crystallographic refinement [55, 56]. We use the same mathematical expression to calculate R-free and R-factor [55, 56], but R-free employs a test set consisting of a small percentage (in the range of 5-10%) of X-ray data omitted from a structure refinement. Fig. (**3C**) shows the histogram plot for the R-factor distribution, as seen for the R-factor, R-free is also right skewed, but with a peak displaced to higher values.

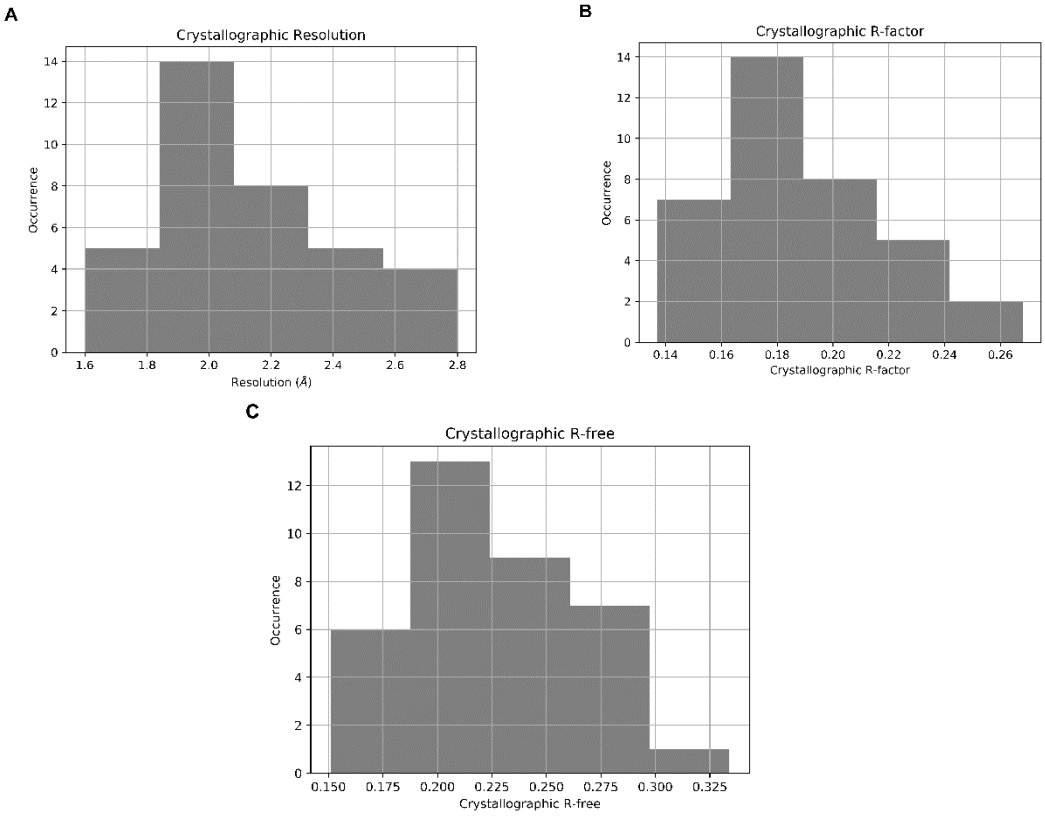


Fig. (3). Crystallographic information for the InhA dataset. A) Resolution. B) R-factor. C) R-free.

Table 2. Analysis of the crystallographic data available for InhA from *Mycobacterium tuberculosis*.

Crystallographic Parameter	Mean Values	Minimum	PDB	Maximum	PDB
Crystallographic high-resolution limit (Å)	2.121	1.6	4OHU	2.8	2B36
Crystallographic low-resolution limit (Å)	39.218	10	2B35	91.29	2X22
Crystallographic R-factor	0.186	0.137	4TZK	0.268	2B35
Crystallographic R-free	0.228	0.151	4TZK	0.334	2B35
Mean B-value (active ligand) ($Å^2$)	33.279	5.333	2X23	60.813	5CP8
Mean occupancy (active ligand)	0.977	0.705	4BII	1	1P44
Mean B-value (whole structure) ($Å^2$)	34.621	10.596	2X23	60.374	4COD
Mean occupancy (whole structure)	0.991	0.934	4OIM	1	1P44
Mean B-value (main-chain protein atoms) ($Å^2$)	32.807	9.551	2X23	61.265	4COD
Mean occupancy (main-chain protein atoms)	0.994	0.925	4OIM	1	1P44
Mean B-value (side-chain protein atoms) ($Å^2$)	35.283	11.188	2X23	60.231	4COD
Mean occupancy (side-chain protein atoms)	0.989	0.938	4OIM	1	1P44

We expect this general behavior when comparing R-factor with R-free, R-free values tend to be higher since it is calculated using X-ray diffraction data omitted from the crystallographic refinement. The crystallographic R-free shows a variation from 0.15 to 0.334, against a variation from 0.137 to 0.268 for the R-factor.

Considering R-factor and R-free values, the worse structure is the 2B35 [54]. The structures 2B35 and 2B36 are the only ones to show no water molecules in the final models, which is most likely due to the poor quality of the electron density maps of these structures. The structure 2B36 was the most reduced resolution structure in the InhA dataset.

Mean B-values can be used to assess the flexibility of a determined region of the protein structure [55]. Fig. (**4A**), Fig. (**4B**), and Fig. (**4C**) show the distribution for the mean-values of active ligand, main-chain, and side-chain atoms, respectively. For instance, from Table **2**, Fig. (**4B**), and (4C) we see that side-chain atoms show higher mean B-value when compared to the main-chain protein atoms. In general, we observe this behavior of the B-values, side-chain protein atoms show higher mean B-factors than those found for main-chain atoms [57]. This difference in B-values is probably due to the flexibility of the side chains, which allow them to oscillate with higher amplitude when compared with main-chain atoms.

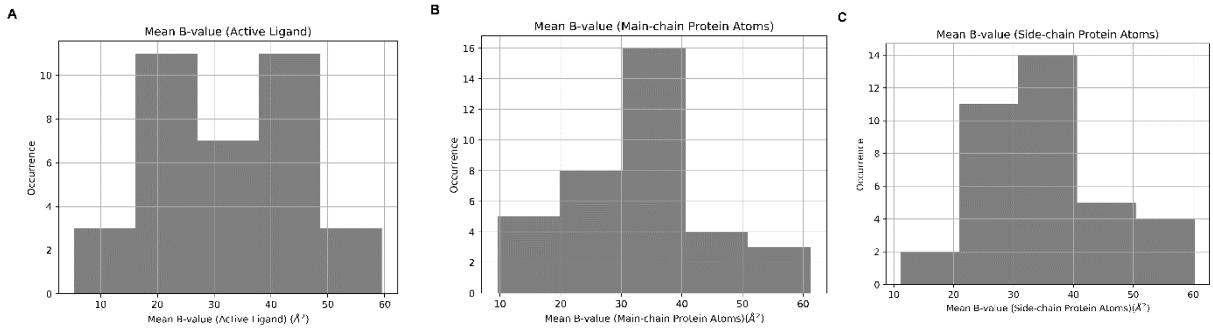


Fig. (4). Mean B-value for the InhA dataset. A) Ligand. B) Main-chain protein atoms. C) Side-chain protein atoms.

Also, ligand B-value indicates the overall quality of the complex structure. For instance, the structures with the lowest and the highest mean B-values in the InhA dataset present the same ligand bound the protein structure (TCU), PDB access codes 2X23 and 5CP8, respectively. Considering that, we have same protein and the same ligand, this difference in the mean B-value for the ligand is most likely due to the overall quality of the X-ray diffraction data.

The structure with the highest mean B-value for the ligand (5CP8) has the reduced resolution, and R-factors (2.4 Å, R-Free= 0.262, and R-factor = 0.206) and the structure 2X23 shows better overall quality (resolution of 1.81 Å, R-Free = 0.203, and R-factor = 0.168). Fig. (**4A**) shows the histogram for the B-values of active ligands in all structures of the InhA dataset. The histogram exhibits a bimodal distribution, with a peak observed around 22 $Å^2$ and the second peak close to 42 $Å^2$.

This bimodal distribution occurs when we have two processes with different distributions merged in one dataset. Analysis of the structures comprising both peaks, reveals that the structures of the first peak (2IDZ, 2NTJ, 2X22, 4BGE, 4BII, 4BQP, 4OXK, 4TRJ, 4TZT, 4U0J, 4U0K) present better crystallographic resolution with a mean value of 1.97 Å. For the structures of the second peak (2B35, 2B37, 2PR2, 3FNE, 3FNF, 3OEW, 4BQR, 4OIM, 4OXN, 4OYR, 5COQ), we have a mean value of 2.24 Å.

In the structures of InhA dataset, the complexes were obtained using either co-crystallization or soaking experiments [58]. In both experimental setups, the ligands showed clear electron density with full occupancy for most of the structure, except for the structure 4BII [59], which showed occupancy factor below 1.0 for the ligand atoms. This low occupancy might be due to the low affinity of the ligand (Pyridomycin) (IC_{50} = 6000 nM) against a mean IC_{50} of 2393.75 nM. Furthermore, the low solubility of the ligand Pyridomycin in the crystallization conditions also contribute to the low occupancy of the ligand.

3.2. Structure of InhA

The structure of InhA shows a typical Rossmann-fold pattern, comprising of a central beta-sheet constituted of parallel strands and flanked by alpha-helices (Fig. (**5**)). This fold is shared with other enzymes from the short-chain dehydrogenase/reductase (SDR) superfamily [60]. Analysis of the binding pocket reveals a cavity with a volume of 532.5 $Å^3$. As we can see in Fig. (**5**), this binding pocket is large enough to accommodate the NADH molecule with an empty volume to be occupied by inhibitors.

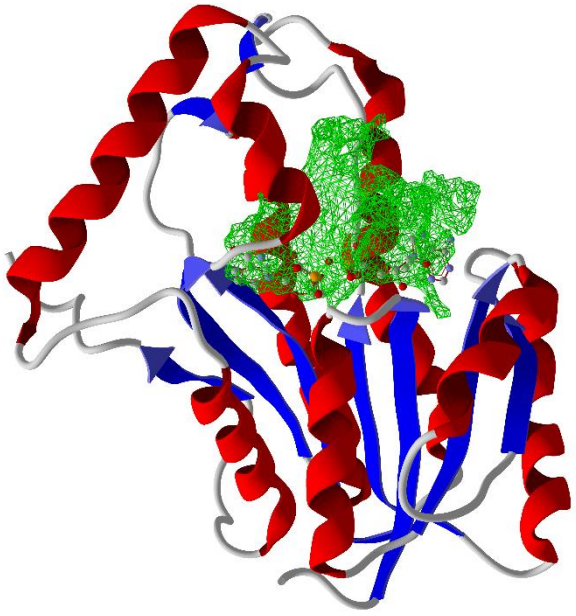


Fig. (5). Crystallographic structure of the complex InhA-NADH (PDB access code: 2AQ8) [34]. We generated this figure with the program MVD [26]. Binding pocket is shown with grid lines. The volume of the cavity was determined using a probe with a radius of 1.2 Å and a grid resolution of 0.8 Å.

3.3. Structural Basis for Binding of Inhibitors to InhA

The richness of structural information about the complexes involving InhA and inhibitors makes possible to evaluate the main structural features responsible for ligand-binding affinity. We used the program SAnDReS to evaluate the intermolecular contact involving all structures in the InhA dataset. Fig. (**6**) shows the bar plot of the number of intermolecular contacts against the residue number. We considered all intermolecular contacts with distance less than 3.5 Å. Taking all structures in the dataset, the residues Ile 21, Phe41, Ser 94, Ile 95, Gly 96, Phe 97, Met 103, Phe 149, Tyr 158(top), Met 161, Pro 193, Ile 194, Ala 198, Met 199, Ile 202, Ile 215, and Leu 218 show most of the intermolecular contacts.

Examination of this set of residues reveals some interesting structural, biological, and chemical features of the inhibition of InhA. The residues Ile 21, Ser 94, Ile 95, Met 103, Met 161, and Ile 194 have been previously identified in isoniazid-resistant *Mycobacterium tuberculosis* [61-63]. This isoniazid-resistant strain is the major health public concern, due to the fail of tradition treatment of tuberculosis [64-71]. Considering the structural approach to this problem, we could say that due the availability of the crystallographic structures for complexes of InhA with inhibitors, our statistical analysis of the major residues involved in the binding of InhA to ligands was surprisingly able to reveal the potential target residues for mutations related isoniazid-resistant *Mycobacterium tuberculosis*.

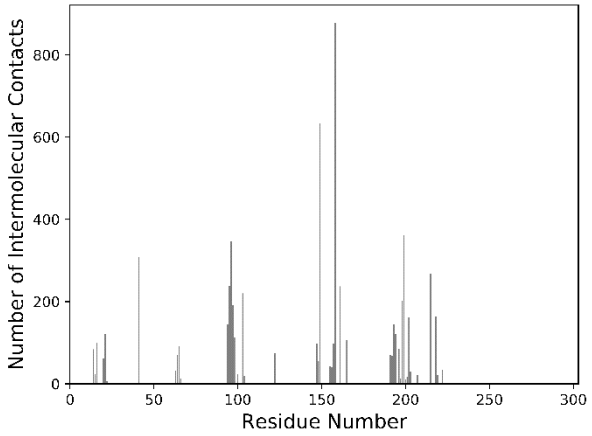


Fig. (6). The occurrence of intermolecular contacts observed for the structures in the InhA dataset. We generated this figure with the program SAnDReS using a cutoff distance of 3.5 Å.

Furthermore, Tyr 158 was the residue that presented the highest number of intermolecular contacts (877 interactions) against 633 contacts observed for Phe 149, the second position residue. A previously published study showed that Y158F InhA mutant is much less sensitive to inhibition by triclosan [61], which corroborates the importance of the Tyr 158 for inhibitor binding. The structure of the complex of InhA with triclosan is available (PDB access code: 3FNF) [72, 73]. Fig. (**7**) shows the intermolecular interactions for triclosan and InhA generated with the program LigPlot+.

The crystallographic structure (3FNF) reveals only one intermolecular hydrogen bond of InhA and triclosan involving the hydroxyl group of the side chain of the residue Tyr 158. The program LigPlot+ identified intermolecular contacts of triclosan and NADH, but they don't involve intermolecular hydrogen bonds. LigPlot+ identified van der Waals contacts comprising the residues Gly 96, Phe 97, Met 103, Phe 149, Met 161, Lys 165, Ala 198, and Glu 219. Among these residues, the Met 103 and Met 161 have been found mutated in isoniazid-resistant *Mycobacterium tuberculosis*. Taking together, we could say that the hydroxyl of the side chain of Tyr 158 is of pivotal importance for triclosan biding since the Y158F mutant has been shown to be less sensitive to this inhibitor [61].

Computational methods have been shown to evaluate ligand-binding affinity with a higher correlation between experimental and theoretical affinities [74-80]. Also, application of supervised machine learning methods can generate even better computational models to predict binding affinity [81-83]. We applied the scoring functions available in the programs AutoDock Vina [84] and MVD [26] to predict binding affinity (log(IC_{50})) to all crystallographic structures in the InhA dataset (supplementary materials 1 and 2). We should highlight that the crystallographic coordinates of the inhibitors were used to evaluate the binding affinity. We did not carry out any docking simulations for computational evaluation of the binding affinity. Statistical analysis of these data shows that the Spearman's rank correlation coefficient ranges from -0.526 to 0.557. We found the highest correlation for the scoring function Ligand Efficiency 1 (LE1), which is calculated by the MolDock Scoring function divided by the number of atoms in the ligand. The result obtained for the LE1 is better than previously published studies [52] for InhA.

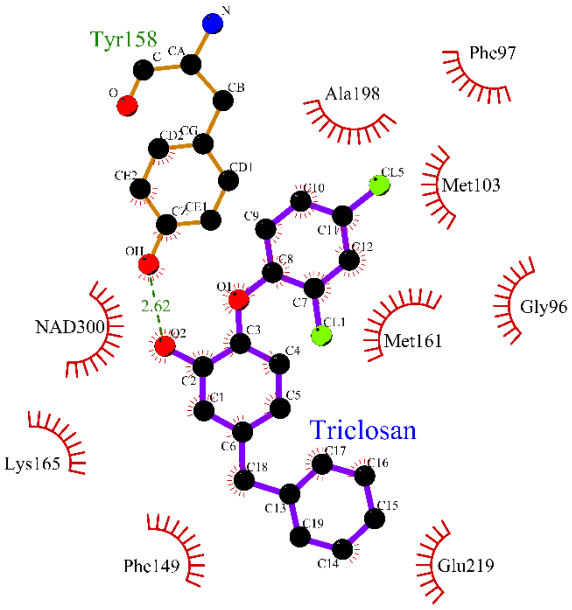


Fig. (7). Analysis of protein-ligand interactions for the structure 3FNF [72, 73] using LigPlot+.

Analysis of the main residues involved in intermolecular contacts (Fig. (**2**) and Fig. (**7**)) of the inhibitors with InhA reveals some structural features that explain the reason for the success of the LE1 in predicting binding affinity. Considering the major residues participating in the binding to the inhibitors, we find participation of polar and hydrophobic and polar side chains. This balance is evaluated in an elegant equation implemented in the LE1 scoring function [26], where a term for electrostatic interactions is considered. Also, terms for hydrogen bond and hydrophobic interactions are present in the scoring function. The weights for each term were obtained using linear regression taking complexes from dataset of high-resolution crystallographic structures for which binding affinity data is available [26]. LE1 scoring function has been shown superior predictive power for other protein targets when compared with Plants, AutoDock 4, and AutoDock Vina scoring functions [24, 28, 29, 85-90].

3.4. Future developments

The increasing number of crystallographic structures available for complexes involving InhA and inhibitors is most likely due to the easiness in obtaining purified InhA in quantity and purity necessary to carry out crystallographic experiments [34]. The availability of high intense X-ray sources found in synchrotron facilities also contributes to increasing the number of structures of InhA and the overall quality of the structural information [91]. Moreover, the development of computational models to predict

binding affinity together with docking simulation program makes possible to test up to millions of potentials new inhibitors. Considering the integration of X-ray diffraction studies, computational methods, and experimental evaluation of binding affinity, we expect that the increasing number of structures of InhA will pave the way to the design more specific inhibitors of InhA. The use of modern docking programs [26, 84, 92] and integration of automatic workflow to virtual screening studies have the potential of speeding up the identification of new potential inhibitors for InhA.

Furthermore, considering that we have an ensemble of crystallographic structures for which binding affinity is available, we could use this structural data to build machine-learning models targeted to the InhA [24, 93, 94]. As a potential source of inhibitors, we may take small molecule structures available in the ZINC database [95-100].

4. CONCLUSION

In the present work, we described a dataset of crystallographic structures available for InhA in complex with inhibitors. Analysis of these structures revealed the central residues responsible for binding affinity. The mapping of the intermolecular contacts showed the importance of the residue Tyr 158 for ligand binding. This structural feature has been confirmed by a mutational study [61]. Furthermore, several of the top residues involved in intermolecular contacts (Ile 21, Ser 94, Ile 95, Met 103, Met 161, and Ile 194) have been found in isoniazid-resistant *Mycobacterium tuberculosis* [61-63]. Also, application of ligand efficiency scoring function available in program MVD has been shown to evaluate binding affinity with a moderate uphill relationship between the predicted and experimental binding affinity with superior predictive power when compared with a previous study [46-52]. The combination of computational methods (docking, virtual screening, and machine learning models), X-ray diffraction studies and experimental evaluation of the binding affinity has the potential to bring a new generation of InhA inhibitors that could treat multiresistant and extensively drug-resistant tuberculosis.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

This work was supported by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) (Process Numbers: 472590/2012-0 and 308883/2014-4). MBA acknowledges support from CAPES. GBF acknowledges the receipt of a fellowship from Programa de Educação Tutorial-Biologia (PUCRS) (Brazil). WFA is a senior researcher for CNPq (Brazil).

5. REFERENCES

[1] Kaczor, A.A.; Polski, A.; Sobótka-Polska, K.; Pachuta-Stec, A.; Makarska-Bialokoz, M.; Pitucha, M. Novel Antibacterial Compounds and their Drug Targets - Successes and Challenges. *Curr. Med. Chem.*, **2017**, *24*(18), 1948-1982.

[2] Ichikawa, S.; Yamaguchi, M.; Matsuda, A. Antibacterial Nucleoside Natural Products Inhibiting Phospho-MurNAc-Pentapeptide Translocase; Chemistry and Structure-Activity Relationship. *Curr. Med. Chem.*, **2015**, *22*(34), 3951-3979.

[3] Chiarelli, L.R.; Mori, G.; Esposito, M.; Orena, B.S.; Pasca, M.R. New and Old Hot Drug Targets in Tuberculosis. *Curr. Med. Chem.*, **2016**, *23*(33), 3813-3846.

[4] Meneghetti, F.; Villa, S.; Gelain, A.; Barlocco, D.; Chiarelli, L.R.; Pasca, M.R.; Costantino, L. Iron Acquisition Pathways as Targets for Antitubercular Drugs. *Curr. Med. Chem.*, **2016**, *23*(35), 4009-4026.

[5] Dos Santos Fernandes, G.F.; Jornada, D.H.; De Souza, P.C.; Chin, C.M.; Pavan, F.R.; Dos Santos, J.L. Current Advances in Antitubercular Drug Discovery: Potent Prototypes and New Targets. *Curr. Med. Chem.*, **2015**, *22*(27), 3133-3161.

[6] Fanzani, L.; Porta, F.; Meneghetti, F.; Villa, S.; Gelain, A.; Lucarelli, A.P.; Parisini, E. *Mycobacterium tuberculosis* Low Molecular Weight Phosphatases (MPtpA and MPtpB): From Biological Insight to Inhibitors. *Curr. Med. Chem.*, **2015**, *22*(27), 3110-3132.

[7] Coracini, J.D.; de Azevedo, W.F. Jr. Shikimate kinase, a protein target for drug design. *Curr. Med. Chem.*, **2014**, *21*(5), 592-604.

[8] Inturi, B.; Pujar, G.V.; Purohit, M.N. Recent Advances and Structural Features of Enoyl-ACP Reductase Inhibitors of *Mycobacterium tuberculosis*. *Arch. Pharm. (Weinheim)*, **2016**, *349*(11), 817-826.

[9] Punkvang, A.; Saparpakorn, P.; Hannongbua, S.; Wolschann, P.; Pungpo, P. Elucidating drug-enzyme interactions and their structural basis for improving the affinity and potency of isoniazid and its derivatives based on computer modeling approaches. *Molecules*, **2010**, *15*(4), 2791-2813.

[10] Punkvang, A.; Saparpakorn, P.; Hannongbua, S.; Wolschann, P.; Beyer, A.; Pungpo, P. Investigating the structural basis of arylamides to improve potency against *M*. *tuberculosis* strain through molecular dynamics simulations. *Eur. J. Med. Chem.*, **2010**, *45*(12), 5585-5593.

[11] Pan, P.; Tonge, P.J. Targeting InhA, the FASII enoyl-ACP reductase: SAR studies on novel inhibitor scaffolds. *Curr. Top. Med. Chem.*, **2012**, *12*(7), 672-693.

[12] Smith, S.; Witkowski, A.; Joshi, A.K. Structural and functional organization of the animal fatty acid synthase. *Prog. Lipid Res.*, **2003**, *42*(4), 289-317.

[13] Parsons, J.B.; Rock, C.O. Is bacterial fatty acid synthesis a valid target for antibacterial drug discovery?. *Curr. Opin. Microbiol*, **2011**, *14*(5), 544-549.

[14] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.*, **2000**, *28*(1), 235-242.

[15] Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J.D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **2002**, *58*(Pt 6 No 1), 899-907.

[16] Westbrook, J.; Feng, Z.; Chen, L.; Yang, H.; Berman, H.M. The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **2003**, *31*(1), 489-491.

[17] Benson, M.L.; Smith, R.D.; Khazanov, N.A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H.A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.*, **2008**, *36*(Database issue), D674-D678.

[18] Ahmed, A.; Smith, R.D.; Clark, J.J.; Dunbar Jr., J.B.; Carlson, H.A. Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Res.*, **2015**, *43*(Database issue), D465-D469.

[19] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, **2007**, *35*(Database issue), D198-D201.

[20] Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **2016**, *44*(D1), D1045-D1053.

[21] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **2004**, *47*(12), 2977-2980.

[22] Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **2015**, *31*(3), 405-412.

[23] Fadel, V.; Bettendorff, P.; Herrmann, T.; de Azevedo W.F. Jr.; Oliveira, E.B.; Yamane, T.; Wüthrich, K. Automated NMR structure determination and disulfide bond identification of the myotoxin crotamine from *Crotalus durissus terrificus*. *Toxicon*, **2005**, *46*(7),759-767.

[24] Xavier, M.M.; Heck, G.S.; de Avila, M.B.; Levin, N.M.; Pintro, V.O.; Carvalho, N.L.; Azevedo, W.F. Jr. SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions. *Comb. Chem. High Throughput Screen*, **2016**, *19*(10), 801-812.

[25] Humphrey, W.; Dalke, A; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, **1996**, *14*(1), 33-38.

[26] Thomsen, R.; Christensen, M.H. MolDock: a new technique for high-accuracy molecular docking. *J. Med. Chem.*, **2006**, *49*(11), 3315-3321.

[27] Heberlé, G.; de Azevedo, W.F. Jr. Bio-inspired algorithms applied to molecular docking simulations. *Curr. Med. Chem.*, **2011**, *18*(9), 1339-1352.

[28] De Azevedo, W.F. Jr. MolDock applied to structure-based virtual screening. *Curr. Drug Targets*, **2010**, *11*(3), 327-334.

[29] Azevedo, L.S.; Moraes, F.P.; Xavier, M.M.; Pantoja, E.O.; Villavicencio, B.; Finck, J.A.; Proenca, A.M.; Rocha, K.B.; de Azevedo, W.F. Jr. Recent Progress of Molecular Docking Simulations Applied to Development of Drugs. *Curr. Bioinform.*, **2012**, *7*(4), 352-365.

[30] Dias, M.V.; Vasconcelos, I.B.; Prado, A.M.; Fadel, V.; Basso, L.A.; de Azevedo, W.F. Jr.; Santos, D.S. Crystallographic studies on the binding of isonicotinyl-NAD adduct to wild-type and isoniazid resistant 2-trans-enoyl-ACP (CoA) reductase from *Mycobacterium tuberculosis*. *J. Struct. Biol.*, **2007**, *159*(3), 369-380.

[31] Wallace, A.C.; Laskowski, R.A.; Thornton, J.M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **1995**, *8*(2), 127-134.

[32] Laskowski, R.A.; Swindells, M.B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model*, **2011**, *51*(10), 2778-2786.

[33] Schön, T.; Miotto, P.; Köser, C.U.; Viveiros, M.; Böttger, E.; Cambau, E. *Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin. Microbiol. Infect*, **2017**, *23*(3), 154-160.

[34] Oliveira, J.S.; Pereira, J.H.; Canduri, F.; Rodrigues, N.C.; de Souza, O.N.; de Azevedo, W.F. Jr.; Basso, L.A.; Santos, D.S. Crystallographic and pre-steady-state kinetics studies on binding of NADH to wild-type and isoniazid-resistant enoyl-ACP(CoA) reductase enzymes from *Mycobacterium tuberculosis. J. Mol. Biol.*, **2006**, *359*(3), 646-666.

[35] Sharma, S.K.; Kumar, G.; Kapoor, M.; Surolia, A. Combined effect of epigallocatechin gallate and triclosan on enoyl-ACP reductase of *Mycobacterium tuberculosis*. *Biochem. Biophys. Res. Commun.*, **2008**, *368*(1), 12-17.

[36] Ghorab, M.M.; El-Gaby, M.S.A; Soliman, A.M.; Alsaid, M.S.; Abdel-Aziz, M.M.; Elaasser, M.M. Synthesis, docking study and biological evaluation of some new thiourea derivatives bearing benzenesulfonamide moiety. *Chem. Cent. J.*, **2017**, *11*(1), 42.

[37] Chiarelli, L.R.; Mori, G.; Esposito, M.; Orena, B.S.; Pasca, M.R. New and Old Hot Drug Targets in Tuberculosis. *Curr. Med. Chem.*, **2016**, *23*(33), 3813-3846.

[38] Dhumal, S.T.; Deshmukh, A.R.; Bhosle, M.R.; Khedkar, V.M.; Nawale, L.U.; Sarkar, D.; Mane, R.A. Synthesis and antitubercular activity of new 1,3,4-oxadiazoles bearing pyridyl and thiazolyl scaffolds. *Bioorg. Med. Chem. Lett.*, **2016**, *26*(15), 3646-3651.

[39] Desai, N.C.; Somani, H.; Trivedi, A.; Bhatt, K.; Nawale, L.; Khedkar, V.M.; Jha, P.C.; Sarkar, D. Synthesis, biological evaluation and molecular docking study of some novel indole and pyridine based 1,3,4-oxadiazole derivatives as potential antitubercular agents. *Bioorg. Med. Chem. Lett.*, **2016**, *26*(7), 1776-1783.

[40] Herrera Acevedo, C.; Scotti, L.; Feitosa Alves, M.; Formiga Melo Diniz, M.D.F.; Scotti, M.T. Computer-Aided Drug Design Using Sesquiterpene Lactones as Sources of New Structures with Potential Activity against Infectious Neglected Diseases. *Molecules*, **2017**, *22*(1), 79.

[41] Alves, M.F.; Scotti, M.T.; Scotti, L.; Mendonça, F.J.B.; Filho, J.M.B.; de Melo, S.A.L.; Dos Santos, S.G.; Diniz, M.F.F.M. Secondary Metabolites from Cissampelos, A Possible Source for New Leads with Anti-Inflammatory Activity. *Curr. Med. Chem.*, **2017**, *24*(16), 1629-1644.

[42] Lorenzo, V.P.; Lúcio, A.S.; Scotti, L.; Tavares, J.F.; Filho, J.M.; Lima, T.K.; Rocha, J.D.; Scotti, M.T. Structure- and Ligand-Based Approaches to Evaluate Aporphynic Alkaloids from Annonaceae as Multi-Target Agent Against *Leishmania donovani*. *Curr. Pharm. Des.*, **2016**, *22*(34), 5196-5203.

[43] Scotti, L.; Mendonca Junior, F.J.; Ishiki, H.M.; Ribeiro, F.F.; Singla, R.K.; Barbosa Filho, J.M.; Da Silva, M.S.; Scotti, M.T. Docking Studies for Multi-Target Drugs. *Curr. Drug Targets*, **2017**, *18*(5), 592-604.

[44] Scotti, L.; Scotti, M.T. Computer Aided Drug Design Studies in the Discovery of Secondary Metabolites Targeted Against Age-Related Neurodegenerative Diseases. *Curr. Top. Med. Chem.*, **2015**, *15*(21), 2239-2252.

[45] Scotti, L.; Bezerra Mendonça Junior, F.J.; Magalhaes Moreira, D.R.; da Silva, M.S.; Pitta, I.R.; Scotti, M.T. SAR, QSAR and docking of anticancer flavonoids and variants: a review. *Curr. Top. Med. Chem.*, **2012**, *12*(24), 2785-2809.

[46] Shilpi, J.A.; Ali, M.T.; Saha, S.; Hasan, S.; Gray, A.I.; Seidel, V. Molecular docking studies on InhA, MabA and PanK enzymes from *Mycobacterium tuberculosis* of ellagic acid derivatives from *Ludwigia adscendens* and *Trewia nudiflora*. *In Silico Pharmacol*, **2015**, *3*(1), 10.

[47] Stigliani, J.L.; Bernardes-Génisson, V.; Bernadou, J.; Pratviel, G. Cross-docking study on InhA inhibitors: a combination of Autodock Vina and PM6-DH2 simulations to retrieve bio-active conformations. *Org. Biomol. Chem.*, **2012**, *10*(31), 6341-6349.

[48] Punkvang, A.; Saparpakorn, P.; Hannongbua, S.; Wolschann, P.; Pungpo, P. Elucidating drug-enzyme interactions and their structural basis for improving the affinity and potency of isoniazid and its derivatives based on computer modeling approaches. *Molecules*, **2010**, *15*(4), 2791-2813.

[49] Stigliani, J.L.; Arnaud, P.; Delaine, T.; Bernardes-Génisson, V.; Meunier, B.; Bernadou, J. Binding of the tautomeric forms of isoniazid-NAD adducts to the active site of the *Mycobacterium tuberculosis* enoyl-ACP reductase (InhA): a theoretical approach. *J Mol. Graph. Model.*, **2008**, *27*(4), 536-545.

[50] Saharan, V.D.; Mahajan, S.S. Development of gallic acid formazans as novel enoyl acyl carrier protein reductase inhibitors for the treatment of tuberculosis. *Bioorg. Med. Chem. Lett.*, **2017**, *27*(4), 808-815.

[51] Bhatt, J.D.; Chudasama, C.J.; Patel, K.D. Pyrazole clubbed triazolo[1,5-a]pyrimidine hybrids as an anti-tubercular agents: Synthesis, in vitro screening and molecular docking study. *Bioorg. Med. Chem.*, **2015**, *23*(24), 7711-7716.

[52] Kinnings, S.L.; Liu, N.; Tonge, P.J.; Jackson, R.M.; Xie, L.; Bourne, P.E. A machine learning based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.*, **2011**, *51*(2), 408-419.

[53] Li, H.J.; Lai, C.T.; Pan, P.; Yu, W.; Liu, N.; Bommineni, G.R.; Garcia-Diaz, M.; Simmerling, C.; Tonge, P.J. A structural and energetic model for the slow-onset inhibition of the *Mycobacterium tuberculosis* enoyl-ACP reductase InhA. *ACS Chem. Biol.*, **2014**, *9*(4), 986-993.

[54] Sullivan, T.J.; Truglio, J.J.; Boyne, M.E.; Novichenok, P.; Zhang, X.; Stratton, C.F.; Li, H.J.; Kaur, T.; Amin, A.; Johnson, F.; Slayden, R.A.; Kisker, C.; Tonge, P.J. High affinity InhA inhibitors with activity against drug-resistant strains of *Mycobacterium tuberculosis*. *ACS Chem. Biol.*, **2006**, *1*(1), 43-53.

[55] Brünger, A.T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **1992**, *355*(6359), 472-475.

[56] Brünger, A.T. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta. Crystallogr. D. Biol. Crystallogr.*, **1993**, *49*(1), 24-36.

[57] Delatorre, P.; de Azevedo Jr., W.F. Simulation of electron density maps for twodimensional crystal structures using Mathematica. *J. Appl. Cryst.*, **2001**, *34*(5), 658-660.

[58] de Azevedo, W.F. Jr.; Canduri, F.; Basso, L.A.; Palma, M.S.; Santos, D.S. Determining the structural basis for specificity of ligands using crystallographic screening. *Cell Biochem. Biophys.*, **2006**, *44*(3), 405-411.

[59] Hartkoorn, R.; Pojer, F.; Read, J.A.; Gingell, H.; Neres, J.; Horlacher, O.; Altmann, K.H.; Cole, S. Pyridomycin bridges the NADH- and substrate-binding pockets of the enoyl reductase InhA. *Nat. Chem. Biol.*, **2014**, *10*(2), 96.

[60] Persson, B.; Kallberg, Y.; Oppermann, U.; Jornvall, H. Coenzyme-based functional assignments of short-chain dehydrogenases/reductases (SDRs). *Chem. Biol. Interact*. **2003**; *143–144*, 271–278.

[61] Parikh, S.L.; Xiao, G.; Tonge, P.J. Inhibition of InhA, the enoyl reductase from *Mycobacterium tuberculosis*, by triclosan and isoniazid. *Biochemistry*, **2000**, *39*(26), 7645-7650.

[62] Seifert, M.; Catanzaro, D.; Catanzaro, A.; Rodwell, T.C. Genetic Mutations Associated with Isoniazid Resistance in *Mycobacterium tuberculosis*: A Systematic Review. *PLoS One*, **2015**, *10*(3), e0119628.

[63] Basso, L.A.; Zheng, R.; Musser, J.M.; Jacobs, W.R. Jr.; Blanchard, J.S. Mechanisms of isoniazid resistance in *Mycobacterium tuberculosis*: enzymatic characterization of enoyl reductase mutants identified in isoniazid-resistant clinical isolates. *J. Infect. Dis.*, **1998**, *178*(3), 769-775.

[64] Bedewi Omer, Z.; Mekonnen, Y.; Worku, A.; Zewde, A.; Medhin, G.; Mohammed, T.; Pieper, R.; Ameni, G. Evaluation of the GenoType MTBDRplus assay for detection of rifampicin- and isoniazid-resistant *Mycobacterium tuberculosis* isolates in central Ethiopia. *Int. J. Mycobacteriol*, **2016**, *5*(4), 475-481.

[65] Ahmad, B.; Idrees, M.; Ahmad, K.; Bashir, S.; Jamil, S. Molecular characterisation of isoniazid resistant clinical isolates of *Mycobacterium tuberculosis* from Khyber Pakhtunkhwa, Pakistan. *J. Pak. Med. Assoc.*, **2017**, *67*(8), 1224-1227.

[66] Takawira, F.T.; Mandishora, R.S.D.; Dhlamini, Z.; Munemo, E.; Stray-Pedersen, B. Mutations in rpoB and katG genes of multidrug resistant *Mycobacterium tuberculosis* undetectable using genotyping diagnostic methods. *Pan. Afr. Med. J.*, **2017**, *27*(1), 145.

[67] Squeglia, F.; Romano, M.; Ruggiero, A.; Berisio, R. Molecular Players in Tuberculosis Drug Development: Another Break in the Cell Wall. *Curr. Med. Chem.*, **2017**, *24*(36), 3954-3969.

[68] Sgaragli, G.; Frosini, M. Human Tuberculosis I. Epidemiology, Diagnosis and Pathogenetic Mechanisms. *Curr. Med. Chem.*, **2016**, *23*(25), 2836-2873.

[69] Chiarelli, L.R.; Mori, G.; Esposito, M.; Orena, B.S.; Pasca, M.R. New and Old Hot Drug Targets in Tuberculosis. *Curr. Med. Chem.*, **2016**, *23*(33), 3813-3846.

[70] Sgaragli, G.; Frosini, M. Human Tuberculosis II. M. *tuberculosis* Mechanisms of Genetic and Phenotypic Resistance to Anti-Tuberculosis Drugs. *Curr. Med. Chem.*, **2016**, *23*(12), 1186-1216.

[71] Sharma, R.; Kaur, A.; Sharma, A.K.; Dilbaghi, N.; Sharma, A.K. Nano-Based Anti-Tubercular Drug Delivery and Therapeutic Interventions in Tuberculosis. *Curr. Drug Target*, **2017**, *18*(1), 72-86.

[72] Freundlich, J.S.; Wang, F.; Vilcheze, C.; Gulten, G.; Langley, R.; Schiehser, G.A.; Jacobus, D.P.; Jacobs, W.R.; Sacchettini, J.C. Triclosan Derivatives: Towards Potent Inhibitors of Drug-Sensitive and Drug-Resistant *Mycobacterium tuberculosis*. *ChemMedChem*, **2009**, *4*(2), 241-248.

[73] Holas, O.; Ondrejcek, P.; Dolezal, M. *Mycobacterium tuberculosis* enoyl-acyl carrier protein reductase inhibitors as potential antituberculotics: development in the past decade. *J. Enzyme Inhib. Med. Chem.*, **2015**, *30*(4), 629-648.

[74] Canduri, F.; Perez, P.C.; Caceres, R.A.; de Azevedo W.F. Jr. Protein kinases as targets for antiparasitic chemotherapy drugs. *Curr. Drug Targets*, **2007**, *8*(3), 389-398.

[75] de Azevedo, W.F. Jr.; Dias, R. Evaluation of ligand-binding affinity using polynomial empirical scoring functions. *Bioorg. Med. Chem.*, **2008**, *16*(20), 9378-9382.

[76] de Azevedo, W.F. Jr.; Dias, R. Computational methods for calculation of ligand-binding affinity. *Curr. Drug Targets*, **2008**, *9*(12), 1031-1039.

[77] Dias, R.; de Azevedo, W.F. Jr. Molecular docking algorithms. *Curr. Drug Targets*, **2008**, *9*(12), 1040-1047.

[78] Dias, R.; Timmers, L.F.; Caceres, R.A.; de Azevedo, W.F. Jr. Evaluation of molecular docking using polynomial empirical scoring functions. *Curr. Drug Targets*, **2008**, *9*(12), 1062-1070.

[79] de Azevedo, W.F. Jr.; Dias, R. Experimental approaches to evaluate the thermodynamics of protein-drug interactions. *Curr. Drug Targets*, **2008**, *9*(12), 1071-1076.

[80] De Azevedo, W.F. Jr. Structure-based virtual screening. *Curr. Drug Targets*, **2010**, *11*(3), 261-263.

[81] de Ávila, M.B.; Xavier, M.M.; Pintro, V.O.; de Azevedo, W.F. Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2. *Biochem. Biophys. Res. Commun.*, **2017**, *494* (1-2), 305-310.

[82] Pintro, V.O.; Azevedo, W.F. Optimized Virtual Screening Workflow. Towards Target-Based Polynomial Scoring Functions for HIV-1 Protease. *Comb. Chem. High Throughput Screen*, **2017**, *20*(9), 820-827.

[83] Heck, G.S.; Pintro, V.O.; Pereira, R.R.; de Ávila, M.B.; Levin, N.M.B.; de Azevedo, W.F. Supervised Machine Learning Methods Applied to Predict Ligand-Binding Affinity. *Curr. Med. Chem.*, **2017**, *24*(23), 2459-2470.

[84] Trott, O.; Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **2010**, *31*(2), 455-461.

[85] Levin, N.M.; Pintro, V.O.; de Ávila, M.B.; de Mattos, B.B., De Azevedo, W.F. Jr. Understanding the Structural Basis for Inhibition of Cyclin-Dependent Kinases. New Pieces in the Molecular Puzzle. *Curr. Drug Targets*, **2017**, *18*(9), 1104-1111.

[86] de Azevedo, Jr. W.F. Opinion Paper: Targeting Multiple Cyclin-Dependent Kinases (CDKs): A New Strategy for Molecular Docking Studies. *Curr. Drug Targets*, **2016**, *17*(1), 2.

[87] Teles, C.B.; Moreira-Dill, L.S.; Silva Ade, A.; Facundo, V.A.; de Azevedo, W.F. Jr.; da Silva, L.H.; Motta, M.C.; Stábeli, R.G.; Silva-Jardim, I. A lupane-triterpene isolated from Combretum leprosum Mart. fruit extracts that interferes with the intracellular development of *Leishmania* (L.) *amazonensis in vitro*. *BMC Complement Altern. Med.*, **2015**, *15*(1), 165.

[88] de Ávila, M.B.; de Azevedo, W.F. Data Mining of Docking Results. Application to 3-Dehydroquinate Dehydratase. *Curr. Bioinform.*, **2014**, *9*(4), 361-379.

[89] Moraes, F.P.; de Azevedo, W.F. Jr. Targeting imidazoline site on monoamine oxidase B through molecular docking simulations. *J. Mol. Model*, **2012**, *18*(8), 3877-3886.

[90] Vianna, C.P.; de Azevedo, W.F. Jr. Identification of new potential *Mycobacterium tuberculosis* shikimate kinase inhibitors through molecular docking simulations. *J. Mol Model*, **2012**, *18*(2), 755-764.

[91] Canduri, F.; de Azevedo, W.F. Protein crystallography in drug discovery. *Curr. Drug Targets*, **2008**, *9*(12), 1048-1053.

[92] Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R.; Olson, A. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **1998**, *19*(14), 1639-1662.

[93] Amaral, M.E.A.; Nery, L.R.; Leite, C.E.; de Azevedo Junior, W.F.; Campos, M.M. Pre-clinical effects of metformin and aspirin on the cell lines of different breast cancer subtypes. *Invest New Drugs*, **2018**, doi: 10.1007/s10637-018-0568-y.

[94] Levin, N.M.B.; Pintro, V.O.; Bitencourt-Ferreira, G.; Mattos, B.B.; Silvério, A.C.; de Azevedo, Jr. W.F. Development of CDK-targeted scoring functions for prediction of binding affinity. *Biophys. Chem.*, **2018**, 235, 1–8. https://doi.org/10.1016/j.bpc.2018.01.004.

[95] Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model*, **2012**, *52*(7), 1757-1768.

[96] Irwin, J.J.; Shoichet, B.K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*, **2005**, *45*(1), 177-182.

[97] Sterling, T.; Irwin, J.J. ZINC 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model*, **2015**, *55*(11), 2324-2337.

[98] Gozalbes, R.; Pineda-Lucena, A. Small molecule databases and chemical descriptors useful in chemoinformatics: an overview. *Comb. Chem. High Throughput Screen*, **2011**, *14*(6), 548-458.

[99] Ghasemi, J.B.; Shiri, F.; Pirhadi, S.; Heidari, Z. Discovery of new potential antimalarial compounds using virtual screening of ZINC database. *Comb. Chem. High Throughput Screen*, **2015**, *18*(2), 227-234.

[100] Patel, P.; Singh, A.; Patel, V.K.; Jain, D.K.; Veerasamy, R.; Rajak, H. Pharmacophore Based 3D-QSAR, Virtual Screening and Docking Studies on Novel Series of HDAC Inhibitors with Thiophen Linker as Anticancer Agents. *Comb. Chem. High Throughput Screen*, 2016, *19*(9), 735-751.

ANEXO C – Artigo publicado na revista Biochemical and Biophysical Research Communication. 2018. Fator de Impacto: 2,985.

Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2

Maurício Boff de Ávila [a,b], Mariana Morrone Xavier [a], Val Oliveira Pintro [a], Walter Filgueira de Azevedo Jr. [a,b *]

[a] *Laboratory of Computational Systems Biology, School of Sciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre, RS 90619-900, Brazil*
[b] *Graduate Program in Cellular and Molecular Biology, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre, RS 90619-900, Brazil*

Abstract

Here we report the development of a machine-learning model to predict binding affinity based on the crystallographic structures of protein-ligand complexes. We used an ensemble of crystallographic structures (resolution better than 1.5 Å resolution) for which half-maximal inhibitory concentration (IC_{50}) data is available. Polynomial scoring functions were built using as explanatory variables the energy terms present in the MolDock and PLANTS scoring functions. Prediction performance was tested and the supervised machine learning models showed improvement in the prediction power, when compared with PLANTS and MolDock scoring functions. In addition, the machine-learning model was applied to predict binding affinity of CDK2, which showed a better performance when compared with AutoDock4, AutoDock Vina, MolDock, and PLANTS scores.

Keywords: Bioinformatics, CDK2, Kinase, Drug design, Docking, Machine Learning.

1. INTRODUCTION

Computational analysis of protein-ligand interactions is of pivotal importance for *in silico* drug design. Among the most used computational methods to assess protein-ligand interactions, we could say that the field of scoring function still needs additional improvement. New developments in this field with integration of supervised machine learning (SML) techniques and classical scoring functions have been shown to improve the predictive po- wer of scoring functions [1e7]. Recent published machine learning approaches to predict ligand-binding affinity showed superior predictive performance, when compared with classical scoring functions such as PLANTS, MolDock, AutoDock4, and AutoDock Vina scoring functions [8e10]. Furthermore, combination of ma- chine learning techniques with classical scoring functions opens the possibility to explore a wide spectrum of machine learning models, where the terms used in the classical scoring functions are used to develop a function targeted to the biological system being analyzed [1,8].

One major development in the field of machine learning is the availability of scikit-learn library [11], which allows fast progress in creation of programs to generate machine-learning models using Python programming language. This approach was used in the development of the program SAnDReS (Statistical Analysis of Docking Results and Scoring Functions) [8], which allows building machine-learning models targeted to the biological system of interested.

Here we report the application of an integrated computational methodology to develop scoring functions using SML techniques available in the program SAnDReS [8]. In this approach, we use a dataset of crystallographic structures for which experimental information about binding affinity is known. Our focus is on complexed crystallographic structures, where the active ligand (inhibitor) is not covalently bound to the protein. Application of SML techniques to this dataset generated computational models with better predictive power when compared with standard scoring functions such as MolDock and PLANTS scoring functions [12].

In order to submit our SML models to additional tests, we applied the SML model to predict ligand-binding affinity for a dataset composed of high-resolution structures of cyclin-dependent kinase 2 (EC 2.7.11.22). The structures of CDK2 were not used in the high-resolution dataset, which provides a reliable test set for the predictive performance of the SML model. CDK2 has been chosen because it is an important protein target for development of anticancer drugs [3]. Since the pioneering work of Prof. Sung Hou-Kim at University of California at Berkeley to solve the first structure of CDK2 using X-ray diffraction crystallography [13], there have been over 400 structures of CDK2 determined by X-ray diffraction crystallography [14]. Many of them with inhibitors for which IC_{50} information is available in the crystallographic structure. Application of SML models to CDK2 showed superior predictive performance when compared with AutoDock4 [15], AutoDock Vina [16], MolDock, and PLANTS scores [12].

2. METHODS

2.1 Datasets

We used a dataset composed of an ensemble of high-resolution crystallographic structures solved to resolution better than 1.5 Å, and for which there is experimental data for half-maximal inhibitory concentration (IC50) for the active ligands. The structures and binding information were downloaded from the Protein Data Bank [17]. Repeated ligands were deleted from this dataset and ended up with 173 unique structures (search carried out on July 19th, 2017). This dataset will be referred to as HRIC50 dataset. In order to further validate our SML model, we created a second dataset composed of 11 CDK2 structures, not used in the HRIC50 dataset. We used CDK2 crystallographic structures solved to resolution better than 2.0 Å and for which IC50 data was available. This dataset will be referred to as CDK2IC50 dataset. Table 1 shows the PDB access codes for the structures of both datasets.

Table 1
List of PDB access codes used for both datasets.

$HRIC_{50}$	2GG3,2GG7,2GG9,2HU6,2I5F,2IKG,2NMZ,2NNG,2OW6,2PDG, 2PIY,2PZN,2QCF,2R3I,2W14,2W3B,2W9H,2WUU,2WZX,2X5O, 2XPC,2XU3,2XU4,2Y1O,2Y68,2YC3,2YEX,2YJ2,2YJ8,2YJ9,2YJC, 2YK9,2YKE,2YKJ,3B28,3B7E,3B8Z,3BCJ,3BLB,3CBP,3DCR,3DD0, 3DN5,3EJS,3EJT,3EJU,3ESS,3EWZ,3EX3,3F66,3FCI,3FS6,3GHV, 3GHW,3H5B,3HHA,3HJ0,3HNB,3HS4,3HYG,3I06,3I33,3I6C,3I6O, 3IOG,3IU7,3KFA,3KIG,3KKU,3KL6,3KWZ3L14,3M0I,3M4H,3NKK, 3NTZ,3NU0,3NU3,3NWB,3NXO,3NXX,3NZB,3OND,3OT3,3OVX, 3OZS,3OZT,3PA3,3PKA,3PKB,3PX8,3R6T,3RL4,3S1Y,3S71,3SPK, 3TEM,3U2C,3UHM,3VF3,3VHV,3VW9,3WFG,3ZSJ,3ZXH,4A6V, 4A6W,4BW1,4DHR,4DRI,4DRN,4DRO,4DRQ,4E4A,4F3I,4FH2, 4FLK,4FYO,4GCJ,4GQR,4GV1,4HCT,4HCU,4HCV,4HWW,4HXQ, 4HXS,4HY4,4HYI,4IGH,4IKU,4JHT,4KEB,4L7G,4M5R
$CDK2IC_{50}$	1GII, 1OIR, 2B53, 2B54, 2R3H, 3IGG, 3LE6, 3PXZ, 3PY0, 3RZB, 4RJ3

2.2 Re-docking

In this work, all docking simulations were carried out following the strategy described elsewhere [8]. Briefly, we adopted as docking engine the program Molegro Virtual Docker (MVD) [12]. In doing so, we have the possibility to explore 32 different docking protocols [18]. It is considered a combination of four scoring functions (MolDock score, MolDock score with Grid, PLANTS score, and PLANTS score with Grid) and four search algorithms (Differential Evolution, Simplex Evolution, Iterated Simplex, and Iterated Simplex with Ant Colony Optimization (ACO). In addition, we may carry out docking simulations with or without water molecules. Scoring functions with grid option pre-calculate potential-energy values on an evenly spaced cubic grid in order to speed up docking simulations [12].

Here, our focus was on the crystallographic structures of complexes between enzymes and inhibitors for which IC50 information is available (HRIC50 dataset). Among the structures, we chose the one with the highest crystallographic resolution and applied the 32 docking protocols for this structure. Our goal is to use the most reliable crystallographic structure to test all 32 docking protocols and apply the best protocol to the rest of structures in the HRIC50 dataset. For each docking protocol, we generated a total 1000 poses. Besides the MolDock and PLANTS scores, we also analyzed docking results using the additional scoring functions and the terms of these functions implemented in the program MVD, as described in Table 2. Details about these scoring functions are described else- where [12,18,19]. The most promising docking protocol was selected using as criteria the lowest root-mean square deviation (RMSD).

2.3 Ensemble docking

Here we intend to evaluate the docking accuracy for an ensemble of structures. Differently from the re-docking analysis, here we have the docking RMSD for 173 structures (HRIC50 dataset). We used the best docking protocol, selected for the highest resolution structure, and applied it to all structures in the HRIC50 dataset.

2.4 Scoring Function Analysis

The goal in this step is to test the ability of scoring functions in predicting log(IC50), where IC50 is the half-maximal inhibitory concentration. Experimental information about IC50 was obtained from the Protein Data Bank (PDB) [17]. PDB merges binding affinity data from three other databases PDBbind [20], Binding MOAD (MOAD) [21], and BindingDB [22]. For both datasets, all ligands were prepared using default charge values for the program MVD [12] and protein atomic charges were defined according to default parameters of MVD. For the CDK2IC50 dataset, we also applied AutoDock4 [15] and AutoDock Vina [16] scoring functions to predict ligand-binding affinity. Gasteiger partial charges were assigned to ligands and protein atoms in the CDK2IC50 dataset using Auto- DockTools4 [15] for evaluation of binding affinities using Auto- Dock4 and AutoDock Vina scores.

For both datasets, the MolDock and PLANTS scores implemented in the program MVD [12] were used to predict ligand-binding affinity. Additional scoring functions available in the MVD such as re-rank, ligand efficiency 1 (LE1) and ligand efficiency 3 (LE3) scores were also evaluated. Furthermore, energy terms used to evaluate specific intermolecular interactions, for instance electrostatic energy, were also determined. All scoring functions and energy terms are indicated in Table 2.

Table 2
List of all scoring functions used in this study.

Scoring Functions and Energy Terms	Description
MolDock Score	Protein-ligand Scoring Function. This scoring function is the sum of internal ligand energies, protein interaction energy and soft penalties [12]
PLANTS Score	Protein ligand Scoring Function [12]
Re-rank Score	Protein ligand Scoring Function [12]
Energy Term 1	Interaction energy between the ligand and the target molecule(s) (Interaction) [12]
Energy Term 2	Interaction energy between the ligand and the co-factor (Cofactor) [12]
Energy Term 3	Interaction energy between the ligand and the protein (Protein) [12]
Energy Term 4	Interaction energy between the ligand and the water molecules (Water) [12]
Energy Term 5	Internal energy of the ligand (Internal) [12]
Energy Term 6	Short-range electrostatic protein-ligand interactions (r < 4.5 Å) (Electro) [12]
Energy Term 7	Long-range electrostatic protein-ligand interactions (r > 4.5 A) (ElectroLong) [12]
Energy Term 8	Hydrogen bonding energy (HBond) [12]
LE1 Score	Ligand Efficiency 1: MolDock Score divided by Heavy Atoms count [12]
LE3 Score	Ligand Efficiency 3: Rerank Score divided by Heavy Atoms count [12]
Docking Score	Score evaluated before post-processing (either PLANTS or MolDock). Only used for re-docking [12].
Displaced Water Score .	Energy contributions from non-displaced and displaced water interactions [12]
AutoDock4 Scoring Function	This scoring function makes use of five energetic terms: the torsional term, the hydrogen bonding interactions, the electrostatic potential, the desolvation energy, and the van der Waals interactions [15]
AutoDock Vina Scoring Function	Vina makes use of the following energy terms: Gauss1, Gauss2, repulsion, hydrophobic, hydrogen bond, and torsion. They are defined elsewhere [16]

2.5 Polynomial Scoring functions

In this step, we applied the machine learning box interface of SAnDReS to explore the scoring function virtual space (SFVS) [1,8]. This virtual space is a mathematical construction that links the protein space, here represented by the proteins in the HRIC50 dataset, with a sub-set of ligands of the chemical space. The sub-set of ligands of the chemical space is composed of the inhibitors for the proteins in the ensemble of structures. For a sub-set of protein structures of the protein space, SAnDReS makes use of SML techniques to find the most adequate polynomial scoring function to predict log(IC_{50}). In this work, we used as independent variables of the polynomial equations the terms in the MolDock and PLANTS scores.

2.6 Supervised Machine learning techniques

SAnDReS has seven SML techniques based on scikit-learn library [11]. In the regression analysis the goal is to minimize residual sum of squares (RSS), as defined below,

$$RSS = \sum_{i=1}^{N} \left(y_i - y_{calc,i} \right)^2$$

where the sum is taken for all structures in the dataset, yi is the experimental value of log(IC_{50}) and ycalc,i is the predicted value of log(IC_{50}) as determined by regression methods. The expression for predicted values is given by the following equation,

$$y_{calc} = \beta_0 + \sum_{j=1}^{M} \beta_j x_j$$

where β's represent the relative weight for each explanatory variable and M is the number of explanatory variables in the machine learning model. Among the regression methods available in the program SAnDReS, we have the elastic net, that minimizes the following penalized RSS,

$$RSS = \sum_{i=1}^{N}\left(y_i - (\beta_0 + \sum_{j=1}^{M}\beta_j x_j)\right)^2 + \lambda_1 \sum_{j=1}^{M}\left|\beta_j\right| + \lambda_2 \sum_{j=1}^{M}\left|\beta_j\right|^2$$

where λ1, λ2 > 0. The ordinary linear regression method is obtained with λ1=λ2=0, when λ2=0 and λ1 > 0, we have the least absolute shrinkage and selection operator (Lasso), for λ1=0 and λ2 > 0, we have the Ridge method. In addition, the cross validation can be implemented for Lasso, Ridge, and elastic net. The cross-validated methods use part of the training set to fit the machine learning model and a test set to evaluate the prediction error as defined elsewhere [11]. Considering the ordinary linear regression method and the penalized methods (Lasso, Elastic net and Ridge) without and with cross validation, we have a total of seven regression methods implemented in the program SAnDReS [8]. They were all applied for a training set taken from HRIC50 dataset.

3. RESULTS AND DISCUSSION

3.1 Docking analysis

Using the highest crystallography resolution as a selection criterion, we identified the PDB access code 1US0 [23] as the structure with the highest resolution. This crystallographic structure was employed for re-docking simulations, using the 32 docking protocols previously described elsewhere [8]. The best overall performance was achieved with the protocol 31, which uses as search algorithm the Iterated Simplex with Ant Colony Optimization and PLANTS score (supplementary material 1). Fig. 1 shows the scatter plot for RMSD vs PLANTS score for 1000 poses generated with protocol 31. The Spearman's rank correlation coefficient (r) was 0.673 (p-value<0.001) with RMSD for the lowest score pose of 0.594 Å. The best protocol was used to carry out docking simulation for the rest of the entries in the HRIC50 dataset.

Analysis of the correlation between scoring functions and docking RMSD for all structures in the HRIC50 dataset (supplementary material 2) indicated that the highest correlation coefficient

(r=0.556 and p-value1 < 0.001; R2=0.336 and p- value2 < 0.001) was observed for PLANTS score with a docking accuracy of 63.9%.

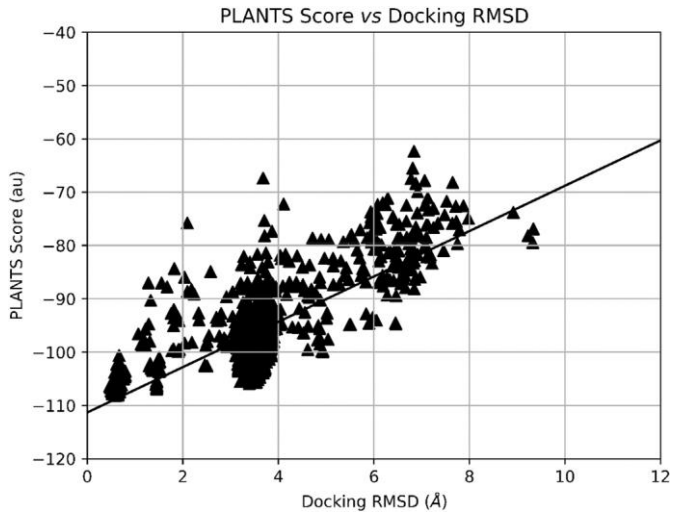


Fig. 1. Scatter plot for re-docking results for 1000 poses generated with the best docking protocol (au means arbitrary units).

3.2 Scoring functions

Results for correlation between MVD scoring functions and log(IC50) for the structures in the HRIC50 dataset indicated low correlation between predicted and experimental binding affinities (supplementary material 3). The highest correlation was observed for energy term 1 (r 0.312 and p-value < 0.001). Nevertheless, p- value < 0.05 was also observed for PLANTS, MolDock, and Re-rank scores. Squared correlation analysis generated poor results, with R2 < 0.1. Since there is no direct significant relationship between the two variables (scoring functions/energy terms and log(IC50)), the Spearman's rank order correlation coefficient is more adequate to investigate the correlation between these variables [24].

We applied SAnDReS to build new machine learning models using as explanatory variables the energy terms available for MolDock and PLANTS scoring functions and as response variable the log(IC50). We have a total of 8 energy terms, considering the combination of 8 energy terms taken 3 at a time, we have a total of 56 combinations. Furthermore, considering that SAnDReS builds 511 polynomial equations for each combination, we ended up with 28,616 machine-learning models.

Application of polynomial scoring function approach generated machine learning models with better predictive power, when compared with original scoring functions/energy terms as shown in Table 3 for polynomial equation number 60 using elastic net CV as regression method. Fig. 2 shows the scatter plot for polynomial equation 60, with test set data. The polynomial equation is shown below,

$$\log(IC_{50}) = 5.763674 + 0.000069(x.y) + 0.000185(x.z) - 0.001090(y.z)\ 0.000040(x^2)$$

where x is Interaction energy between the pose and the protein, y is Internal energy of the ligand (Internal), and z is Hydrogen bonding energy (HBond). These energy terms are described elsewhere [12].

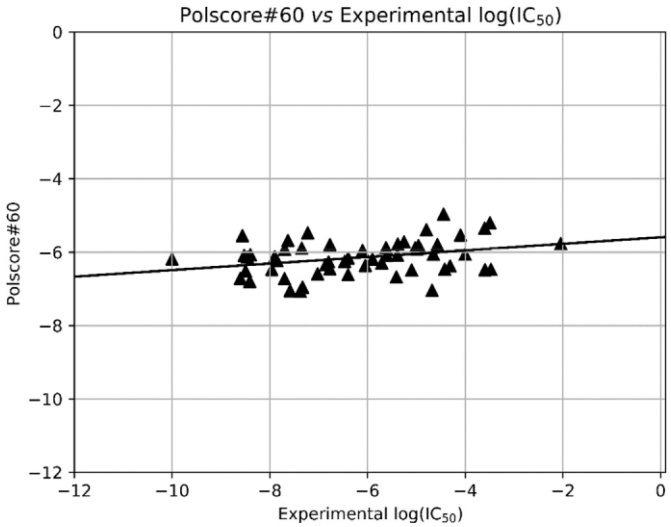


Fig. 2. Scatter plot for experimental and predicted log(IC_{50}) for a test set taken from $HRIC_{50}$ dataset.

Table 3
Results for training and test sets for $HRIC_{50}$ dataset.

Scoring Functions and Energy Terms	r (training set)	p-value (training set)	r (test set)	p-value (test set)
PLANTS Score	0.266	$3.797.10^{-03}$	0.167	$2.185.10^{-01}$
MolDock Score	0.284	$1.939.10^{-03}$	0.224	$9.678.10^{-02}$
Rerank Score	0.227	$1.371.10^{-02}$	0.109	$4.219.10^{-01}$
Term 1	0.334	$2.305.10^{-04}$	0.215	$1.109.10^{-01}$
Term 2	0.130	$1.623.10^{-01}$	0.211	$1.192.10^{-01}$
Term 3	0.340	$1.795.10^{-04}$	0.147	$2.810.10^{-01}$
Term 4	0.214	$2.032.10^{-02}$	0.083	$5.455.10^{-01}$
Term 5	—0.077	$4.104.10^{-01}$	0.155	$2.541.10^{-01}$
Term 6	0.107	$2.514.10^{-01}$	0.179	$1.871.10^{-01}$
Term 7	0.134	$1.511.10^{-01}$	0.101	$4.568.10^{-01}$
Term 8	0.067	$4.746.10^{-01}$	0.237	$7.889.10^{-02}$
Polscore0000060	0.401	$7.243.10^{-06}$	0.328	$1.363.10^{-02}$

3.3 Scoring function virtual space

We could think that SAnDReS works as a scoring function builder, which allows us to explore the SFVS. The SFVS is a mathematical construction where all scoring functions exist, with SAnDReS we can scan this virtual space and find the most adequate model that links the chemical space with the protein space.

We applied the AutoDock4, AutoDock Vina, MolDock, and PLANTS scores to predict binding affinity for CDK2IC50 dataset. We used the crystallographic positions of the inhibitors to calculate binding affinity. Analysis of the predictive power of all classical scoring functions indicated Spearman's rank correlation coefficient ranges from —0.773 to 0.682 (supplementary material 4). Application of polynomial scoring function 60 to CDK2IC50 dataset generated a correlation of 0.845 (p-value < 0.001), which indicates a better predictive power, when compared with classical scoring

functions. Table 4 shows the experimental and predicted binding affinity for all structures in the CDK2IC50 dataset, where we could see the good agreement between predicted and experimental log(IC50).

Table 4
Experimental and predicted log(IC_{50}) for all structures in the CDK2IC_{50} dataset.

PDB Access Code	Active Ligand Code	Resolution (Å)	IC_{50} (nM)	Log (IC_{50})	Predicted log(IC_{50})
1OIR	HDY	1.91	32	—7.495	—7.495
2B53	D23	2.00	600	—6.222	—6.221
2B54	D05	1.85	20	—7.699	—7.699
2R3H	SCE	1.50	20000	—4.699	—3.839
3IGG	EFQ	1.80	80.75	—7.093	—6.196
3LE6	2BZ	2.00	35	—7.456	—6.277
3PXZ	JWS	1.70	5900	—5.229	—5.277
3PY0	SU9	1.75	79.25	—7.101	—6.678
3RZB	02Z	1.90	100000	—4.000	—5.779
4RJ3	3QS	1.63	93	—7.032	—6.544

3.4 Decoys and Actives

We generated a dataset using decoys and active ligands for CDK2 in order to test the ability to identify active ligands present in a dataset to be used for virtual screening. The decoys were randomly taken from DUD-E database [25] using SAnDReS [8]. The actives were the ligands for the structures in Table 1 for the CDK2IC50 dataset. We have a total of 1100 ligands where 11 were actives and the rest of the ligands (1089) were decoys. We applied the protocol 31 from MVD for virtual screening. The performance was evaluated using area under curve (AUC) for receiver operator characteristic (ROC) curve and enrichment factor (EF) as defined elsewhere [8]. The EF for polynomial equation60 was 175.00 with AUC of 79.450%, which are higher than previously published benchmark for CDK2 (AUC=79.10% and EF=13.9) [25]. Fig. 3 shows the ROC curve for polynomial equation 60. Taken together, we could say that the SML model was able to predict binding affinity with better predictive power.

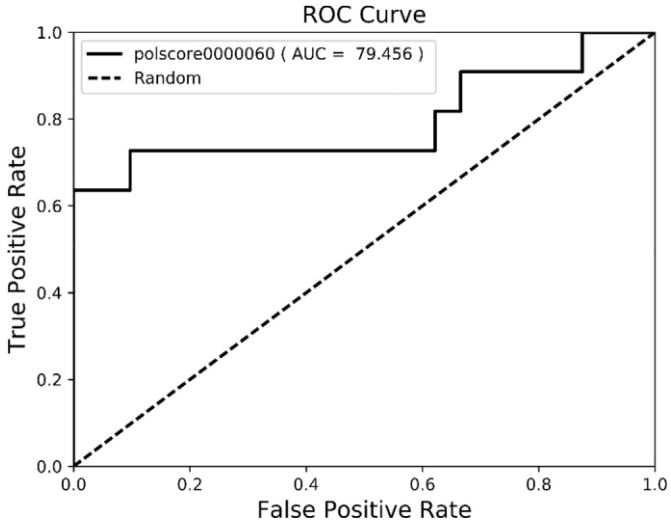


Fig. 3. ROC curve for polynomial equation 60.

3.5 Molecular fork

In the polynomial equation 60, the hybrid term for internal energy of the ligand the intermolecular hydrogen bond terms shows the highest relative weight obtained by regression methods. It has been shown, that the CDK2 inhibition is highly dependent on participation of intermolecular hydrogen bonds involving C-O group on Glu 81, N-H and C-O groups in Leu 83. This pattern has been named molecular fork [26 e 30] and participates in intermolecular interactions for all structures in the CDK2IC50 dataset. It is tempting to speculate that the success of polynomial equation 60 in predicting binding affinity is partially due to the predominance of the intermolecular hydrogen bond term on it, which was able to capture the essence of the intermolecular interaction for CDK2 structures.

Application of machine learning methods to build new models to predict binding affinity has been shown here to generate a polynomial equation with improved prediction power when compared with classical scoring functions, such as PLANTS and MolDock scores. Furthermore, the use of the machine learning model (polynomial equation 60) to calculate log(IC50) for an dataset of CDK2 structures which were not used in the training set, generated correlation between predicted and experimental affinities higher than the ones obtained using classical scoring functions (AutoDock4, AutoDock Vina, MolDock and PLANTS). We could say that our approach has the ability to explore the SFVS, building a computational model with superior predictive power. Such computational methodology has a great potential to be used in the early stages of drug development, since it allows us to explore the SFVS and find the one most adequate to the biological system being investigated.

ACKNOWLEDGMENTS

This work was supported by grants from CNPq (Brazil) (308883/ 2014-4). MBA acknowledges support from CAPES. VOP acknowledges support from PUCRS/BPA fellowship. WFA is senior researcher for CNPq (Brazil) (Process Number: 308883/2014-4).

SUPLEMENTARY DATA

Supplementary data related to this article can be found at https://doi.org/10.1016/j.bbrc.2017.10.035.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

REFERENCES

[1] G.S. Heck, V.O. Pintro, R.R. Pereira, M.B. de A´vila, N.M.B. Levin, W.F. de Azevedo, Supervised machine learning methods applied to predict ligand-binding affinity, Curr. Med. Chem. 24 (2017) 2459e2470.

[2] C. Fan, Y. Huang, Identification of novel potential scaffold for class I HDACs inhibition: an in-silico protocol based on virtual screening, molecular dy- namics, mathematical analysis and machine learning, Biochem. Biophys. Res. Commun. 491 (2017) 800e806.

[3] N.M.B. Levin, V.O. Pintro, M.B. de Avila, B.B. de Mattos, W.F. De Azevedo Jr., Understanding the structural basis for inhibition of cyclin-dependent kinases. New pieces in the molecular puzzle, Curr. Drug Targets 18 (2017) 1104e1111.

[4] A.E. Klon, M. Glick, M. Thoma, P. Acklin, J.W. Davies, Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results, J. Med. Chem. 47 (2004) 2743e2749.
[5] H. Fukunishi, R. Teramoto, J. Shimada, Hidden active information in a random compound library: extraction using a pseudo-structure-activity relationship model, J. Chem. Inf. Model 48 (2008) 575e582.
[6] M.H. Seifert, Robust optimization of scoring functions for a target class, J. Comput. Aided Mol. Des. 23 (2009) 633e644.
[7] F. Klepsch, P. Vasanthanathan, G.F. Ecker, Ligand and structure-based classi- fication models for prediction of P-Glycoprotein inhibitors, J. Chem. Inf. Model 54 (2014) 218e229.
[8] M.N. Xavier, G.S. Heck, M.B. Avila, N.M.B. Levin, V.O. Pintro, N.L. Carvalho, W.F. Azevedo, SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions, Comb. Chem. High. Throughput Screen 19 (2016) 801e812.
[9] M.A. Khamis, W. Gomaa, W.F. Ahmed, Machine learning in computational docking, Artif. Intell. Med. 63 (2015) 135e152.
[10] M. Wo´jcikowski, P.J. Ballester, P. Siedlecki, Performance of machine-learning scoring functions in structure-based virtual screening, Sci. Rep. 7 (2017) 46710.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Verplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825e2830.
[12] R. Thomsen, M.H. Christensen, MolDock: a new technique for high-accuracy molecular docking, J. Med. Chem. 49 (2006) 3315e3321.
[13] H.L. De Bondt, J. Rosenblatt, J. Jancarik, H.D. Jones, D.O. Morgan, S.H. Kim, Crystal structure of cyclin-dependent kinase 2, Nature 363 (1993) 595e602.
[14] W.F. de Azevedo, Opinion paper: targeting multiple cyclin-dependent kinases (CDKs): a new strategy for molecular docking studies, Curr. Drug Targets 17 (2016) 2.
[15] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility, J. Comput. Chem. 30 (2009) 2785e2791.
[16] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multi- threading, J. Comput. Chem. 31 (2010) 455e461.
[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235e242.
[18] W.F. De Azevedo Jr., MolDock applied to structure-based virtual screening, Curr. Drug Targets 11 (2010) 327e334.
[19] L.S. Azevedo, F.P. Moraes, M.M. Xavier, E.O. Pantoja, B. Villavicencio, J.A. Finck, A.M. Proenca, K.B. Rocha, W.F. de Azevedo, Recent progress of molecular docking simulations applied to development of drugs, Curr. Bioinform 7 (2012) 352e365.
[20] R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures, J. Med. Chem. 47 (2004) 2977e2980.
[21] L. Hu, M.L. Benson, R.D. Smith, M.G. Lerner, H.A. Carlson, Binding MOAD (mother of all databases), Proteins 60 (2005) 333e340.
[22] X. Chen, M. Liu, M.K. Gilson, BindingDB: a web-accessible molecular recog- nition database, Comb. Chem. High. Throughput Screen 4 (2001) 719e725.
[23] E.I. Howard, R. Sanishvili, R.E. Cachau, A. Mitschler, B. Chevrier, P. Barth, V. Lamour, M. Van Zandt, E. Sibley, C. Bon, D. Moras, T.R. Schneider, A. Joachimiak, A. Podjarny, Ultrahigh resolution drug design I: details of in- teractions in human aldose reductase-inhibitor complex at 0.66 A, Proteins 55 (2004) 792e804.

[24] M. Otyepka, V. Krystof, L. Havlícek, V. Siglerova´, M. Strnad, J. Koca, Docking-based development of purine-like inhibitors of cyclin-dependent kinase-2, J. Med. Chem. 43 (2000) 2506e2513.
[25] M.M. Mysinger, M. Carchia, J.J. Irwin, B.K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, J. Med. Chem. 55 (2012) 6582e6594.
[26] W.F. de Azevedo Jr., F. Canduri, N.J. da Silveira, Structural basis for inhibition of cyclin-dependent kinase 9 by flavopiridol, Biochem. Biophys. Res. Commun. 293 (2002) 566e571.
[27] W. Filgueira de Azevedo Jr., R.T. Gaspar, F. Canduri, J.C. Camera Jr., N.J. Freitas da Silveira, Molecular model of cyclin-dependent kinase 5 complexed with roscovitine, Biochem. Biophys. Res. Commun. 297 (2002) 1154e1158.
[28] F. Canduri, H.B. Uchoa, W.F. de Azevedo Jr., Molecular models of cyclin- dependent kinase 1 complexed with inhibitors, Biochem. Biophys. Res. Commun. 324 (2004) 661e666.
[29] P.C. Perez, R.A. Caceres, F. Canduri, W.F. de Azevedo Jr., Molecular modeling and dynamics simulation of human cyclin-dependent kinase 3 complexed with inhibitors, Comput. Biol. Med. 39 (2009) 130e140.
[30] E. Schonbrunn, S. Betzi, R. Alam, M.P. Martin, A. Becker, H. Han, R. Francis, R. Chakrasali, S. Jakkaraj, A. Kazi, S.M. Sebti, C.L. Cubitt, A.W. Gebhard, L.A. Hazlehurst, J.S. Tash, G.I. Georg, Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases, J. Med. Chem. 56 (2013) (2013) 3768e3782.

ANEXO C – Artigo publicado na revista Chemical Biology and Drug Design. 2018. Fator de Impacto: 2,548.

RESEARCH ARTICLE

WILEY

Development of machine learning models to predict inhibition of 3-dehydroquinate dehydratase

Maurício Boff de Ávila[1,2] | **Walter Filgueira de Azevedo Jr**[1,2]

[1]Laboratory of Computational Systems Biology, School of Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil

[2]Graduate Program in Cellular and Molecular Biology, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil

Correspondence
Walter Filgueira de Azevedo, Laboratory of Computational Systems Biology, School of Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil.
Email: walter@azevedolab.net

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 308883/2014-4

In this study, we describe the development of new machine learning models to predict inhibition of the enzyme 3-dehydroquinate dehydratase (DHQD). This enzyme is the third step of the shikimate pathway and is responsible for the synthesis of chorismate, which is a natural precursor of aromatic amino acids. The enzymes of shikimate pathway are absent in humans, which make them protein targets for the design of antimicrobial drugs. We focus our study on the crystallographic structures of DHQD in complex with competitive inhibitors, for which experimental inhibition constant data is available. Application of supervised machine learning techniques was able to elaborate a robust DHQD-targeted model to predict binding affinity. Combination of high-resolution crystallographic structures and binding information indicates that the prevalence of intermolecular electrostatic interactions between DHQD and competitive inhibitors is of pivotal importance for the binding affinity against this enzyme. The present findings can be used to speed up virtual screening studies focused on the DHQD structure.

KEYWORDS
3-dehydroquinate dehydratase, crystallographic structures, drug design, machine learning, protein–ligand interactions, systems biology

1 | INTRODUCTION

Shikimate pathway (SP) is an essential metabolic route found in bacteria, higher plants and apicomplexan parasites.[1] In bacteria, seven enzymes participate in this pathway which catalyzes the formation of chorismate, a natural precursor of aromatic amino acids (Phe, Tyr, Trp), folic acid, and ubiquinone from phosphoenolpyruvate and D-erythrose-4-phosphate.[2,3] The vital role played by the SP in bacteria and the absence of this pathway in humans makes it a target for the development of new antimicrobial drugs.

Our study is about the third enzyme of the route, 3-dehydroquinate dehydratase (DHQD) (Enzyme Classification (EC) 4.2.1.10), which catalyzes the reversible reaction of elimination of water (dehydration) from 3-dehydroquinate to form 3-dehydroshikimate.[4] DHQD exists in two forms (type I (DHQDI) and type II (DHQDII)) that differ in structure but catalyze the same reaction. DHQDI has a structure composed of eight-stranded α/β barrel. On the other hand, DHQDII shows a five-stranded parallel β-sheet flanked by four α-helices.[5] DHQDI catalyzes dehydration by *syn* elimination with the loss of hydrogen from C2. This reaction involves the utilization of an imine intermediate.[6] DHQDII is present in two routes, shikimate (bacteria) and quinate (fungi), and catalyzes dehydration reaction through anti-elimination taunting loss of hydrogen from C2.

Our primary goal here is to propose a computational model to predict inhibition of DHQD. Instead of using docked structures (poses)[7–18] obtained from docking simulations, we based our modeling on the crystallographic position of the ligands[19] derived from an ensemble of crystallographic structures available for DHQD. We tested the predictive power of our proposed scoring function against a dataset composed of decoys and active ligands. Our results indicated that DHQD-targeted scoring function shows higher predictive power when compared to MOLDOCK, PLANTS,[20–22]

ANEXO D – Artigo publicado na revista Current Medicinal Chemistry. 2017. Fator de Impacto: 4,184.

REVIEW ARTICLE

Supervised Machine Learning Methods Applied to Predict Ligand-Binding Affinity

Gabriela S. Heck[a], Val O. Pintro[a], Richard R. Pereira[a], Mauricio B. de Ávila[a,b], Nayara M.B. Levin[a,b] and Walter F. de Azevedo Jr.[a,b,*]

[a]*Laboratory of Computational Systems Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil;* [b]*Graduate Program in Cellular and Molecular Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil*

Abstract: ***Background:*** Calculation of ligand-binding affinity is an open problem in computational medicinal chemistry. The ability to computationally predict affinities has a beneficial impact in the early stages of drug development, since it allows a mathematical model to assess protein-ligand interactions. Due to the availability of structural and binding information, machine learning methods have been applied to generate scoring functions with good predictive power.

Objective: Our goal here is to review recent developments in the application of machine learning methods to predict ligand- binding affinity.

Method: We focus our review on the application of computational methods to predict binding affinity for protein targets. In addition, we also describe the major available databases for experimental binding constants and protein structures. Furthermore, we explain the most successful methods to evaluate the predictive power of scoring functions.

Results: Association of structural information with ligand-binding affinity makes it possible to generate scoring functions targeted to a specific biological system. Through regression analysis, this data can be used as a base to generate mathematical models to predict ligand-binding affinities, such as inhibition constant, dissociation constant and binding energy.

Conclusion: Experimental biophysical techniques were able to determine the structures of over 120,000 macromolecules. Considering also the evolution of binding affinity information, we may say that we have a promising scenario for development of scoring functions, making use of machine learning techniques. Recent developments in this area indicate that building scoring functions targeted to the biological systems of interest shows superior predictive performance, when compared with other approaches.

ARTICLE HISTORY

Received: December 20, 2017
Revised: May 23, 2017
Accepted: June 06, 2017

DOI:
10.2174/0929867324666170623092503

Keywords: Machine Learning, medicinal chemistry, binding affinity, regression, drug, enzyme.

1. INTRODUCTION

The application of machine learning (ML) technique is not new to the studies of computational medicinal chemistry and systems biology. A recent literature search in PubMed conducted on May 22nd 2017 using the keywords "machine learning" and "biology" returned 2266 scientific publications, as shown in Fig. (1). The oldest report dates back to 1985 [1]. In this list of publications, the first report to use the term "Machine learning" in the paper title came out in 1998 [2]. There are examples of application of such methods to a wide variety of biological problems. For instance, the use of artificial neural networks to model complex biological data [3], the application of a weighted variant of the K-nearest neighbor (KNN) to analyze the protein

*Address correspondence to this author at the Laboratory of Computational Systems Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil;
E-mail: walter.junior@pucrs.br

ANEXO E – Artigo publicado na revista Current Drug Target. 2017.
Fator de Impacto: 2,532.

REVIEW ARTICLE

Understanding the Structural Basis for Inhibition of Cyclin-Dependent Kinases. New Pieces in the Molecular Puzzle

Nayara M. Bernhardt Levin[a,b], Val Oliveira Pintro[a], Maurício Boff de Ávila[a,b], Bruna Boldrini de Mattos[a] and Walter Filgueira de Azevedo Jr.[a,b,*]

[a]*Laboratory of Computational Systems Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil;* [b]*Graduate Program in Cellular and Molecular Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681. Porto Alegre-RS 90619-900, Brazil*

ARTICLE HISTORY

Received: August 26, 2016
Revised: November 07, 2016
Accepted: November 08, 2016

DOI:
10.2174/1389450118666161116130155

Abstract: ***Background:*** Cyclin-dependent kinases (CDKs) comprise an important protein family for development of drugs, mostly aimed for use in treatment of cancer but there is also potential for development of drugs for neurodegenerative diseases and diabetes. Since the early 1990s, structural studies have been carried out on CDKs, in order to determine the structural basis for inhibition of this protein target.

Objective: Our goal here is to review recent structural studies focused on CDKs. We concentrate on latest developments in the understanding of the structural basis for inhibition of CDKs, relating structures and ligand-binding information.

Method: Protein crystallography has been successfully applied to elucidate over 400 CDK structures. Most of these structures are complexed with inhibitors. We use this richness of structural information to describe the major structural features determining the inhibition of this enzyme.

Results: Structures of CDK1, 2, 4-9, 12 13, and 16 have been elucidated. Analysis of these structures in complex with a wide range of different competitive inhibitors, strongly indicate some common features that can be used to guide the development of CDK inhibitors, such as a pattern of hydrogen bonding and the presence of halogen atoms in the ligand structure.

Conclusion: Nowadays we have structural information for hundreds of CDKs. Combining the structural and functional information we may say that a pattern of intermolecular hydrogen bonds is of pivotal importance for inhibitor specificity. In addition, machine learning techniques have shown improvements in predicting binding affinity for CDKs.

Keywords: Cyclin-dependent kinase, binding affinity, drug design, machine learning.

1. INTRODUCTION

Several protein kinases (enzyme classification (EC) 2.7.-.-) have been described as participating in intracellular regulatory pathways (reviews in [1-10]). Our focus in this paper is on cyclin-dependent kinases (CDKs) (EC 2.7.11.22). We could say that, from the enzymology point of view, CDKs are enzymes that catalyze the phosphor transfer to a protein, as indicated in the catalyzed reaction below,

ATP + a protein ⇔ ADP + a phosphoprotein.

CDKs are also known as serine-threonine kinases, due to the specificity of their substrates.

*Address correspondence to this author at the Laboratory of Computational Systems Biology, Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil; Tel/Fax: ????????????????; E-mail: walter.junior@pucrs.br

Functional studies were able to determine the significance of CDKs in the cell cycle progression [11, 12]. As functional studies progressed, new CDKs were identified and their biological roles established [13, 14]. Because of the importance of CDKs in the regulation of the cell division cycle, these enzymes have been the object of extensive investigation. Furthermore, due to their central role in cell cycle progression, CDKs have caught attention as a target for development of anticancer drugs. CDK inhibitors have also shown potential for the treatment of inflammatory disorders [15] and neurodegenerative diseases [16]. Screening studies aimed to identify CDK inhibitors were initially focused on starfish due to the abundance of CDK1/cyclin D in starfish oocytes [17], which allowed the identification of Roscovitine, a CDK inhibitor with half maximal inhibitory concentration (IC_{50}) in nanomolar range [18, 19].

In parallel with CDK functional studies, several research groups have tried to obtain x-ray diffracting crystals to de-

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br