

UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA DE SÃO CARLOS

“SOBRE OS ESTUDOS METABÓLICOS DE FÁRMACOS EMPREGANDO-SE ATIVIDADE ENZIMÁTICA DE CYP450 VISANDO-SE ESTABELEECER CORRELAÇÕES ENTRE ESTRUTURA E ATIVIDADE”

Renato de Lima Bauab

Dissertação apresentada ao Instituto de Química de São Carlos, Universidade de São Paulo, para obtenção do Título de Mestre em Físico-Química, Curso de Pós-Graduação em Química, Área de Concentração Físico-Química.

São Carlos, São Paulo, junho, 2011

Este exemplar foi revisado e alterado em relação à
Versão original, sob a exclusiva responsabilidade
do autor.

São Carlos, 07/12/2011

Renato de Lima Bauab

UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA DE SÃO CARLOS

“SOBRE OS ESTUDOS METABÓLICOS DE FÁRMACOS EMPREGANDO-SE ATIVIDADE ENZIMÁTICA DE CYP450 VISANDO-SE ESTABELEECER CORRELAÇÕES ENTRE ESTRUTURA E ATIVIDADE”

Renato de Lima Bauab

Carlos Alberto Montanari
Orientador

Dissertação apresentada ao Instituto de Química de São Carlos, Universidade de São Paulo, para obtenção do Título de Mestre em Físico-Química, Curso de Pós-Graduação em Química, Área de Concentração Físico-Química.

São Carlos, São Paulo, junho, 2011

Renato de Lima Bauab

SOBRE OS ESTUDOS METABÓLICOS DE FÁRMACOS EMPREGANDO-SE ATIVIDADE ENZIMÁTICA DE CYP450 VISANDO-SE ESTABELECEER CORRELAÇÕES ENTRE ESTRUTURA E ATIVIDADE./ Renato Lima Bauab, - São Carlos: IQSC, 2011

110 p.

Dissertação (Mestrado) Instituto de Química de São Carlos, 2011.

Orientador: Prof. Dr. Carlos Alberto Montanari

1. Metabolismo de Fármacos. I. Título

Dedicatória

Ao Meu pai Luiz Henrique Bauab (in memorian),
À minha mãe Márcia M. de Lima Bauab, pela
oportunidade de estar aqui.

Aos que tiveram sua presença desde a pia batismal,
Luiz Celso H. Teles e Ponciana Z. Teles.

Ao meu paizão Marcos Roberto Semenzim por sua
participação em minha vida.

Agradecimentos

A Deus, pela criação e pelas muitas oportunidades.

À Minha família, pela criação, pelo incentivo e por todo o apoio.

À minha irmã, Juliana L. Bauab Gonçalves, pela amizade, pelo amor e pelo apoio que sempre me deu.

Ao Prof. Dr. Wagner Luiz Polito, por toda a base e o suporte que sempre me deu, mesmo antes da minha carreira de químico e pela amizade.

Ao Prof. Dr. Carlos Alberto Montanari, por ter acreditado em mim desde meu início no Grupo de Química Medicinal de Produtos Naturais (NEQUIMED-PN), por todo o suporte no desenvolvimento do trabalho, pelas práticas esportivas e pela amizade.

À Dr^a. Maria Luiza C. Montanari, por todo o carinho que sempre manifestou para com os membros do NEQUIMED-PN, pela amizade e pelos “puxões de orelha”.

Aos membros da República, Eduardo Baucia, Bruno Mazzotti, Heitor Polidoro, Felipe Suzuki, Rodrigo Junqueira, Guto Dacol e a todos os agregados, pela amizade, companhia e convívio.

Aos amigos do NEQUIMED, Igor, Josmar, Helton, Juliana, Renato, Vinícius, Geraldo, pelo convívio, pelas discussões, pela ajuda e por todos os momentos que passamos juntos.

Aos meus muitos amigos de causa, de São Carlos em especial, que sempre me acompanharam ao longo de minha estada estiveram ao meu lado em muitos momentos.

Ao meu grande amigo Dr. Renato Ferreira de Freitas, por todo o convívio, o auxílio e o empenho, que viabilizaram a conclusão deste trabalho.

Às funcionárias da Pós-Graduação, Sílvia e Andréia que sempre me ajudaram na resolução das questões burocráticas.

Ao Prof. Dr. Oswaldo “Barba” Baptista Duarte Filho, pela colação de grau fora da data que possibilitou minha matrícula no programa de Pós-Graduação

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pelo apoio financeiro.

Ao meu mestre Dr. Celso Charuri, por todo o conhecimento que me permitiu chegar onde cheguei e ser quem sou.

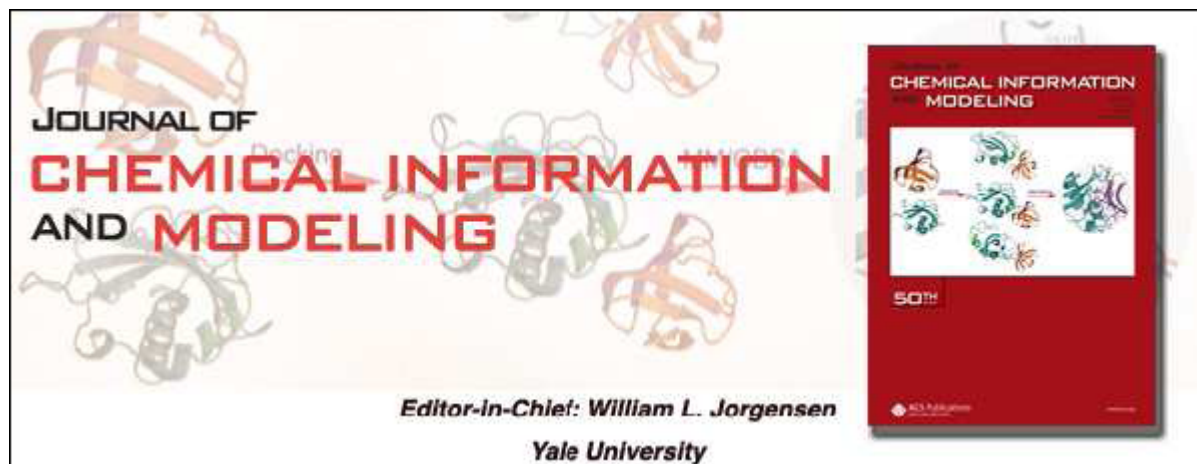
“... a criação do visível é totalmente dependente daquelas essências que não são visíveis a olho nu. Entretanto, as idéias centrais, que ocupam as mentes da maioria das pessoas – humanas – não têm peso suficiente para merecerem comentários ou serem investidas com uma realidade permanente.”

Conde de Saint Germain

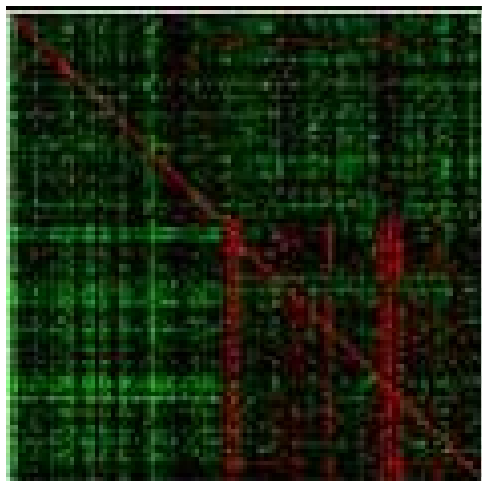
Esta Dissertação foi desenvolvida com o auxílio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da FAPESP pelas contribuições ao Grupo de Pesquisa.

Parte desta Dissertação foi publicada .

Informações a seguir e Transcrição no Anexo I dessa Dissertação.



One of the ten Most-accessed articles from a leading journal in chemical informatics and molecular modeling:



Novel Application of 2D and 3D-Similarity Searches To Identify Substrates among Cytochrome P450 2C9, 2D6, and 3A4

R. F. Freitas, R. L. Bauab and C. A. Montanari

DOI: [10.1021/ci900074t](https://doi.org/10.1021/ci900074t)

SUMÁRIO

ÍNDICE DE FIGURAS	I
ÍNDICE DE TABELAS.....	IV
LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS	V
RESUMO	vi
ABSTRACT	vii
1. INTRODUÇÃO	1
1.1. MODELOS <i>IN SILICO</i>	2
1.2. FARMACOCINÉTICA.....	4
1.3. METABOLISMO DE FÁRMACOS	5
1.4. ENZIMAS DO CITOCROMO P450	6
1.4.1. CITOCROMO P450 1A2.....	7
1.4.2. CITOCROMO P450 2C9.....	7
1.4.3. CITOCROMO P450 2C19.....	7
1.4.4. CITOCROMO P450 2D6.....	8
1.4.5. CITOCROMO P450 3A4.....	8
1.5. RECEPTORES DO 5HT ₃	10
2. OBJETIVOS	11
2.1. OBJETIVOS GERAIS.....	11
2.2. OBJETIVOS ESPECÍFICOS	11
3. FERRAMENTAS E MÉTODOS.....	12
3.1. METASITE	12
3.2. MOLPRINT2D	12
3.1.3 UNITY 2D FINGERPRINTS ³⁶	13
3.1.4 ROCS ³⁹	13
3.2 QUANTIFICAÇÃO DE DESEMPENHO.....	13
4. RESULTADOS.....	15
4.1. SETRONS	15
4.1.1. ALOSETRON	15
4.1.2. DOLASETRON	18
4.1.3. GRANISETRON.....	21
4.1.4. ONDANSETRON.....	23
4.1.5. PALONOSETRON	26
4.1.6. TROPISETRON	28
4.2. BASE DE DADOS.....	30
4.2.1. CYP2C9 x CYP2D6	33

4.2.1.1. DIVERSIDADE DE COMPOSTOS	34
4.2.1.2. ANÁLISE ROC-AUC.....	37
4.2.2. CYP2C9 x CYP3A4.....	43
4.2.2.1. DIVERSIDADE DOS COMPOSTOS	44
4.2.2.2. ANÁLISES AUC-ROC	46
4.2.3. CYP2D6-CYP3A4	51
4.2.3.1. DIVERSIDADE DOS COMPOSTOS	52
4.2.3.2. ANÁLISE ROC-AUC.....	54
4.3. CORRELAÇÃO ENTRE A ESTRUTURA DOS SUBSTRATOS E A SELETIVIDADE DAS ISOFORMAS	59
4.4. SOBREPOSIÇÕES DO ROCS.....	64
5. CONCLUSÃO.....	68
6. REFERÊNCIAS BIBLIOGRÁFICAS.....	70
ANEXO I	77

ÍNDICE DE FIGURAS

FIGURA 1: METABÓLITO EXPERIMENTAL DO ALOSETRON POR OXIDAÇÃO VIA CYP.....	16
FIGURA 2: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP1A2 PARA O ALOSETRON.....	16
FIGURA 3: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2C9 PARA O ALOSETRON.....	17
FIGURA 4: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP3A4 PARA O ALOSETRON.....	17
FIGURA 5: METABÓLITOS EXPERIMENTAIS DO DOLASETRON E DO HIDRODOLASETRON POR OXIDAÇÃO VIA CYP.	18
FIGURA 6: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2D6 PARA O R(+)-HYDRODOLASETRON.....	19
FIGURA 7: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP3A4 PARA O R(+)-HYDRODOLASETRON.....	19
FIGURA 8: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2D6 PARA O S(-)-HYDRODOLASETRON.....	20
FIGURA 9: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP3A4 PARA O S(-)-HYDRODOLASETRON.....	20
FIGURA 10: METABÓLITOS EXPERIMENTAIS DO GRANISETRON POR OXIDAÇÃO VIA CYP.....	21
FIGURA 11: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP1A2 PARA O GRANISETRON.....	22
FIGURA 12: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP3A4 PARA O GRANISETRON.....	22
FIGURA 13: METABÓLITOS EXPERIMENTAIS DO ONDANSETRON POR OXIDAÇÃO VIA CYP.....	23
FIGURA 14: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP1A2 PARA O R-ONDANSETRON.....	24
FIGURA 15: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2D6 PARA O R-ONDANSETRON.....	24
FIGURA 16: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP1A2 PARA O S-ONDANSETRON.....	25
FIGURA 17: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2D6 PARA O S-ONDANSETRON.....	25
FIGURA 18: METABÓLITOS EXPERIMENTAIS DO PALONOSETRON POR OXIDAÇÃO VIA CYP.....	26
FIGURA 19: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP1A2 PARA O PALONOSETRON.....	27
FIGURA 20: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP2D6 PARA O PALONOSETRON.....	27
FIGURA 21: PROBABILIDADE P_{SM}^I (BARRAS ESCURAS) E ACESSIBILIDADE E_I (BARRAS CLARAS) NA CYP3A4 PARA O PALONOSETRON.....	28

FIGURA 22: METABÓLITOS EXPERIMENTAIS DO TROPISETRON POR OXIDAÇÃO VIA CYP.	29
FIGURA 23: PROBABILIDADE P_{SM}^j (BARRAS ESCURAS) E ACESSIBILIDADE E_i (BARRAS CLARAS) NA CYP2D6 PARA O TROPISETRON.	29
FIGURA 24: PROBABILIDADE P_{SM}^j (BARRAS ESCURAS) E ACESSIBILIDADE E_i (BARRAS CLARAS) NA CYP3A4 PARA O TROPISETRON.....	30
FIGURA 25: DIAGRAMAS DE REPRESENTAÇÃO DOS CONJUNTOS DE SELETIVIDADE DE COMPOSTOS. CADA SETA RESPRESENTA UM CONJUNTO DE SELETIVIDADE. O NÚMERO DE COMPOSTOS DE CADA CONJUNTO TAMBÉM ESTÁ REPORTADO E A DEIREÇÃO DAS SETAS DEFINEM A SELETIVIDADE.....	31
FIGURA 26: VALOR DA MÉDIA DE TANIMOTO PARA A SISTEMÁTICA DE SIMILARIDADE POR COMPARAÇÃO PAREADA DA BASE DE DADOS DAS CYP2C9-CYP2D6. NA FIGURA ESTÃO AS ESTRUTURAS 2D DOS DOIS COMPOSTOS ATÍPICOS À BASE DE DADOS.....	33
FIGURA 27: MATRIZES DE TANIMOTO PARA AS COMPARAÇÕES DUPLAS SELETIVAS PARA A BASE DE DADOS CYP2C9-CYP2D6.....	36
FIGURA 28: (A) CURVAS ROC COMPARANDO O DESEMPENHO DOS MÉTODOS DE BUSCA POR SIMILARIDADES PARA DISCRIMINAR ENTRE OS SUBSTRATOS DA CYP2C9 E OS DA CYP2D6; (B) ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP2C9 QUE PROVERAM OS MELHORES VALORES DE AUC PARA CADA MÉTODO.....	38
FIGURA 29: (A) CURVAS ROC COMPARANDO O DESEMPENHO DOS MÉTODOS DE BUSCA POR SIMILARIDADE PARA DISCRIMINAR SUBSTRATOS ENTRE A CYP2C9 E A CYP2D6; (B) ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP2D6 QUE FORNECERAM O MELHOR VALOR DE AUC PARA CADA MÉTODO.....	41
FIGURA 30: MÉDIA DE TANIMOTO PARA A SISTEMÁTICA DE COMPARAÇÃO POR SIMILARIDADE DUPLA DA BASE DE DADOS CYP2C9-CYP3A4. NA FIGURA ESTÃO AS ESTRUTURAS 2D DOS DOIS COMPOSTOS QUE SÃO ATÍPICOS À BASE DE DADOS.	43
FIGURA 31: MATRIZES DE TANIMOTO DA COMPARAÇÃO PAREADA SELETIVA DA BASE DE DADOS CYP2C9-CYP3A4.	45
FIGURA 32: (A) CURVAS ROC COMPARANDO O DESEMPENHO DOS MÉTODOS DE BUSCA POR SIMILARIDADE PARA DISCRIMINAR OS SUBSTRATOS DA CYP2C9 DOS DA CYP3A4; (B) ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP2C9 QUE FORNECERAM O MELHOR VALOR DE AUC PARA CADA MÉTODO.....	48
FIGURA 33: (A) CURAS ROC COMPARANDO O DESEMPENHO DOS MÉDOTOS DE BUSCA POR SIMILARIDADE PARA DISCRIMINAR SUBSTRATOS DA CYP3A4 DOS DA CYP2C9; (B) ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP3A4 QUE FORNECERAM O MELHOR VALOR DE AUC PARA CADA MÉTODO.....	49
FIGURA 34: VALOR DA MÉDIA DE TANIMOTO PARA A COMPARAÇÃO SISTEMÁTICA DE SIMILARIDADE PAREADA PARA A BASE DE DADOS CYP2D6-CYP3A4.....	52
FIGURA 35: MATRIZES DE TANIMOTO PARA A COMPARAÇÃO SELETIVA POR PAREAMENTO DA BASE DE DADOS CYP2D6-CYP3A4.....	54
FIGURA 36: (A) CURVAS ROC COMPARANDO O DESEMPENHO DOS MÉTODOS DE BUSCA POR SIMILARIDADE PARA DISCRIMINAR SUBSTRATOS DA ENZIMA CYP2D6 DOS DA CYP3A4; (B)	

ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP2D6 QUE PROVERAM O MELHOR VALOR DE AUC PARA CADA MÉTODO.....	56
FIGURA 37: (A) CURVAS ROC COMPARANDO O DESEMPENHO DOS MÉTODOS DE BUSCA POR SIMILARIDADE PARA DISCRIMINAR SUBSTRATOS DA ENZIMA CYP3A4 DOS DA CYP2D6; (B) ESTRUTURAS QUÍMICAS DOS SUBSTRATOS DA CYP3A4 QUE PROVERAM O MELHOR VALOR DE AUC PARA CADA MÉTODO.....	58
FIGURA 38: SOBREPOSIÇÃO DOS 10 PRIMEIROS SUBSTRATOS DA CYP2C9 COM O SUPROFENO, CONFORME O MÉTODO DO ROCS-COMBO. OS SÍTIOS DE METABOLISMO ESTÃO REPRESENTADOS EM LARANJA. OS ÁTOMOS DE CARBONO DO SUPROFENO ESTÃO COLORIDOS EM VERDE.....	65
FIGURA 39: SOBREPOSIÇÃO DOS 10 PRIMEIROS SUBSTRATOS DA CYP2D6 COM A FENFORMINA, CONFORME O MÉTODO DO ROCS-COMBO. OS SÍTIOS DE METABOLISMO ESTÃO REPRESENTADOS EM LARANJA. OS ÁTOMOS DE CARBONO DA FENFORMINA ESTÃO COLORIDOS EM VERDE.	66
FIGURA 40: SOBREPOSIÇÃO DOS 10 PRIMEIROS SUBSTRATOS DA CYP3A4 COM A CICLOBENZAPRINA, CONFORME O MÉTODO DO ROCS-COMBO. OS SÍTIOS DE METABOLISMO ESTÃO REPRESENTADOS EM LARANJA. OS ÁTOMOS DE CARBONO DA CICLOBENZAPRINA ESTÃO COLORIDOS EM VERDE.....	67

ÍNDICE DE TABELAS

TABELA 1: DISTRIBUIÇÃO DOS POLIMORFISMOS DA CYP2D6 DENTRO DAS ETNIAS.....	8
TABELA 2: MÉDIA E DESVIO PADRÃO DE ALGUMAS PROPRIEDADES 1D DA BASE DE DADOS. ...	32
TABELA 3: PARÂMETROS ESTATÍSTICOS DA ANÁLISE ROC-AUC PARA A BASE DE DADOS CYP2C9-2D6.	38
TABELA 4: PARÂMETROS ESTATÍSTICOS PARA A ANÁLISE ROC-AUC PARA A BASE DE DADOS CYP2C9-CYP3A4.	47
TABELA 5: PARÂMETROS ESTATÍSTICOS DA ANÁLISE ROC-AUC PARA A BASE DE DADOS CYP2D6-CYP3A4.	55
TABELA 6: PONTOS FARMACOFÓRICOS DA ESTRUTURA DE REFERÊNCIA QUE PROVERAM A MELHOR SEPARAÇÃO ENTRE SUBSTRATOS E NÃO-SUBSTRATOS PARA AS TRÊS CYP450 ESTUDADAS AQUI. ALGUNS EXEMPLOS TAMBÉM SÃO MOSTRADOS ABAIXO.....	61

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

ADME – Absorção, distribuição, metabolismo e excreção

CYP1A2 – Enzima da família do citocromo P450 da subclasse 1A2

CYP2C9 – Enzima da família do citocromo P450 da subclasse 2C9

CYP2C19 – Enzima da família do citocromo P450 da subclasse 2C19

CYP2D6 – Enzima da família do citocromo P450 da subclasse 2D6

CYP3A4 – Enzima da família do citocromo P450 da subclasse 3A4

ADME/Tox – Absorção, distribuição, metabolismo, excreção e toxidez

NCE – Nova entidade química

SAR – Relações entre estrutura química e atividade farmacológica

QSAR – Relações quantitativas entre estrutura química e atividade farmacológica

SPR – Relações entre estrutura química e propriedade farmacológica

QSPR – Relações quantitativas entre estrutura química e propriedade farmacológica

LBDD – Método de planejamento baseado na estrutura do ligante

SBDD – Método de planejamento baseado na estrutura do receptor

FDA – Administração Federal de Alimentos e Medicamentos dos Estados Unidos

CYP – Enzima genérica da família do citocromo P450

AUC – Área sob a curva

ROC – Curva de característica receptora operante

Se – Sensibilidade pelo método ROC

Sp – Especificidade pelo método ROC

P_{sm}^i – Probabilidade de metabolismo da molécula pela enzima calculado pelo MetaSite

E_j – Acesoibilidade da molécula ao sítio catalítico da enzima calculado pelo MetaSite

HBA – Aceitador de ligações de hidrogênio

HBD – Doador de ligações de hidrogênio

MW – Peso molar

RB – Ligações rotacionais

Log P – Logarítimo do coeficiente de partição em octanol/água

PSA – Área da superfície polar

Tc – Coeficiente de Tanimoto

NSAI – Fármacos anti-inflamatórios não esteroideal

RESUMO

“SOBRE OS ESTUDOS METABÓLICOS DE FÁRMACOS EMPREGANDO-SE ATIVIDADE ENZIMÁTICA DE CYP450 VISANDO-SE ESTABELECEER CORRELAÇÕES ENTRE ESTRUTURA E ATIVIDADE”

Carlos Alberto Montanari
Orientador

Renato de Lima Bauab

Palavras-Chave: *Química Medicinal, Fármacos, Farmacoterapia, Quiminformática, Farmacocinética, Metabolismo, Enzimas, Cítocromos*

A Química Medicinal é ciência multidisciplinar com ação direta sobre conhecimentos específicos focalizando Química, Biologia, Medicina, Fisiologia, entre outras áreas de estudos no domínio fundamental e tecnológico. Esta ciência atua ainda entre várias interfaces científicas tais como a Bioquímica, Biofísica, Biologia Molecular, Química Biológica e outras.

A investigação no metabolismo de fármacos é a primeira e essencial fase na moderna farmacologia, uma vez que os parâmetros Farmacocinéticos são os mais relevantes dados iniciais a serem considerados no início da Fase 3 em testes clínicos com humanos. Em sua totalidade os dados farmacocinéticos são conhecidos como ADME (**A**bsorção, **D**istribuição, **M**etabolismo e **E**xcreção) e, ao lado dos Parâmetros Toxicológicos responde por no mínimo 70 % das avaliações finais negativas de fármacos (não recomendáveis) durante a Fase 3 em testes clínicos com humanos.

Esta dissertação emprega *Métodos Quiminformáticos* para obter parâmetros metabólicos de fármacos conhecidos com o objetivo de se chegar a modelos de metabolismo preditivos baseados em correlações entre estrutura e atividade e, por intermédio desta avaliação, desenvolver abordagem similar para fármacos desconhecidos tentando obter modelos metabólicos preditivos baseando-se em correlações de estrutura e reatividade, envolvendo as enzimas citocromo P450 dentro do grupo de enzimas CYP1A2, CYP2C9, CYP2C19, CYP2D6 e CYP3A4. A validação do modelo *in silico* foi desenvolvida por meio de estudos comparativos de perfis metabólicos empregando-se o critério de superposição de dados de compostos com estruturas de referência que mostrem as melhores correlações de estrutura e reatividade considerando enzimas CYP2C9, CYP2D6 e CYP3A4.

Os modelos obtidos podem ser muito úteis na previsão de metabolismo considerando enzimas CYP2C9, CYP2D6 e CYP3A4 para novos tipos de possíveis fármacos, pois, o comportamento referente a tendências de metabolismo de novas entidades químicas pode levar a análises por antecipação de reações enzimáticas. Certamente, este estudo preditivo na Fase I do estudo de fármacos na farmacoterapia reduzirá drasticamente o perfil temporal e o impacto de custos no desenvolvimento de novas substâncias bio-ativas no planejamento da gênese de novos fármacos.

ABSTRACT

“ON THE METABOLIC STUDIES OF DRUGS BY USING ENZYMATIC ACTIVITIES OF CYP450'S TO ENDS UP CORRELATIONS BETWEEN STRUCTURE AND ACTIVITIES”

Carlos Alberto Montanari
Orientador

Renato de Lima Bauab

Keys Words: *Medicinal Chemistry, Drugs, Pharmacotherapy, Chemoinformatic, Pharmacokinetics, Metabolism, Enzymes, Cytochromes*

The Medicinal Chemistry is a multidisciplinary science with direct action over specific knowledge focusing Chemistry, Biology, Medicine, Physiology, among others domains of fundamental and technologic studies. This science also acts between several scientific interfaces like Biochemistry, Biophysics, Molecular Biology, Biologic Chemistry, and others

The investigation on drugs metabolism is the first and essential phase on modern Pharmacotherapy since the pharmacokinetics' parameters are the most relevant impute on the beginning of the Phase 3 on Human Clinical Tests.

The overall pharmacokinetics data base is known as ADME (**A**bsorption, **D**istribution, **M**etabolism and **E**xcretion) and side by side with the Toxicological Parameters responds at least of 70 % of the total final evaluation of drugs negatively evaluated (not recommended) during the Phase 3 on Human Clinical Tests.

This dissertation employs *Chemoinformatic Methods* to obtain metabolic parameters of known drugs with the mean objective of to ends up a predictive metabolic pattern based on correlation of structure and activity, and by mean of this evaluation, to perform similar approaches on unknown drugs trying to get predictive metabolic pattern based on correlation of structure and reactivity, involving the cytochrome enzymes P450 on the group of CYP1A2, CYP2C9, CYP2C19, CYP2D6 e CYP3A4. The pattern *in silico* validation was developed by mean of a comparative studies of metabolic profile ad by using the superposition criteria of the reference structures compounds data base having better correlation of structure and reactivity considering the enzymes CYP2C9, CYP2D6 e CYP3A4.

The obtained pattern can be useful on metabolism prediction considering enzymes CYP2C9, CYP2D6 e CYP3A4 for new kinds of possible drugs, since this behavior concerning metabolic trends of newer chemical entities can arise anticipated analysis of enzymatic reactions. Surely, this predictive studies on Phase 1 of drugs on Pharmacotherapy will reduces drastically the time profile and the costs impacts on developing of new bioactive substances on planning genesis of new drugs.

1. INTRODUÇÃO

Métodos de busca por similaridade são amplamente utilizados em técnicas modernas de ensaio virtual na pesquisa farmacofórica e são sustentados pelo princípio da similaridade, postulando que moléculas estruturalmente similares possuem propriedades similares¹. Alguns motivos para a propagação rápida dos métodos de busca por similaridade são:

1. São baratos, computacionalmente, permitindo buscas em enormes bases de dados;
2. Podem ser utilizados quando existe muito pouco ou nenhuma informação sobre o alvo e apenas um ou dois ativos².

Desde sua introdução, os métodos de busca por similaridade focalizam-se na varredura de base de dados visando identificar novos compostos estruturalmente similares a uma molécula com atividade conhecida (composto de referência ou de consulta). Então, verifica-se se a molécula da base de dados que ainda não foi avaliada é também favorável a ser ativa³. Entretanto, a afinidade por uma enzima alvo é apenas um critério utilizado para selecionar ou descartar um composto durante a busca de novos fármacos. Esta busca constitui a fase farmacodinâmica de um fármaco. Outro requerimento primordial a um fármaco é um perfil farmacocinético apropriado, que incorpora os estudos das propriedades ADME/Tox (absorção, distribuição, metabolismo, excreção e toxidez). Estes estudos tornaram-se essenciais por reduzirem o índice de falhas em estágios mais avançados do processo de desenvolvimento de fármacos. De fato, estimativas mostram que aproximadamente 39% dos fracassos na etapa investigativa do desenvolvimento de fármacos devem-se a índices farmacocinéticos insatisfatórios⁴.

Outro parâmetro essencial que um fármaco deve demonstrar é a seletividade por

sua enzima alvo relativa a outras estruturalmente parecidas⁵. Uma baixa seletividade costuma apresentar um baixo índice terapêutico, o que geralmente resulta em efeitos tóxicos de um composto. Entretanto, avaliar a seletividade de um composto é mais complexo em relação à busca por compostos ativos, comparativamente, pois requer que compostos quimicamente similares sejam discriminados entre si por suas diferentes atividades em relação a múltiplos membros de uma família alvo⁶. Neste contexto, estudos de seletividade de compostos tornaram-se cada vez mais populares na química biológica e na quimiogenômica em que, pequenas moléculas são utilizadas como sondas seletivas para famílias proteicas e/ou membros individuais de uma família a fim de se avaliarem suas funções^{7,8}. Entretanto, muito pouca informação está disponível para predições de seletividade e apenas muito poucos estudos computacionais avançaram ao ponto de atingir a seletividade de ligantes dentro de uma família alvo⁹.

1.1. MODELOS *IN SILICO*

Identificado(s) o(s) alvo(s), a descoberta e o desenvolvimento de ligantes em NCE (New Chemical Entity) são desenvolvidos pelo uso de coleções virtuais e reais de compostos baseadas em bancos de dados, coleções combinatórias e produtos naturais^{10,11,12}. Os modelos *in silico* são amplamente usados nesta fase para definir o espaço químico-biológico ligante-similar com a inclusão de informações referentes à afinidade pelo alvo, potência e propriedades farmacocinéticas constituídas de absorção, distribuição, metabolismo e excreção (ADME). Os estudos da viabilidade sintética são incluídos e algumas moléculas sintetizadas são usadas para validar os modelos iniciais. Os métodos em quiminformática¹³ estabelecem-se pelo gerenciamento, manipulação e otimização das propriedades úteis no processo de descoberta e desenvolvimento de novos fármacos. Eles são fundamentais para:

- (i) a representação e comunicação de dados,
- (ii) o planejamento e organização de banco de dados¹⁴,
- (iii) a predição da estrutura e propriedade,
- (iv) a caracterização das propriedades de fármaco ou protótipo (substância matriz),
- (v) o estabelecimento da similaridade e diversidade molecular,
- (vi) o planejamento e otimização de coleções de compostos,
- (vii) o ensaio virtual e busca em banco de dados,
- (viii) a classificação e seleção de compostos,
- (ix) a geração de relações qualitativas e quantitativas estrutura-atividade (SAR e QSAR) e estrutura-propriedade (SPR/QSPR),
- (x) a geração de modelos estatísticos e descritores (quimiometria),
- (xi) a predição de características de compostos *in vitro* e *in vivo*.

Os métodos em quiminformática¹⁵ também são usados na identificação e otimização da NCE. Todavia, neste estágio da descoberta de fármacos as relações entre estrutura-atividade e estrutura-propriedade¹⁶ (SAR/SPR) são geradas para pequenas séries de moléculas bioativas com o objetivo primeiro de serem usadas como filtros de seleção de novas moléculas não antes testadas contra o alvo biológico de interesse. Os estudos em fase pré-clínica são então realizados apenas para aquelas moléculas com maiores valores agregados. Dessas, as moléculas candidatas a fármacos entram em fases clínicas e aquela(s) aprovada(s) pelos órgãos governamentais que regulam o uso de medicamentos, entra(m) na terapêutica para o tratamento da doença alvo.

A aplicação da química medicinal¹⁷ como ciência fundamental para a descoberta e o desenvolvimento de NCE estabelece-se pela interação de substâncias químicas

com alvos biológicos e pode resultar em um fenômeno de ação-perturbação do sistema biológico escolhido. A sua descrição de forma qualitativa é realizada e as classes de compostos são identificadas com potencial atividade farmacológica. O parâmetro biológico quantitativo obtido através da medida da resposta farmacológica é usado para o estabelecimento de relações quantitativas estrutura-atividade (QSAR). Para isto, há delineamento da estrutura e subestrutura¹⁸ moleculares por determinação experimental ou cálculo de parâmetros físico-químicos. Todos os estudos são norteados pela aplicação combinada e integrada de métodos de planejamento baseado na estrutura do ligante (LBDD) e do receptor (SBDD).

A busca por moléculas química e metabolicamente estáveis representa hoje um dos maiores obstáculos no estudo do metabolismo de fármacos, onde fatores como dureza e moleza metabólica são bases fundamentais. Com a finalidade de diminuir custos (financeiros e de tempo) no desenvolvimento de compostos ativos, recentemente condenados por problemas farmacocinéticos e toxicológicos ocultos, a química medicinal integra informações metabólicas no planejamento de fármacos e desenvolvimento de moléculas. Aspectos de interesse no metabolismo de fármacos neste estágio são: a química e bioquímica das reações metabólicas; as conseqüências de algumas reações na ativação e desativação, intoxicação e desintoxicação; predição do metabolismo de fármacos; desenvolvimento de pró-fármacos; alterações de suas propriedades físico-químicas (acidez, basicidade, lipofilia etc.) resultante da biotransformação¹⁹.

1.2. FARMACOCINÉTICA

De aproximadamente 250 compostos que entram no desenvolvimento pré-clínico, somente cinco são avaliados clinicamente e apenas um é aprovado pela

agência estadunidense FDA. Fatores farmacocinéticos de ADME/Tox (Absorção, Distribuição, Metabolismo e Excreção; Toxicidade) são citados em 70% das falhas destes compostos em fases clínicas. Assim, a identificação de ligantes e, posteriormente, candidatos com perfil farmacocinético inapropriado tornou-se uma das maiores prioridades no planejamento de fármacos. A prévia identificação é normalmente facilitada por testes *in vitro* e modelos *in silico*²⁰.

1.3. METABOLISMO DE FÁRMACOS

O *Metabolismo de Fármacos* é o processo desenvolvido pelo organismo, ocorrendo com rompimento e/ou adição de ligações químicas na molécula de fármacos, com o objetivo de promover sua remoção do sistema. Desta forma, evita-se o acúmulo e, posteriormente, a intoxicação por este xenobiótico.

Atualmente, diversos fármacos utilizados na terapêutica requerem um determinado tipo de metabolismo, podendo ser desde uma simples hidroxilação catalisada por mais de uma enzima, bem como, muitas vezes, por meio de uma via específica, com atuação de apenas uma enzima com capacidade de, por exemplo, catalisar a N-desmetilação. Obrigatoriamente, os fármacos necessitam apresentar algum grau de solubilidade lipídica, pois, sem isto, não são capazes de cruzarem as membranas celulares ricas em lipídeos para atingir seus alvos enzimáticos. Entretanto, o corpo possui dificuldade em eliminar compostos lipossolúveis via urina ou bile. Majoritariamente o metabolismo de fármacos ocorre no interior de células hepáticas e, em menor incidência, em células pulmonares, encefálicas e na região do intestino delgado.

O processo metabólico de fármacos é mediado por diversas enzimas e, atualmente, esses catalisadores biológicos podem ser divididos em 3 grupos (enzimas

de Fase I, Fase II e Fase III). A Fase I é mediada pelas enzimas da família do citocromo P450, promovendo pequenas modificações (por exemplo, Hidroxilação, N-desmetilação, O-desetilação entre outras poucas possíveis transformações responsáveis pela maioria das interações fármaco-fármaco).

Enzimas de Fase II – também conhecidas como enzimas de conjugação – são responsáveis pela conjugação do metabólito com uma molécula endógena pelo processo de glicuronidação e sulfonação de exógenos, que resistiram ao metabolismo de fase I e não são suficientemente polares para serem eliminados pela urina²¹.

O terceiro grupo é composto por enzimas transportadoras, agindo com “bombas” para expurgarem xenobióticos do interior de células para os tubos renais ou dutos biliares. Também, Enzimas da Fase III atuam como barreiras contra fármacos de absorção oral e no intestino, prevenindo sua penetração no cérebro, testículos, ovários glândulas adrenais e placenta²².

1.4. ENZIMAS DO CITOCROMO P450

A citocromo P450 (CYP) é uma proteína essencial para o metabolismo de fármacos, pois, a maioria dos fármacos é metabolizada pelas enzimas CYP no fígado e no intestino. Existem várias subclasses na família da CYP e cada subclasse realiza reações mais específicas. Entretanto, a seletividade dos membros da família da CYP é pequena. Hoje mais de 60% dos fármacos existentes são metabolizados pela CYP3A4²³ e 20% pela CYP2D6²⁴. Sendo que, praticamente, dois terços dos 200 fármacos mais utilizados na terapêutica são metabolizados pelas enzimas do citocromo P450, com um percentual relativo aproximado de 46% pela CYP3A4, 16% pela 2C9, 12% pela 2C19 e CYP2D6 e 9% pela família 1A²⁵.

1.4.1. CITOCROMO P450 1A2

A enzima citocromo P450 1A2 está envolvida no metabolismo de várias substâncias endógenas, como melatonina e estrógenos, e também na ativação de substâncias pró-carcinogênicas como aminas heterocíclicas, arilaminas e aflatoxina B₁. Também, mais de 20 fármacos presentes na terapêutica são metabolizados, exclusivamente pela CYP1A2. Um exemplo disso é a cafeína, presente em diversas bebidas, comidas e medicamentos, é quase totalmente metabolizada pela CYP1A2. A cafeína é utilizada como padrão para a determinação *in vitro* da atividade enzimática desta enzima²⁶.

1.4.2. CITOCROMO P450 2C9

A segunda enzima P450 mais abundante no intestino delgado é a CYP2C9. Ela está diretamente relacionada ao metabolismo de fármacos anti-inflamatórios não corticóides, como por exemplo ibuprofeno, naproxeno, flurbiprofeno; agentes hipoglicêmicos orais, anticonvulsivantes e diuréticos²⁷. Tem sido alvo de grande pesquisa; por metabolizar boa parte dos anticoagulantes presentes no mercado, sua tipagem pode ajudar a prevenir complicações hemorrágicas²⁸.

1.4.3. CITOCROMO P450 2C19

Nos últimos anos a CYP2C19 vem recebendo uma atenção especial por estar associada ao metabolismo de fármacos inibidores da bomba de prótons como o omeprazol^{29,30}. Somente na população caucasiana existem pelo menos 7 diferentes

alelos desta enzima sendo estudados com o intuito de planejar fármacos que não tenham seu metabolismo influenciado por estes alelos.

1.4.4. CITOCROMO P450 2D6

Uma característica alvo de grande atenção na terapêutica atual é o polimorfismo da enzima CYP2D6. Já existem mais de 50 genótipos diferentes para os alelos da CYP2D6 e, sua presença, em seres humanos varia conforme as etnias analisadas.

Essas diferenciações em seu genótipo se manifestam com 5 comportamentos diferentes, são eles: aumento da atividade enzimática, inativação da enzima, ausência da enzima, instabilidade da enzima ou alteração na afinidade por substratos. Dados europeus baseados em seus genótipos sugerem que aproximadamente 70% da população apresentam um baixo metabolismo exercido pela CYP2D6, enquanto 1-10% apresentam um metabolismo ultra-rápido³¹. A Tabela 1 ilustra a distribuição dos alelos entre as diferentes etnias.

Tabela 1: Distribuição dos polimorfismos da CYP2D6 dentro das etnias

Alelo de Maior Variância	Mutação	Conseqüência	Frequência do Alelo (%)			
			Caucasianos	Asiáticos	Africanos Negros	Etíopes e Árabes - Sauditas
CYP2D6*2xn	Duplicação do gene	Aumento na atividade enzimática	1-5	0-2	2	10-16
CYP2D6*4	Defeito de entrelaçamento	Inativação da enzima	12-21	1	2	1-4
CYP2D6*5	Apagamento do gene	Ausência da enzima	2-7	6	4	1-3
CYP2D6*10	P34S, S486T	Enzima instável	1-2	51	6	3-9
CYP2D6*17	T107I, R296C, S486T	Afinidade por substratos alterada	0	0	20-35	3-9

1.4.5. CITOCROMO P450 3A4

A CYP3A4 é uma das enzimas mais importantes porque encontra-se em maior quantidade em tecidos críticos como, por exemplo, o fígado e o trato gastrointestinal.

Metaboliza a maior parte dos fármacos incluindo midazolam, carbamazepina, lidocaína, ciclosporina dentre outros. Sua principal forma de atuação é por N-desmetilação, N-desetilação, C(1)-hidroxilação e C(6 β)-hidroxilação. Esta enzima é importantíssima no metabolismo de hormônios esteróides; apresenta atividade superior nas mulheres³². Seu estudo pode ser considerado o de maior importância no metabolismo de fármacos, pois, devido à sua grande capacidade de metabolizar diferentes substratos, influi diretamente em estudos de associação fármaco-fármaco e determinação de dosagem pela taxa metabólica.

Seguramente, hoje em dia, mais de 60% das moléculas com bom potencial para fármaco, estudadas pela indústria farmacêutica são reprovadas nos testes de ADME/Tox (absorção, distribuição, metabolismo, excreção e toxidez). Tornam-se necessários estudos de Modelos Farmacocinéticos, pois, eles podem reduzir o tempo e o custo das pesquisas na busca por novos fármacos. Além disso, podem auxiliar na identificação de grupos ou fragmentos moleculares estáveis metabolicamente e assim, viabilizar o desenvolvimento total e/ou parcial de novos fármacos.

O paradigma moderno da química medicinal empregado neste projeto estabelece-se pelo uso de métodos em quiminformática integrados aos estudos clínico-experimentais já realizados para os atuais fármacos. A idealização de ensaios virtuais com a finalidade de identificar similaridade química no processo metabólico pelas enzimas da família do citocromo P450 estabelece os primeiros requisitos fundamentais para o processo de descrição das propriedades farmacocinéticas de exógenos no corpo humano. Dessa forma, a quiminformática conterà informações sobre o planejamento das coleções de compostos e sobre o gerenciamento dos bancos de dados necessários para a realização de ensaios virtuais mais precisos.

Como a maioria dos fármacos é metabolizada por enzimas do citocromo P450, as interações de novos fármacos em potencial com as CYP1A2, CYP2C9, CYP2C19,

CYP2D6 e CYP3A4 são caracterizados por testes *in vitro* como parte da pesquisa e do desenvolvimento farmacêutico, tipicamente. Isto deve facilitar a descoberta e o desenvolvimento de fármacos mais seguros e com poucos efeitos colaterais, interações fármaco-fármaco mínimas, interações fármaco proteína otimizadas, e propriedades farmacocinéticas previsíveis e até mesmo planejadas, no intuito de melhorar a ação terapêutica do fármaco e ainda, assim, permitir a utilização de dosagens cada vez menores na cura de enfermidades.

A seleção virtual para ligações de novas substâncias químicas com as CYP1A2, CYP2C9, CYP2C19, CYP2D6 e CYP3A4 pode aumentar o processo de seleção farmacêutica por gerar informações efetivas sobre o metabolismo e a toxidez³³.

1.5 RECEPTORES 5HT₃

Os receptores do 5HT₃ começaram a serem pesquisados no início da década de 50 e foram das primeiras 3 famílias de receptores 5HT a serem estudados, entretanto, apenas nos últimos trinta anos os primeiros estudos de interação de ligantes foram publicados. Eles são receptores não seletivos dos canais de íons Na⁺/K⁺ e são encontrados na área postrema do sistema nervoso central, no córtex entorrinal e frontal e no hipocampo. Estudos comprovaram algumas similaridades estruturais entre os receptores 5-HT₃ de humanos e outros membros da superfamília do canal de íons, os receptores nicotínicos de acetilcolina.

A aplicação clínica para receptores do 5HT₃ como alvo terapêutico é de grande interesse devido à sua eficiência no tratamento de náuseas e vômitos induzidos por quimioterapia, radioterapia e anestésicos, além de ser efetivo no tratamento de enxaquecas ou dores associadas à enxaqueca. Diversos estudos sugerem que podem ser utilizados inclusive no tratamento de depressão, demência, ansiedade, dor e antipsicóticos³⁴.

2. OBJETIVOS

2.1. OBJETIVOS GERAIS

Empregar procedimentos de quimioinformática no estudo de previsão de metabolismo de fármacos.

2.2. OBJETIVOS ESPECÍFICOS

2.2.1. Estabelecer um modelo *in silico* para avaliação da variabilidade do perfil metabólico de compostos matrizes pela inserção de fragmentos para as CYPs 1A2, 2C9, 2C19, 2D6 e 3A4, utilizando modelos baseados no sítio catalítico destas enzimas.

2.2.1. Definir uma coleção de fármacos com metabolismo conhecido para as enzimas CYPs 2C9, 2D6 e 3A4 e a partir desta coleção determinar qual a enzima de maior afinidade de um composto alvo e qual o átomo de maior probabilidade de metabolismo utilizando métodos de similaridade molecular.

3. FERRAMENTAS E MÉTODOS

3.1. METASITE³⁵

O MetaSite é um software desenvolvido pela Moldiscovery com a finalidade de calcular o sítio com maior probabilidade de metabolismo de qualquer molécula relacionada com o metabolismo oxidativo de citocromos P450 na fase I. Ele emprega o método de mapeamento de energias de interação 3D para proteínas-alvo e sondas químicas (GRID) e a estrutura 3D de um composto desejado. Este método possui uma precisão superior a 85% para os três átomos mais prováveis de metabolismo.

É uma ferramenta automatizada de fácil utilização e pode ser empregado na análise de uma grande coleção de moléculas por meio de um *script em shell*. Embora o MetaSite possa ser utilizado na criação de um extenso perfil metabólico de moléculas, algumas precauções devem ser tomadas, pois em seu algoritmo está inclusa uma Equação de Normalização para os átomos. Em outras palavras, o perfil metabólico de moléculas não pode ser comparado a menos que os objetos de comparação sejam muito próximos uns dos outros (moléculas *me too*). Por exemplo, modificações de um composto matriz podem ser comparados com sua estrutura primária com a intenção de identificar alguma alteração de comportamento metabólico causada pela inserção de um fragmento.

3.2. MOLPRINT2D³⁶

Este método utiliza uma tabela de conectividade da estrutura molecular para calcular os ambientes atômicos, considerando para isso o tipo de átomos vizinhos. Então, os ambientes atômicos são utilizados como representação molecular, e são calculados em um procedimento de duas etapas:

- Em uma estrutura molecular isenta de hidrogênios, todo átomo pesado possui seu tipo atômico Sybyl assinalado.
- Uma impressão individual para cada átomo pesado é calculada utilizando-se distâncias de zero a duas ligações e contando a ocorrência dos tipos atômicos.

3.1.3 UNITY 2D FINGERPRINTS³⁷

Impressões UNITY 2D são uma mistura de impressões e chaves estruturais. Estas impressões denotam a presença ou ausência de caminhos de conexão combinatória enumerados por uma seqüência de bits. O UNITY permite ao usuário especificar os componentes individuais da sequência de dados, os quais podem seguir dentro de um específico intervalo ou variedade de subestruturas. Neste trabalho foi utilizado o padrão de 988 sistemas de bits³⁸.

3.1.4 ROCS³⁹

O método ROCS (*Rapid Overlay of Chemical Structures*) identifica similaridades entre duas moléculas baseado em suas formas moleculares tridimensionais. O ROCS aproxima os volumes moleculares a funções Gaussianas ao invés de esferas sólidas, resultando assim em Equações Matemáticas analíticas e diferenciáveis, que permitem uma otimização rápida e robusta de sobreposição de volume por variância de suas orientações relativas^{40,41}. Uma função de similaridade mede a “distância de forma” entre o par de moléculas na sobreposição dos volumes otimizados.

3.2 QUANTIFICAÇÃO DE DESEMPENHO

O objetivo de uma estratégia de ensaio virtual é o aprimoramento de ligantes conhecidos sobre ligantes inativos, provavelmente, antes de ensaios *in vitro*^{42,43}.

Neste trabalho, o objetivo de sua aplicação foi um pouco diferente: recuperar

substratos conhecidos de não substratos de uma isoforma específica de CYP. Avaliações de diferentes métodos em discriminar entre substratos e não substratos foram desenvolvidas utilizando a área sob a curva (AUC) de uma curva de característica receptora operante (ROC).

Traçar uma Curva ROC consiste em determinar a sensibilidade (Se) e a especificidade (Sp) em todo limiar de possíveis resultados. Se e Sp fornecem informações sobre o número de ativos verdadeiros selecionados e o número de inativos descartados, respectivamente. Ambos os parâmetros podem variar entre 0 e 1. Uma Curva ROC representa a Se como função de $(1-Sp)$ em toda região de detecção possível. Para distribuições ideais, quando não há sobreposição entre os resultados dos ativos e inativos, a Curva ROC sobe de sua origem até o canto esquerdo superior até a total recuperação dos ativos ($Se=Sp=1$), e continua como uma linha reta horizontal até o canto direito superior da Curva, local em que todos os ativos e inativos estão recuperados⁴⁴, o que corresponde a $Se=1$ e $Sp=0$.

Outra forma de interpretar os resultados de Curvas ROC é a área sob a curva (AUC). Uma distribuição ideal de ativos e inativos apresenta uma AUC com valor igual a 1. Uma classificação randômica de compostos pode ser representada por uma diagonal ascendente da origem até o canto direito superior, e isto corresponde a uma AUC de 0,5. Um teste com melhor desempenho em relação a uma discriminação randômica para ativos e inativos, recuperam-se compostos com uma AUC entre 0,5 e 1.

Geralmente, quanto maior a AUC, mais efetivo é o trabalho da seleção virtual em discriminar compostos ativos de inativos.

4. RESULTADOS

4.1. SETRONS

Foram analisados os perfis metabólicos de seis setrons, são eles: alosetron, dolasetron, granisetron, ondansetron, palonosetron e tropisetron. Algumas pequenas variações foram feitas em suas estruturas primárias conforme sua forma de substrato para CYP disponível na literatura na intenção de se alcançar um resultado mais próximo dos ensaios *in vivo*, estas modificações estão comentadas no perfil das respectivas moléculas abaixo.

4.1.1. ALOSETRON

Conforme Ismail *et al.*⁴⁵ e ao atual registro do Lotronex[®]⁴⁶ na FDA, o metabólito principal do alosetron é sua forma 6-hidroxilada (M1) e as CYPs envolvidas em seu metabolismo são 2C9 (30%), 3A4 (18%) e 1A2(10%) respectivamente. Porém, testes *in vivo* sugerem que a CYP1A2 possui participação mais proeminente no metabolismo do alosetron.

Esta posição de metabolismo corresponde ao H22. Existem pelo menos 15 metabólitos prováveis para o alosetron, porém os oito metabólitos mais significativos correspondem a pouco mais de 1% da dose; entretanto, apenas a forma 6-hidroxi deve ser investigada por representar mais de 15% da dose e sua forma glicuronizada a aproximadamente 14%.

Analisando a saída de dados do Metasite (Figuras 1-4), é claramente visível a correlação entre os índices de acessibilidade dos três citocromos com testes *in vivo* apresentando o maior valor de E_i para M1 contra todos os demais metabólitos, apenas para CYP1A2. Conforme dito anteriormente, o Metasite apenas sugere os sítios mais

prováveis de metabolismo, portanto ele não substitui a importância de dados experimentais.

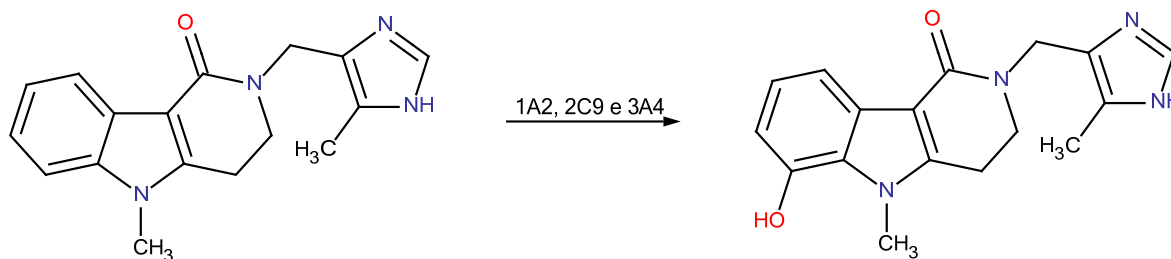


Figura 1: Metabólito experimental do alosetron por oxidação via CYP.

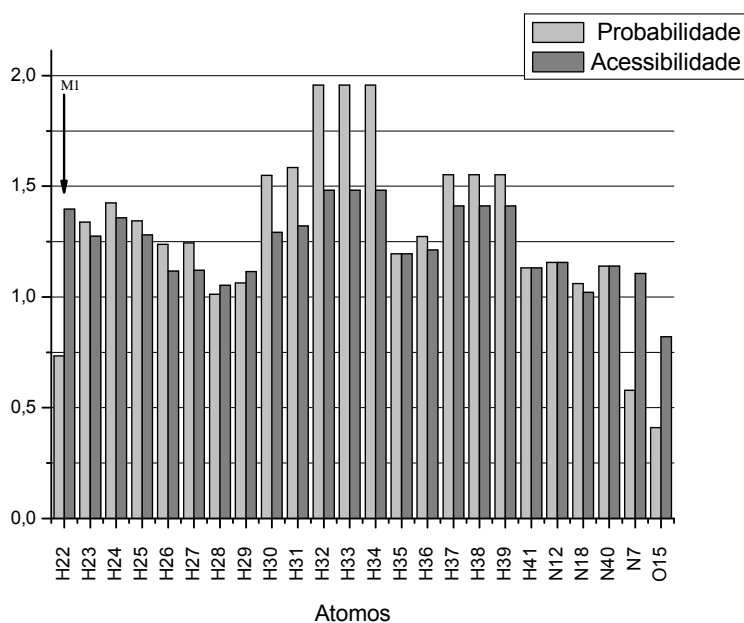


Figura 2: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP1A2 para o alosetron.

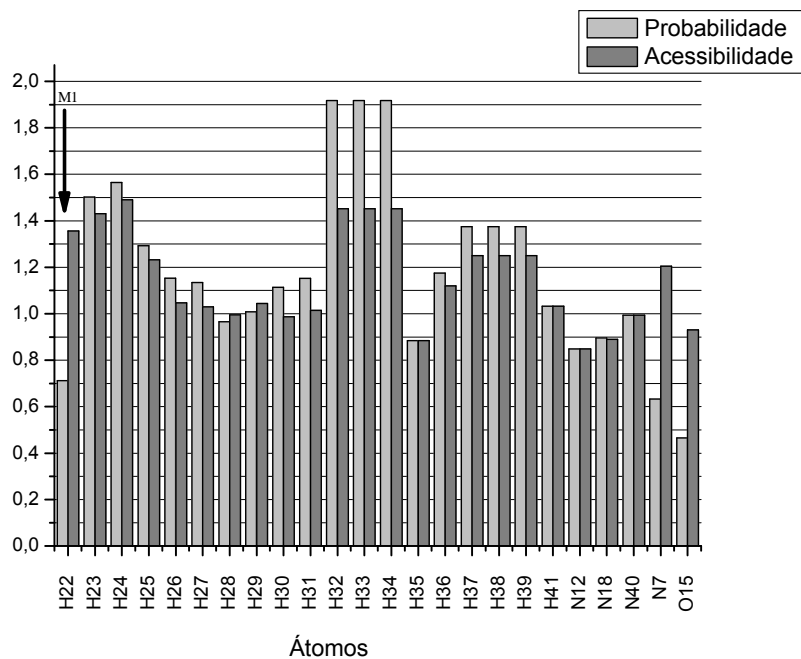


Figura 3: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2C9 para o alosetron.

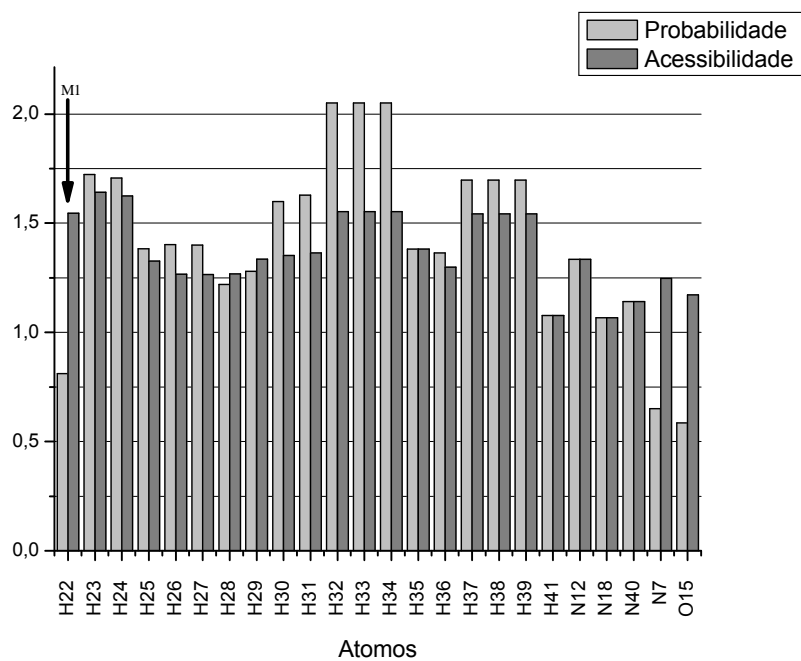


Figura 4: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o alosetron.

4.1.2. DOLASETRON

Anzemet[®] é um pró-fármaco que é rapidamente convertido para o hidrodolasetron via reação de redução catalisada pela enzima carbonil redutase⁴⁷. A maioria do hidrodolasetron formado é o enantiômero R(+) (>85%); é metabolizado pela CYP2D6 e em pequenas quantidades pela CYP3A4, prioritariamente e, os percentuais de seus principais metabólitos de citocromo estão entre 1,9-2,3% como 5-OH-hidrodolasetron e entre 5,8-6,6% como 6-OH-hidrodolasetron⁴⁸.

As Figuras 6-9 mostram o Metasite apontando o metabólito 5-OH-hidrodolasetron como o favorito tanto para a CYP2D6 quanto para a CYP3A4 (i.e., ocupam a primeira posição no gráfico de P_{MS}^i) e o metabólito 6-OH-hidrodolasetron encontra-se em uma margem entre a quarta e a oitava posição no gráfico de P_{MS}^i .

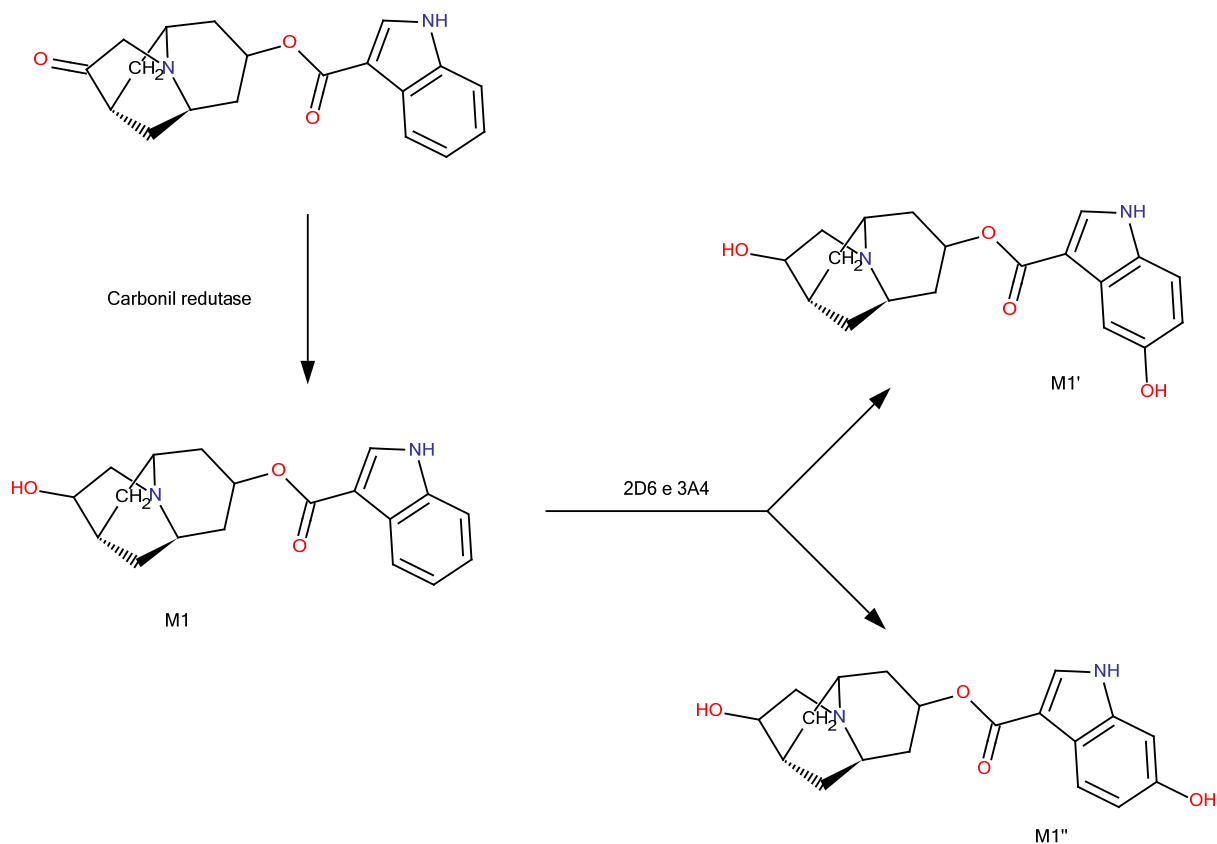


Figura 5: Metabólitos experimentais do dolasetron e do hidrodolasetron por oxidação via CYP.

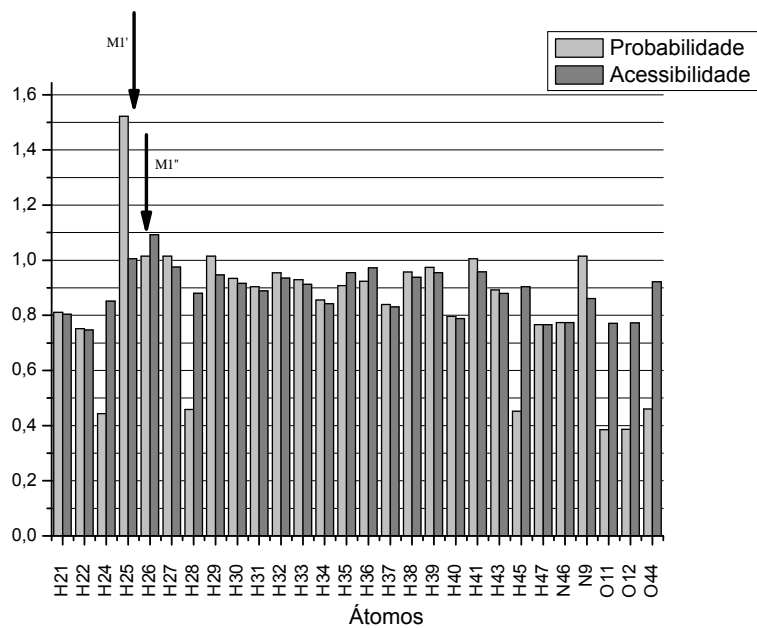


Figura 6: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o R(+)-hydrodolaseton.

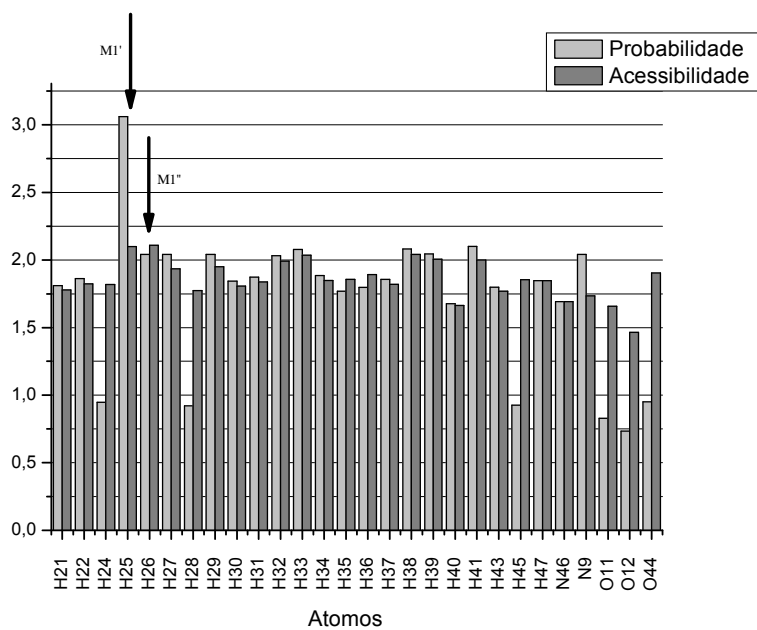


Figura 7: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o R(+)-hydrodolaseton.

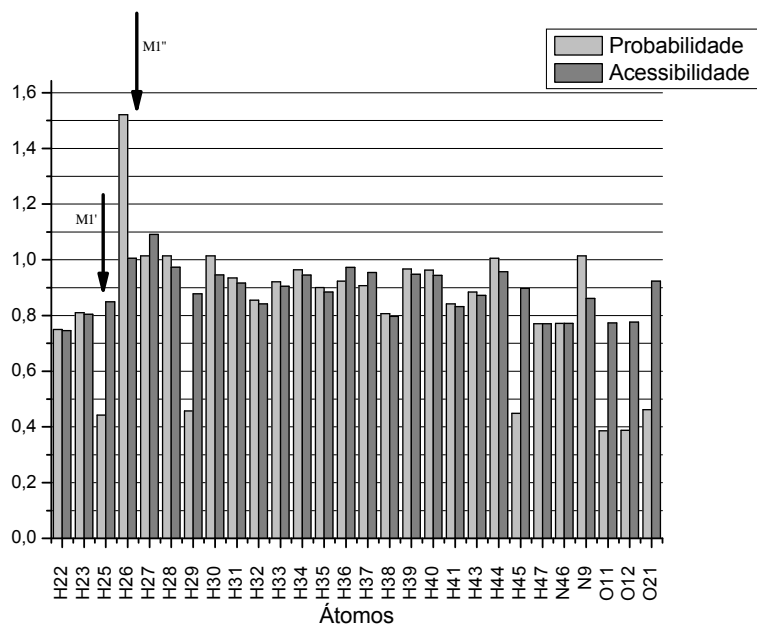


Figura 8: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o S(-)-hydrodolasetron.

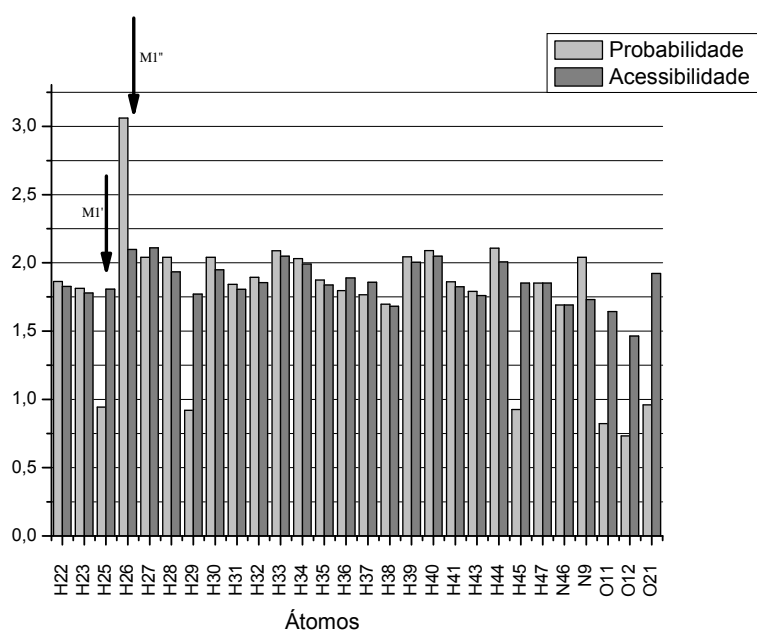


Figura 9: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o S(-)-hydrodolasetron.

4.1.3. GRANISETRON

Este é um fármaco com um comportamento muito interessante, também conhecido como Kytril® e com a fórmula no índice CA. Este fármaco não apresenta qualquer atividade metabólica com a CYP2D6, como esperado⁴⁹. Por outro lado, o granisetron é um substrato para a CYP1A1. Conforme Testa e Krämer⁵⁰, a família da CYP1 de humanos age como as aril-hidrocarboneto hidroxilases e é induzida por hidrocarbonetos aromáticos policíclicos. Esta característica indica a existência de alguma conjugação de densidade eletrônica intramolecular permitindo à CYP1A1 reconhecer o granisetron como seu substrato.

Em uma investigação anterior Nakamura *et al*⁵¹ relataram o granisetron sendo majoritariamente metabolizado para o 7-hidroxi granisetron e uma menor quantia para o 9'-desmetilgranisetron.

Para esta predição metabólica, o Metasite não foi de grande ajuda em estimar a P_{SM} para o 7-hidroxi granisetron, pois, não existe publicada até o momento alguma estrutura para a CYP1A1. Sendo assim, o teste foi realizado utilizando-se a estrutura de seu alelo, a CYP1A2, que não atribuiu uma boa posição para o metabólito 7-hidróxi. A predição apontou para a vigésima segunda posição para o H43 ser oxidado pela via metabólica. Por outro lado, a predição para a CYP3A4 foi boa para H23-25 e apontou o 9'-desmetilgranisetron na primeira posição. Figura 10.

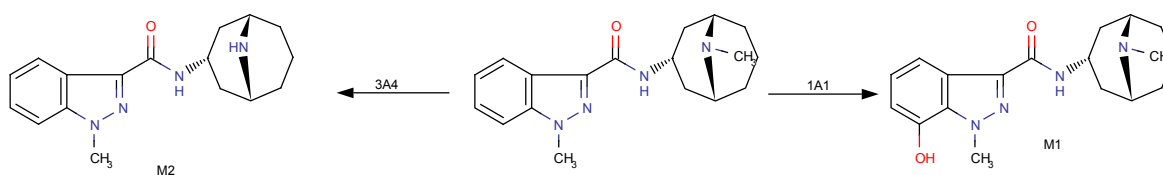


Figura 10: Metabólitos experimentais do granisetron por oxidação via CYP.

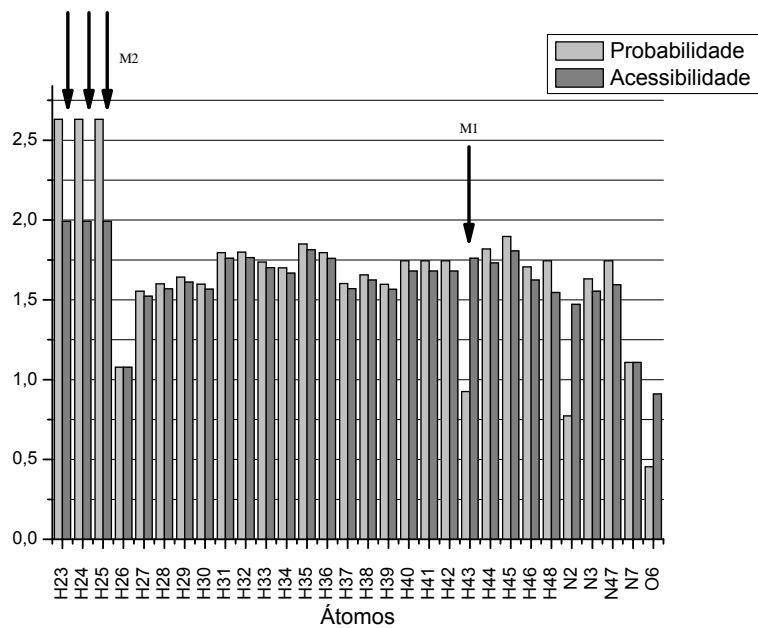


Figura 11: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP1A2 para o granisetron.

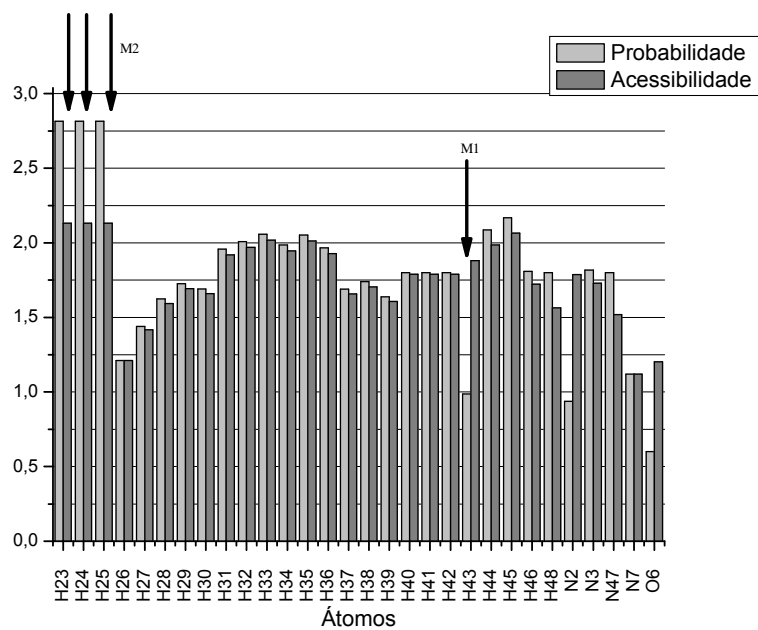


Figura 12: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o granisetron.

4.1.4. ONDANSETRON

Com o primeiro nome comercial de Zofran[®], o ondansetron é comercializado em sua forma racêmica. Tal fármaco atua como substrato para dois citocromos diferentes, CYP1A2, responsável pelo metabólito N-desmetil nos H21-23 e a CYP2D6, que, por sua vez, é responsável pelos isômeros 6,7 e 8 hidróxi-ondansetron nos H26, H25 e H24, respectivamente⁵² e conforme dados experimentais de Fischer *et al*⁵³ não apresenta inibição da CYP3A4, Figura 13.

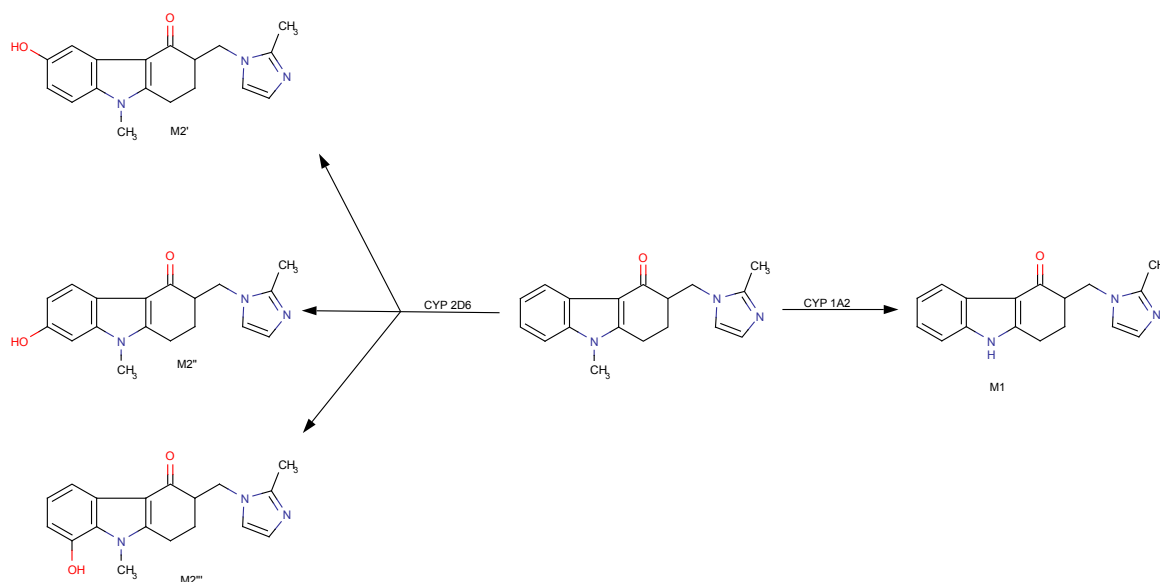


Figura 13: Metabólitos experimentais do ondansetron por oxidação via CYP.

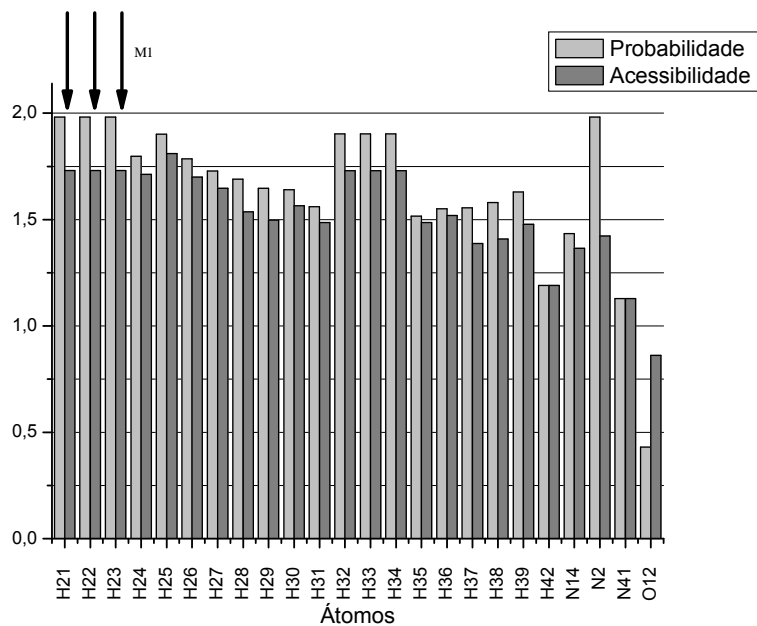


Figura 14: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP1A2 para o R-ondansetron.

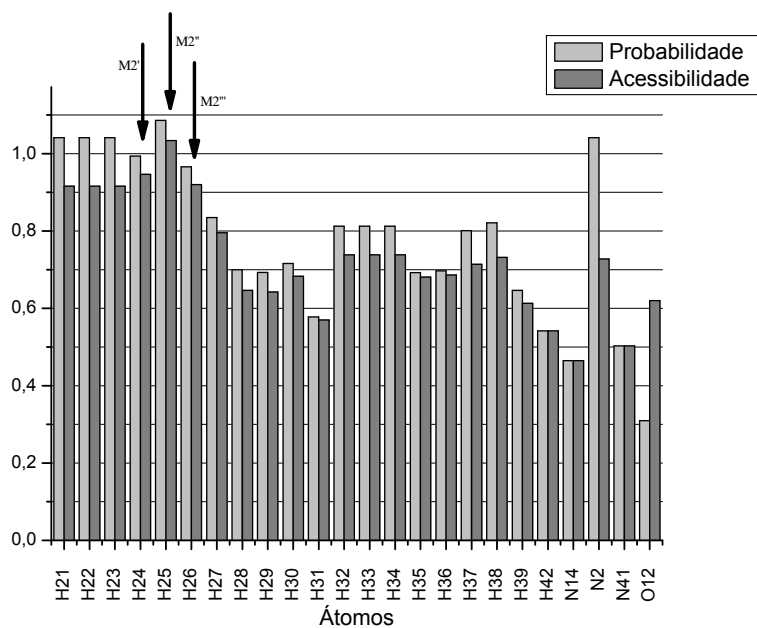


Figura 15: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o R-ondansetron.

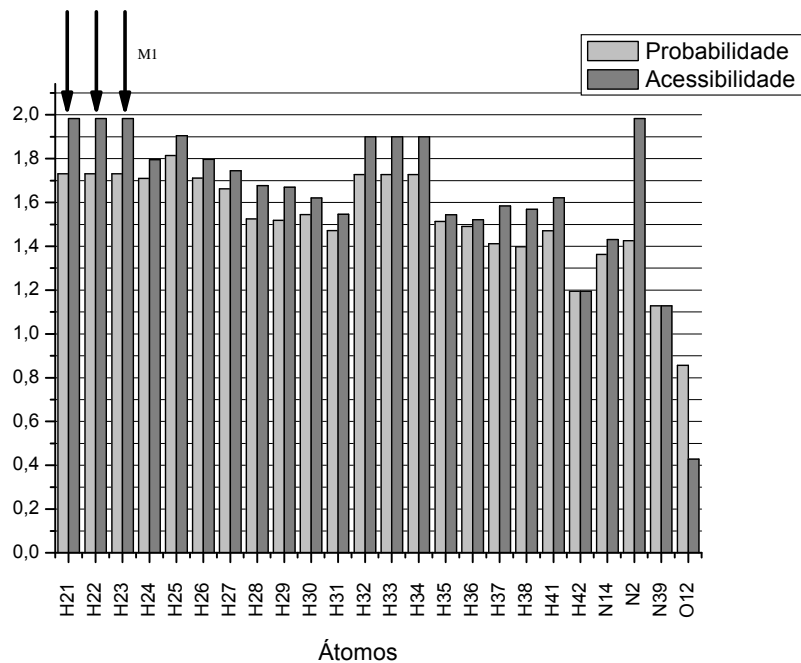


Figura 16: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP1A2 para o S-ondansetron.

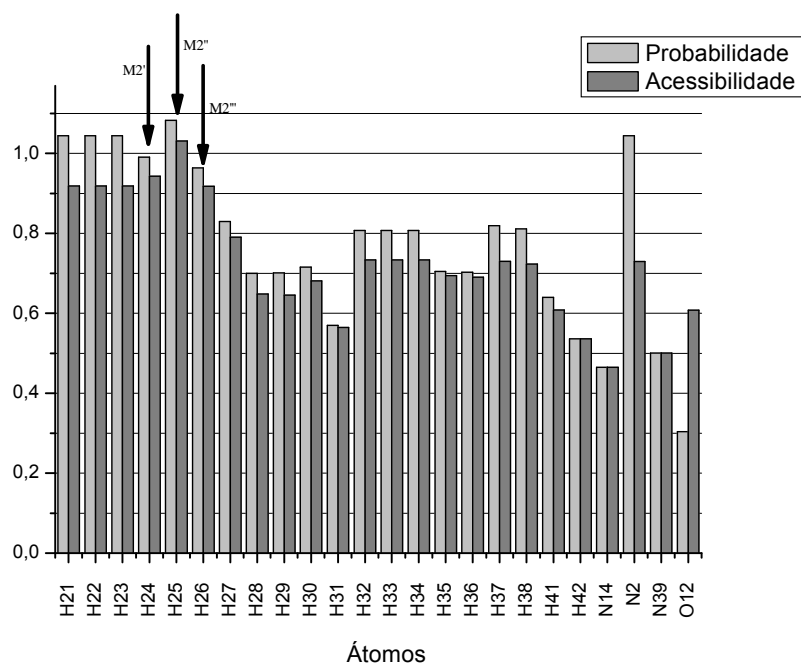


Figura 17: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o S-ondansetron.

4.1.5. PALONOSETRON

Este é um novo fármaco na terapêutica e é um bom exemplo do desenvolvimento racional de fármacos. Apresenta uma elevada meia vida (72 horas) para seu metabolismo de primeiro passo e um bom período de eficácia de 40 horas. Conhecido comercialmente como Aloxi[®] é muito estável ao metabolismo de CYPs e possui apenas um produto de metabolismo mediado por citocromos. Este metabólito é uma hidroxilação estereosseletiva mediada pelas CYP2D6, 3A4 e 1A2, respectivamente⁵⁴.

O Metasite mostrou uma capacidade de predição muito peculiar, pois conforme Stoltz *et al.* existe apenas o isômero (S) do palonosetron hidroxilado no H14, porém o Metasite calculou um melhor P_{SM} para o isômero (R) no H13. Estes resultados são uma boa evidência de que as CYPs possuem alguma propriedade estereosseletiva, embora pequena, como por exemplo, não são completamente promíscuas, como tanto se acredita. Ver Figura 18.

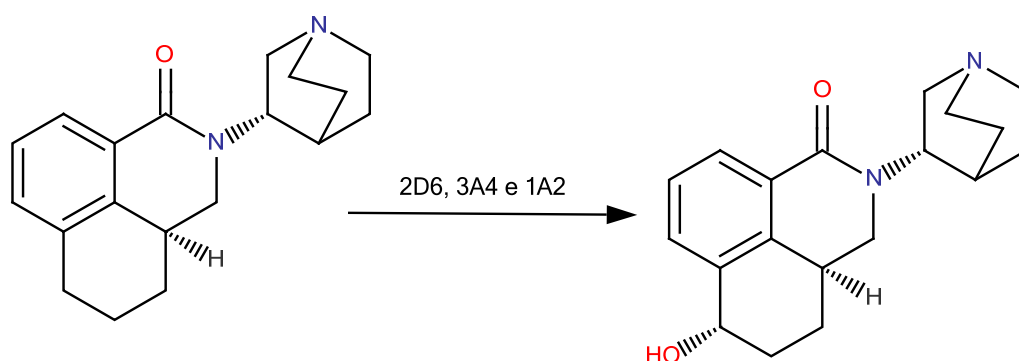


Figura 18: Metabólitos experimentais do palonosetron por oxidação via CYP.

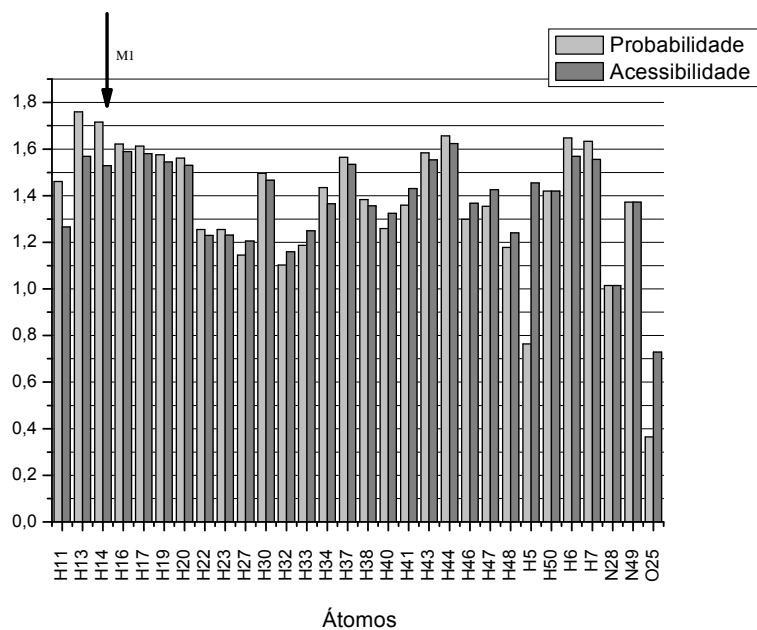


Figura 19: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP1A2 para o palonosetron.

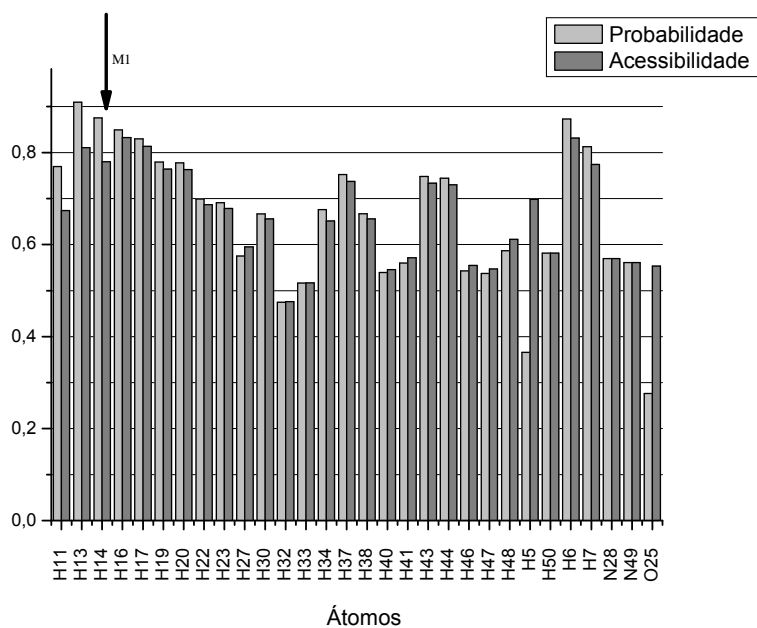


Figura 20: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o palonosetron.

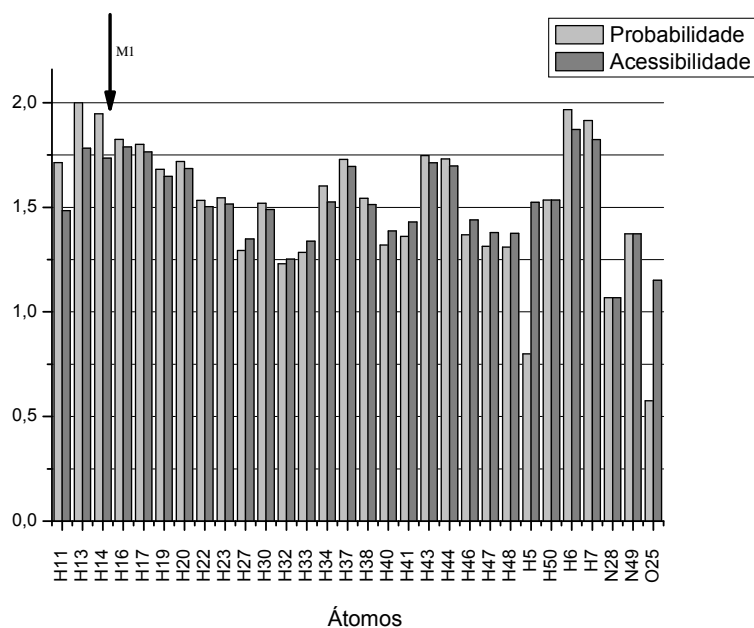


Figura 21: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o palonosetron.

4.1.6. TROPISETRON

Conhecido como Navoban[®], ele apresenta como maior produto metabólico as formas 5 e 6-hidroxiladas, mediadas pela CYP2D6 (aproximadamente 50% do total da quantia administrada) e uma quantidade menor do produto N-desmetilado, mediado pela CYP3A4⁵⁵. Agentes quimioterapêuticos induzem náuseas e vômitos por liberarem a serotonina presente nas células do intestino delgado, que são detectadas pelo sistema nervoso. O tropisetron é um antagonista para os receptores do sistema nervoso e ajuda a prevenir e a tratar este tipo de náusea e vômito. Figura 22.

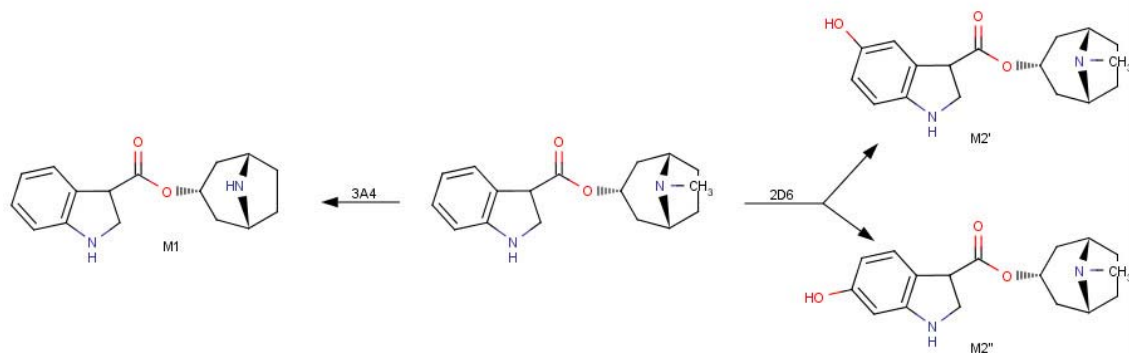


Figura 22: Metabólitos experimentais do tropisetron por oxidação via CYP.

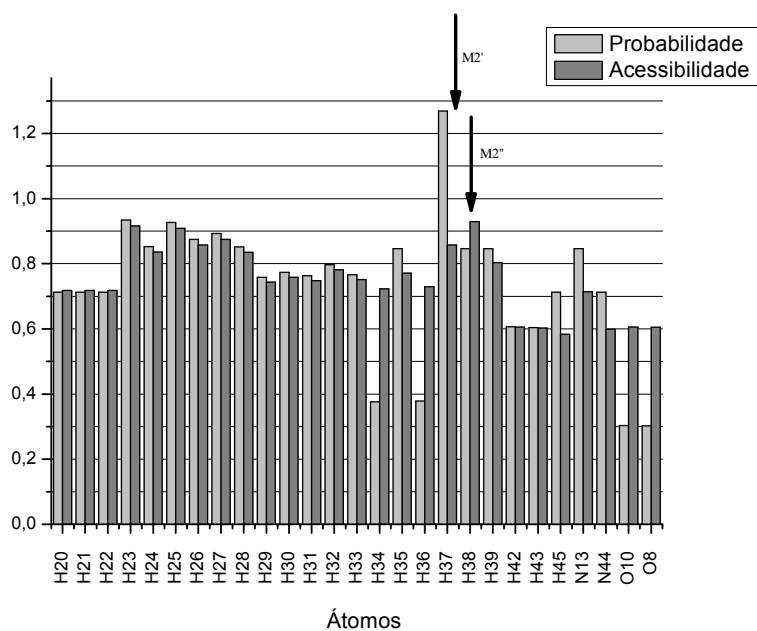


Figura 23: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP2D6 para o tropisetron.

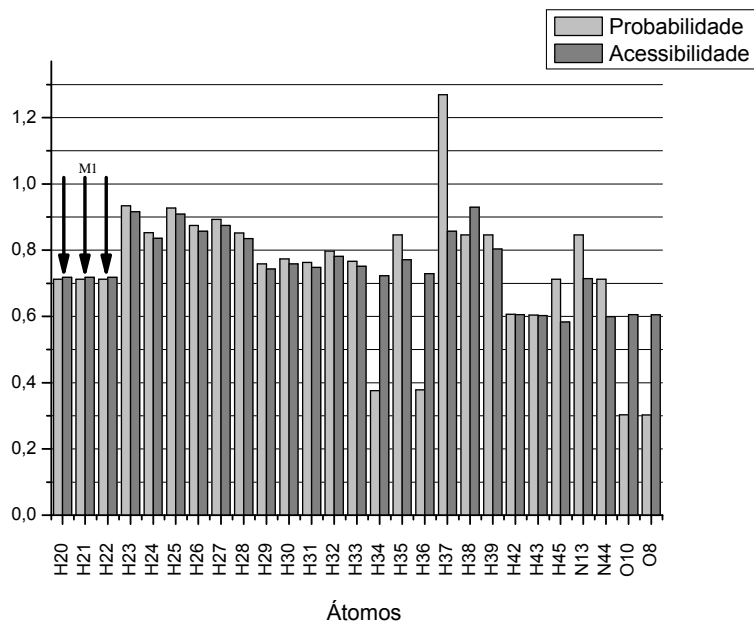


Figura 24: Probabilidade P_{sm}^i (barras claras) e acessibilidade E_i (barras escuras) na CYP3A4 para o tropisetron.

4.2. BASE DE DADOS

Fármacos que atuam como substratos das CYP2C9, CYP2D6 e CYP3A4 foram primeiramente coletados dos trabalhos de Rendic⁵⁶, Terfloth⁵⁷, e Yap⁵⁸. Como foram feitas análises por comparação pareada entre estes 3 citocromos (2C9 *versus* 2D6, 2C9 *versus* 3A4, e 2D6 *versus* 3A4), para considerar-se um composto como sendo substrato de uma enzima é necessário o cumprimento de duas condições:

- (1) ser definido como substrato de apenas uma enzima do par; (2) não ser considerado como inibidor da outra enzima do par. Foi consultada a referência original de cada composto, para confirmar seu enquadramento nestas duas condições. Os compostos foram selecionados um a um, consumindo-se bastante tempo, mas garantiu a alta qualidade dos dados. Com este protocolo foi possível construir uma base seletiva de substratos de dupla comparação. Por exemplo, na Figura 25, 76 compostos são seletivos para a CYP2C9 contra a CYP2D6, e 100 para a CYP2D6 contra a CYP2C9 (direção oposta).

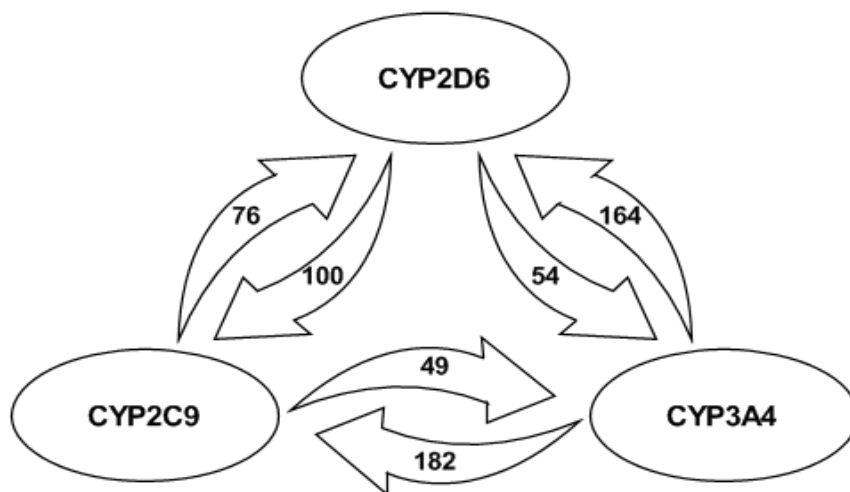


Figura 25: Diagramas de representação dos conjuntos de seletividade de compostos. Cada seta representa um conjunto de seletividade. O número de compostos de cada conjunto também está reportado e a direção das setas definem a seletividade.

As bases de dados geradas são compostas por 625 compostos ao todo, com 439 deles sendo únicos. O conjunto total é composto por 125 substratos da CYP2C9 (20%), 154 da CYP 2D6 (25%) e 346 da CYP3A4 (55%). O número total de compostos difere no número de compostos únicos porque um substrato pode apresentar mais de uma análise por comparação pareada. Por exemplo, o fármaco aceclofenaco foi considerado um substrato para a CYP2C9 em ambas as análises por comparação pareada, CYP2C9-CYP2D6 e CYP2C9-CYP3A4.

As análises de algumas propriedades físico-químicas dos compostos estudados revelaram a presença de valores similares em praticamente todas as propriedades. As moléculas pertencentes ao conjunto da CYP2C9 e CYP3A4 são aquelas com mais propriedades relacionadas, com a massa molar sendo um pouco mais elevado no conjunto da CYP3A4. Entretanto, também, o conjunto da CYP2D6 demonstra muitas propriedades comparáveis com os outros dois conjuntos de dados, a massa molar média e área de superfície polar média possuem os menores valores de todo o conjunto, indicando os compostos da CYP2D6 como sendo menores em comparação

com os das outras duas enzimas. A construção do conjunto de dados teve um efeito pronunciado na eficácia dos testes de varredura virtual⁵⁹. O desempenho dos métodos de varredura virtual em 3D é comparável com um simples método de 1D (massa molar, logP, número de doadores de ligações de hidrogênio, número de aceitadores de ligações de hidrogênio e número de ligações rotacionais)⁶⁰. Este resultado indica que os compostos ativos são muito dissimilares do conjunto de compostos inativos, pois, um simples método de 1D já é bastante eficiente em separar estes dois conjuntos. Logo, os bons resultados encontrados em diversas publicações na área de varredura virtual retrospectiva são puramente relacionados com as diferenças de propriedades simples entre os compostos ativos e inativos. Assim, um conjunto inativo pode ser tão similar quanto possível dos compostos ativos para obter-se uma indicação confiável da utilidade do método de varredura virtual. Para isto, a coleção virtual em estudo é apropriada em acessar a utilidade das ferramentas empregadas neste trabalho. Para isto, não pode haver diferença estatística dentre simples propriedades moleculares de 1D, Tabela 2.

Tabela 2: Média e desvio padrão de algumas propriedades 1D da base de dados.

	CYP2C9		CYP2D6		CYP3A4	
	Média	Desv. Pad.	Média	Desv. Pad.	Média	Desv. Pad.
HBA	5	2	4	2	5	3
HBD	1	1	2	1	2	1
MW	321	88	303	77	369	114
RB	5	3	5	3	5	3
LogP	3	2	3	2	3	2
PSA	69	31	44	23	70	35

4.2.1. CYP2C9 x CYP2D6

Foram utilizados métodos de busca por similaridade 2D e 3D para a realização da sistemática de similaridade por comparação pareada. Foi calculada a média de Tanimoto, que é a média geométrica do coeficiente de Tanimoto, para cada composto utilizado como referência⁶¹. Por exemplo, para o par CYP2C9-CYP2D6 todos os compostos foram utilizados como referência e todo o conjunto foi alinhado contra cada composto de referência. Logo a média de Tanimoto corresponde à similaridade média quando toda a base de dados está alinhada (e pontuada) em relação a um composto de referência específico. Como mostra a Figura 26, o modo padrão do ROCS é aquele no qual a média de Tanimoto apresenta altos valores. A inclusão de características químicas pelo uso do “campo de força da cor” (chamado de *color score*) na superposição e pontuação por similaridade não aumenta o resultado do ROCS e de fato fica pior em relação à sobreposição de forma. Como o campo de força combo (combo score) é a soma do campo de força de cor com a forma, seu valor da média de Tanimoto é intermediária dentre estes campos de força.

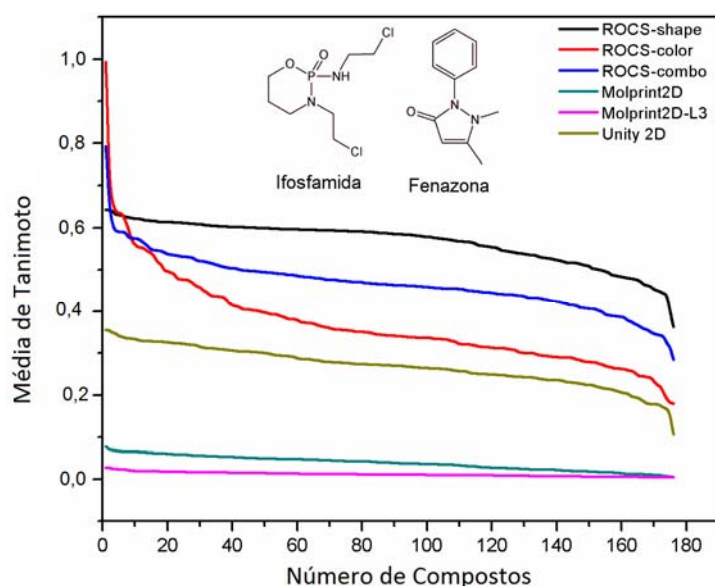


Figura 26: Valor da média de Tanimoto para a sistemática de similaridade por comparação pareada da base de dados das CYP2C9-CYP2D6. Na figura estão as estruturas 2D dos dois compostos atípicos à base de dados.

Curiosamente, dois substratos da CYP2C9, ifosfamida e fenazona, são aparentemente estranhos ao conjunto quando o campo de força de cor é utilizado, obtendo os mais altos valores na média de Tanimoto: 0,992 e 0,705, respectivamente. Atualmente, nem o composto de referência foi capaz de produzir um bom alinhamento com estes dois substratos, pois, a média de similaridade (expressada pelo coeficiente de Tanimoto) para estes compostos alinhados com cada composto de referência foi de 0,154 e 0,265, respectivamente. A ifosfamida é um membro da classe das oxazofosforinas e é um agente alquilante, mostarda nitrogenada utilizado no tratamento do câncer.

As impressões do Unity 2D demonstraram um valor ligeiramente menor para a média de Tanimoto em relação ao *color score* do ROCS. Finalmente, o Molprint2D foi que apresentou os menores valores para as médias de Tanimoto. Entretanto, esta observação não implica a eficiência desta impressão em discriminar os substratos de uma enzima. Como pelo modo padrão, o Molprint2D gera impressões para a molécula utilizando a distância de duas ligações, foi investigada a possibilidade de utilizar mais uma camada para poder elevar os resultados apresentados por este método. Infelizmente, a utilização de três camadas não apresentou um desempenho tão bom quanto as impressões geradas com duas camadas.

4.2.1.1. DIVERSIDADE DE COMPOSTOS

A diversidade estrutural dos compostos do conjunto de dados CYP2C9-CYP2D6 foi avaliada utilizando-se a sistemática de similaridade por comparação pareada. O quadro resultante das matrizes de Tanimoto para cada método de similaridade está ilustrado na Figura 27.

Para o método de busca por similaridade 3D (ROCS), a análise desta figura revelou a diversidade estrutural dos compostos, quantificados pelo coeficiente de

Tanimoto, diferindo consideravelmente conforme o nível de informação utilizado para pontuar os compostos.

Se for baseado puramente na sobreposição de forma, pode-se observar uma alta similaridade inter-compostos na base de dados, assim como não se observam núcleos na matriz de Tanimoto. De maneira contrária, a inclusão de características químicas (*color score*) induz as similaridades por comparação pareada para valores ainda menores (Figura 27), ou seja, os compostos são altamente diversos conforme seus grupos químicos, e provoca mudanças na matriz de Tanimoto, de uma matriz de alta inter-similaridade para uma de alta inter-diversidade. A linha vertical vermelha observada nesta figura pertence à ifosfamida, confirmando que este composto possui uma estrutura incomum, pois ela possui uma alta similaridade anormal com todos os compostos. A comparação das moléculas tanto em complementaridade química quanto em forma (campo de força combo) apresenta uma matriz de Tanimoto semelhante ao campo de força de cor. No geral, apesar desta base de dados ter sido construída com substratos seletivos, a análise da Figura 9 revela a incapacidade do ROCS de encontrar qualquer simples padrão de similaridade e a inexistência de uma correlação clara entre a seletividade e a similaridade/diversidade estrutural para esta base de dados.

Os métodos de busca por similaridade 2D (Molprint2D e Unity 2D) apresentaram resultados interessantes. Conforme a análise prévia do valor da média de Tanimoto, o Molprint2D é o método onde os compostos apresentam maior diversidade estrutural, como observado por sua similaridade por comparação pareada extremamente baixa na matriz de Tanimoto (Figura 27). Entretanto, agora é possível notar uma separação inequívoca entre os substratos da CYP2C9 e da CYP2D6. Na matriz de Tanimoto, foi observado que os substratos da CYP2D6 apresentaram uma alta intra-similaridade, como observado pelo largo núcleo do lado direito da matriz, onde o coeficiente de

Tanimoto é promovido aos maiores valores dentro do núcleo. Em adição, os substratos da CYP2C9 também demonstram uma similaridade por comparação pareada relativamente pronunciada, não como para os substratos da CYP2D6, visualizados como um pequeno núcleo no canto esquerdo superior da matriz de Tanimoto. Utilizando-se distâncias de três ligações na geração das impressões não ocorreu melhora nos resultados do Molprint2D. Entretanto, o resultado piorou, pois os núcleos, os quais estavam claros na análise prévia não são mais evidentes (Figura 27).

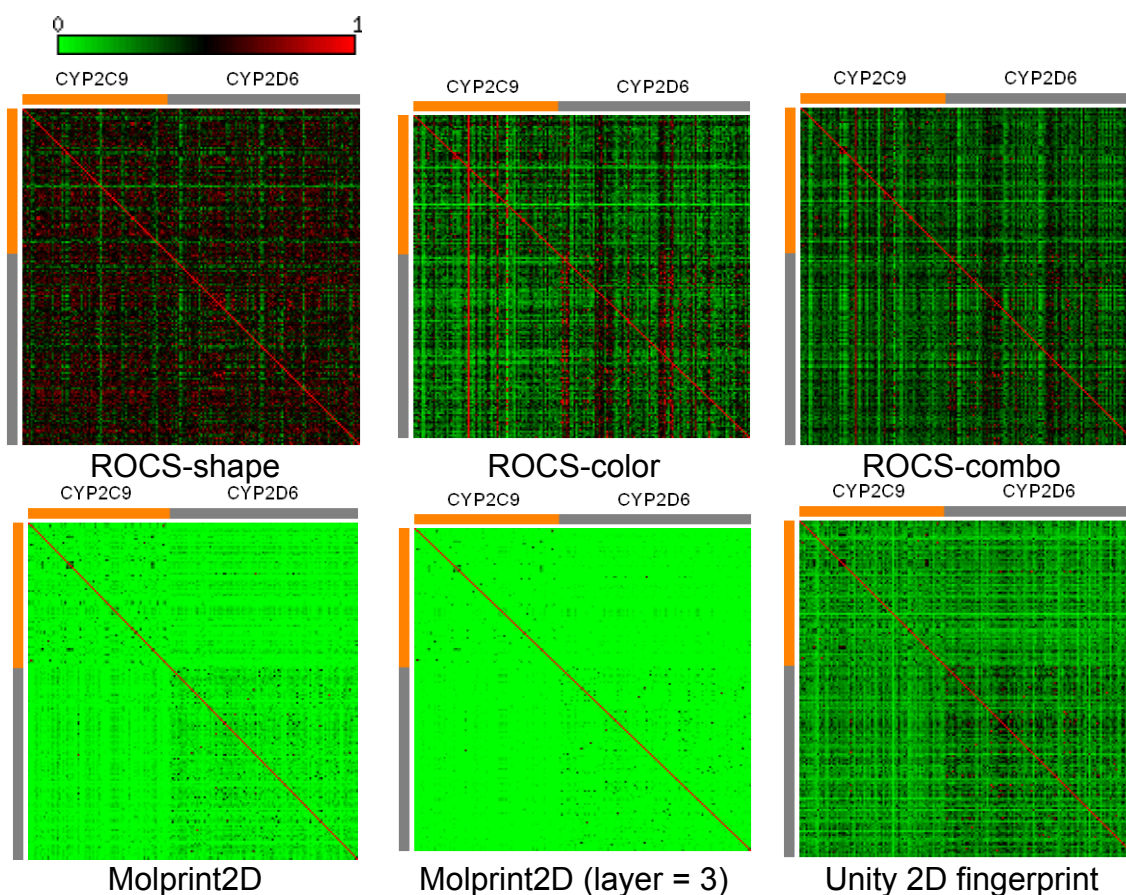


Figura 27: Matrizes de Tanimoto para as comparações duplas seletivas para a base de dados CYP2C9-CYP2D6.

Os valores de similaridade foram obtidos utilizando o coeficiente de Tanimoto (T_c). Para os valores de T_c , é utilizado um código de cor contínuo do verde ao vermelho, conforme o aumento da similaridade (preto indica $T_c = 0,5$). Em cada matriz, a posição dos compostos para um conjunto particular de substratos seletivos está indicado pelas barras de cor.

O método de impressões do Unity2D apresentou um comportamento similar ao observado pelo Molprint2D, com os substratos da CYP2D6 apresentando alta similaridade por comparação pareada. Também foi notado que todas similaridades por comparação pareada são maiores usando as impressões do Unity 2D ao invés do Molprint2D. Entretanto, não foi observado qualquer núcleo para os substratos da CYP2C9.

4.2.1.2. ANÁLISE ROC-AUC

No primeiro momento é aceitável pensar na distinção entre substratos e não substratos como sendo uma tarefa trivial, pois a base de dados foi construída apenas com compostos seletivos para uma isoforma específica. Entretanto, como já foi demonstrado, a separação entre substratos e não substratos para a base de dados CYP2C9-CYP2D6 representa um desafio para os métodos computacionais empregados neste trabalho. O desafio torna-se mais evidente pela análise da Tabela 3, onde estão os resultados da análise ROC-AUC (Área sob a curva de um gráfico ROC) para a base de dados CYP2C9-CYP2D6. Uma análise mais detalhada desta tabela revela a presença de uma grande diversidade estrutural, pois a maioria dos compostos não possui estrutura representativa o suficiente dentro do conjunto, para possibilitar uma separação entre substratos e não substratos.

Então, nem o método baseado em similaridade é capaz de realizar uma distinção melhor, em relação à seleção randômica entre substratos e não substratos para a maior parte dos compostos utilizados como referência. Isto é verdadeiro especialmente para os substratos da CYP2C9, onde apenas um método (*color score* – ROCS) apresentou um valor médio para a área sob a curva maior que 0,600.

Embora os substratos da CYP2C9 tenham apresentado uma alta diversidade química, alguns poucos compostos possuem uma estrutura de forma a permitir sua

separação entre substratos e não substratos para esta enzima. (Figura 28). Para o método do Campo de Força por Forma (ROCS-*shape*) e Molprint2D-L3 (Molprint2D com 3 camadas), os maiores valores para AUC foram obtidos utilizando-se a fenazona e o aceclofenaco, respectivamente, como estruturas de referência. No entanto, estes dois métodos são apenas ligeiramente superiores a uma seleção randômica.

Tabela 3: Parâmetros estatísticos da análise ROC-AUC para a base de dados CYP2C9-2D6.

		CYP2C9				
	n	Média	Desv Pad	Min	Máx	
ROCS- <i>shape</i>	76	0,480	0,044	0,359	0,579	
ROCS- <i>color</i>	76	0,639	0,088	0,360	0,761	
ROCS- <i>combo</i>	76	0,585	0,076	0,373	0,745	
Molprint2D	76	0,489	0,131	0,164	0,702	
Molprint2D-L3	76	0,514	0,044	0,410	0,599	
Unity 2D	76	0,480	0,097	0,263	0,698	
		CYP2D6				
	n	Média	Desv Pad	Min	Máx	
ROCS- <i>shape</i>	100	0,525	0,051	0,394	0,642	
ROCS- <i>color</i>	100	0,583	0,108	0,276	0,792	
ROCS- <i>combo</i>	100	0,583	0,079	0,303	0,798	
Molprint2D	100	0,708	0,083	0,355	0,859	
Molprint2D-L3	100	0,564	0,041	0,485	0,700	
Unity 2D	100	0,684	0,081	0,450	0,854	

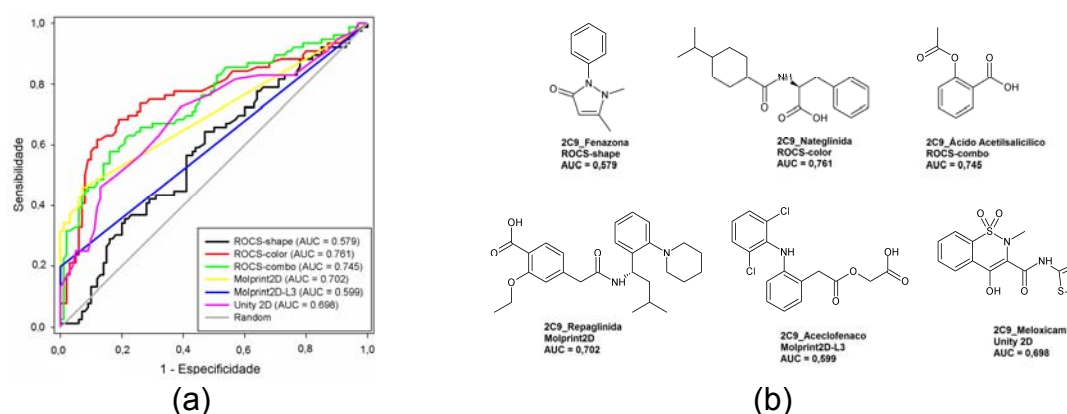


Figura 28: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridades para discriminar entre os substratos da CYP2C9 e os da CYP2D6; (b) Estruturas químicas dos substratos da CYP2C9 que proveram os melhores valores de AUC para cada método.

Todos os outros métodos (ROCS-color, ROCS-combo, Molprint2D, e Unity 2D) revelaram um valor máximo de AUC entre 0,700 e 0,760, significando a probabilidade de uma seleção randômica recuperar uma molécula ativa ser 7 vezes maior em relação a uma inativa, em um total de 10, ou seja 70%. Este é um resultado excepcional, considerando a alta diversidade química dentre os substratos presentes para a CYP2C9. Também fica claro a utilização do método baseado no ligante 3D (ROCS) ser melhor quando utiliza-se algum sinalizador químico ao invés de apenas a forma, pois a AUC foi de 0,579 (ROCS-shape) para 0,761 (ROCS-color).

Um resultado interessante foi cada método ter elegido um substrato diferente para a discriminação observada. Além disto, os maiores valores de AUC obtidos com o campo de força Color do ROCS e com o Molprint2D foram recuperados utilizando a nateglinida e a repaglinida, respectivamente. Estes dois fármacos são utilizados no tratamento do diabetes tipo 2, e pertencem à classe das meglitinidas, substâncias hipoglicêmicas. Este resultado foi uma surpresa, pois o ROCS e o Molprint2D possuem diferentes níveis de sofisticação.

As curvas ROC possuem diversas vantagens sobre os métodos gráficos mais tradicionais. Primeiro, por não dependerem do conjunto de varredura das moléculas. Segundo, provém um espectro inteiro de pares de sensibilidade e especificidade, é o único a fornecer uma imagem completa do teste de acurácia e reportar aspectos duais de qualquer teste, conhecido por habilidade de seleção de compostos ativos e descarte de inativos³¹. Entretanto, as curvas ROC possuem uma reconsideração importante: não são capazes de lidar com problemas de reconhecimento *a priori* (*early recognition problem*)^{62,63}. Por exemplo, se for considerado uma situação onde duas curvas ROC possuem a mesma AUC de exatamente 0,5, porém a primeira curva recuperou metade dos compostos ativos logo no começo da lista de ordenação por resultado e a outra metade até o final, enquanto a segunda curva recuperou todos os compostos ativos no

meio da lista, a primeira curva é nitidamente melhor para análise de problema de reconhecimento *a priori*. Como as estruturas ativas recuperadas no topo da lista de ordenação são o principal apoio para os testes de varredura virtual, isto poupa da atual dificuldade de se varrerem todos os compostos, isto é bem razoável para preferir-se um bom comportamento *a priori*^{36,64}.

Para trabalhar com o problema de reconhecimento *a priori* foi adotada a sugestão de Nicholls *et al.*⁶⁵. Foram utilizados para isto, os três métodos que apresentaram os maiores valores para a AUC (ROCS-color, ROCS-combo e Molprint2D, respectivamente) e foi estipulada uma linha de corte de 20% da base de dados para ser analisado (35 compostos). O número de substratos da CYP2C9 recuperados por estes dois métodos foi muito próximo: ROCS-color (28), ROCS-combo (29) e Molprint2D (30). Entretanto, a área de uma curva ROC traçada na linha de corte ($ROC_{20\%}$) apresentou uma inversão na precisão destes métodos: Molprint2D ($AUC_{20\%} = 0,893$), ROCS-combo ($AUC_{20\%} = 0,675$) e ROCS-color ($AUC_{20\%} = 0,510$), respectivamente. O resultado superior do Molprint2D sobre os outros dois métodos é devido ao fato de os primeiros 24 compostos recuperados pelo Molprint2D serem exclusivamente substratos da CYP2C9, enquanto o ROCS-color e o ROCS-combo apresentarem mais não-substratos distribuídos nas primeiras posições.

Uma análise feita com as dez melhores estruturas de referência segundo cada método demonstrou a nateglinida sendo selecionada por três diferentes métodos (ROCS-shape, ROCS-combo e Molprint2D-L3). O flurbiprofeno não apareceu como a melhor referência para nenhum dos métodos, mas está dentre os dez primeiros para os três métodos (ROCS-shape, ROCS-combo e Molprint2D-L3). Este resultado está de acordo com um trabalho prévio, onde o ROCS foi utilizado tendo o flurbiprofeno como estrutura de referência para predizer a orientação correta no sítio ativo para os substratos da CYP2C9⁶⁶.

Continuando com a análise da Tabela 3, é possível observar os substratos da CYP2D6 apresentando baixa diversidade química. Isto é justificado pela média da AUC para todos os métodos de desempenho melhores que uma seleção randômica (AUC > 0,500). Novamente o ROCS-shape foi quem apresentou o pior desempenho, média (AUC > 0,525) e máxima (AUC > 0,642), o qual foi obtido utilizando-se a lisurida como molécula de referência. Este resultado reforça a hipótese da comparação de compostos (ao menos na presente base de dados) baseados puramente em sua melhor sobreposição de forma não ser efetiva para distinguir entre substratos e não-substratos. A inclusão de parâmetros químicos (ROCS-color) e também a pontuação combinada (ROCS-combo) aumenta sensivelmente o desempenho dos métodos de similaridade 3D. Para ambos os métodos (color e combo), os valores para a AUC (médio e máximo) são maiores que quando usa-se apenas a forma na pontuação dos compostos (Figura 29).

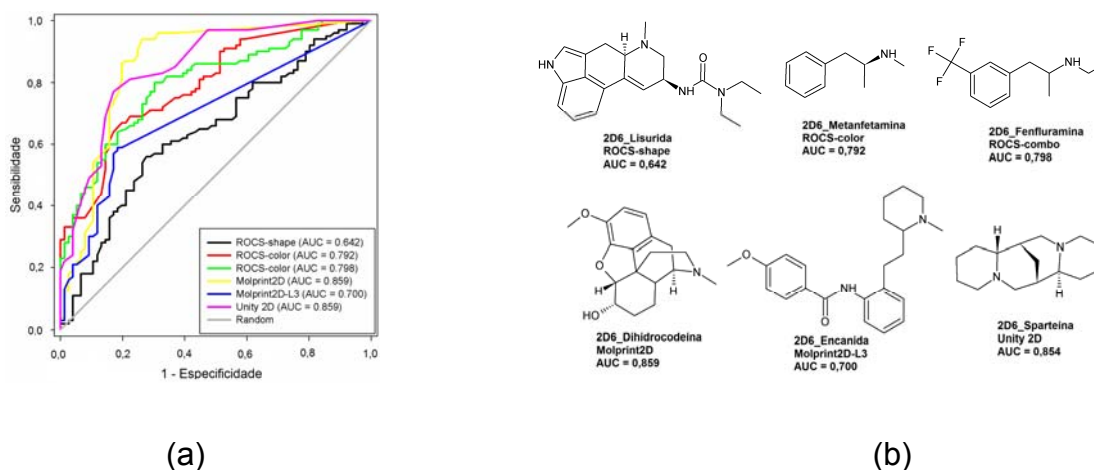


Figura 29: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridade para discriminar substratos entre a CYP2C9 e a CYP2D6; (b) Estruturas químicas dos substratos da CYP2D6 que forneceram o melhor valor de AUC para cada método.

Uma inversão de desempenho foi observada na análise dos substratos da CYP2D6. Agora os métodos de similaridade 2D (Molprint2D e Unity 2D) apresentaram melhor performance que os métodos de 3D, conforme observado pelos altos valores de AUC (ambos, média e máxima) para os métodos utilizados.

Novamente, cada método atingiu o valor máximo para AUC utilizando diferente estruturas como referência (Figura 29). Como anteriormente, o Molprint2D e o Unity 2D foram os melhores métodos para discriminar entre substratos e não-substratos para a CYP2D6, utilizando a di-hidrocodeína (AUC = 0,859) e a asparteína (AUC = 0,854) como estruturas de referência. ROCS-combo (fenfluramina; AUC = 0,798) e ROCS-color (AUC = 0,792) são os dois métodos mais próximos com alta efetividade para separar substratos de não-substratos. Analisando as dez estruturas que forneceram os maiores valores de AUC usando o Molprint2D, nota-se que há cinco fármacos da mesma classe de analgésicos opióides. Um resultado similar foi observado para o método do ROCS-combo: 3 em 10 estruturas que proveram os maiores valores de AUC são fármacos estimulantes do sistema nervoso central.

Como foi mostrado na análise de AUC -ROC para os substratos da CYP2C9, aqui também foi utilizado o método para o problema de reconhecimento *a priori*. A análise de 20% da base de dados, utilizando os três melhores resultados (Molprint2D, Unity 2D e ROCS-combo, respectivamente) surgiu um perfil similar ao encontrado previamente. O número de substratos da CYP2D6 recuperados por este método foi: Molprint2D (29), Unity 2D (32), e ROCS-combo (32), respectivamente. A área da curva ROC com a linha de corte de 20% (AUC_{20%}) mostrou um resultado oposto à curva completa: ROCS-combo (AUC_{20%} = 0,844), Unity2D (AUC_{20%} = 0,693) e Molprint2D (AUC_{20%} = 0,595), respectivamente. O melhor desempenho do ROCS-combo é devido ao fato dos 23 primeiros compostos recuperados por este método são todos substratos da CYP2D6.

4.2.2. CYP2C9 x CYP3A4

A comparação por similaridade dupla também foi realizada utilizando métodos de busca por similaridade 2D e 3D para a base de dados CYP2C9-CYP3A4. Novamente a comparação por compostos baseada exclusivamente na melhor sobreposição de forma (ROCS-shape), com os 3 melhores confôrmeros, foi onde a média de Tanimoto apresentou maiores valores (Figura 30). O resultado da inclusão de um campo de força químico para medir a complementaridade química (ROCS-color) na superposição e pontuação de similaridade não foi tão boa quanto à pontuação baseada na sobreposição de forma. A pontuação combo apresentou um valor intermediário para a média de Tanimoto entre estas pontuações.

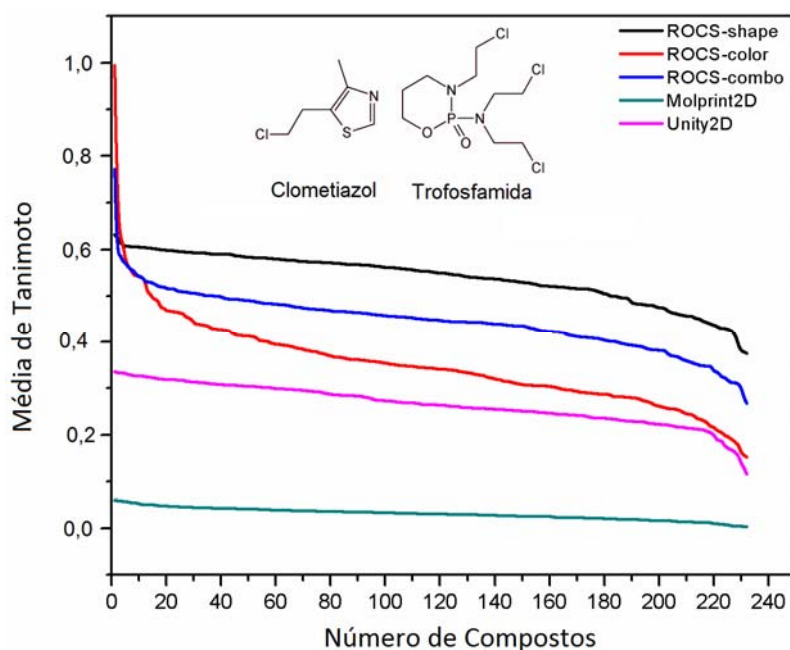


Figura 30: Média de Tanimoto para a sistemática de comparação por similaridade dupla da base de dados CYP2C9-CYP3A4. Na figura estão as estruturas 2D dos dois compostos que são atípicos à base de dados.

A trofosfamida e o clometiazol são dois substratos da CYP3A4 que se apresentaram como atípicos quando a pontuação *color* foi usada (Figura 30). Ambos apresentam uma alta média de Tanimoto: 0,994 e 0,718, respectivamente. A similaridade média para estes compostos, obtida pelo alinhamento deles sobre todos

os compostos de referência foi 0,139 e 0,175, respectivamente, indicando que nenhum composto de referência foi capaz de gerar um alinhamento significativo com estes dois substratos.

A trofosfamida é também um agente alquilante, mostarda nitrogenada, assim como a ifosfamida, e utilizada no tratamento de câncer. Logo, o alto coeficiente de Tanimoto exibido por estas estruturas com cada composto em sua base de dados pode ser associado com algum erro de parametrização do campo de força utilizado pelo ROCS.

O valor da média de Tanimoto apresentado pelas impressões do Unity2D é um pouco menor se comparada à do ROCS-color (Figura 30). Como anteriormente, o Molprint2D foi o método com a menor média de Tanimoto. O uso de uma distância superior a três ligações para gerar as impressões do Molprint2D não aumentou o resultado padrão para este método, como mostrado na análise da base de dados CYP2C9-CYP2D6. Logo, ficou decidido o uso de apenas um modo de padrão operacional do Molprint2D para o restante da análise.

4.2.2.1. DIVERSIDADE DOS COMPOSTOS

A diversidade química para a base de dados CYP2C9-CYP3A4 foi estimada utilizando a sistemática de similaridade por comparação pareada. As matrizes de Tanimoto para cada método de similaridade estão apresentadas na Figura 31. Baseado na sobreposição de forma observa-se uma elevada similaridade inter-compostos na base de dados. Isto está evidenciado pelos pontos vermelhos dispersos na matriz de Tanimoto. Após a inclusão das características químicas (ROCS-color), a similaridade por comparação pareada rebaixou os coeficientes de Tanimoto para valores ainda menores (Figura 31), indicando alta diversidade dos compostos conforme suas funcionalidades químicas. A *matriz de Tanimoto* foi convertida de uma matriz de alta

inter-similaridade para uma de alta inter-diversidade. Este resultado valida a observação sobre os valores das médias de Tanimoto, onde o ROCS-color apresenta valores menores para este parâmetro frente ao ROCS-shape e ao ROCS-combo. A linha vertical vermelha observada nesta figura corresponde à trofosfamida, evidenciando sua alta similaridade atípica com todos os compostos. A comparação das moléculas tanto na forma quanto na complementaridade química (pontuação combo) apresentou uma matriz de Tanimoto onde os coeficientes estão entorno de 0,500 (Figura 31, a cor das células varia do verde escuro ao preto). A ausência de algum nó na matriz de Tanimoto indica as pontuações utilizadas com o ROCS não sendo capazes de fornecer uma separação nítida entre os substratos das enzimas CYP2C9 e CYP3A4.

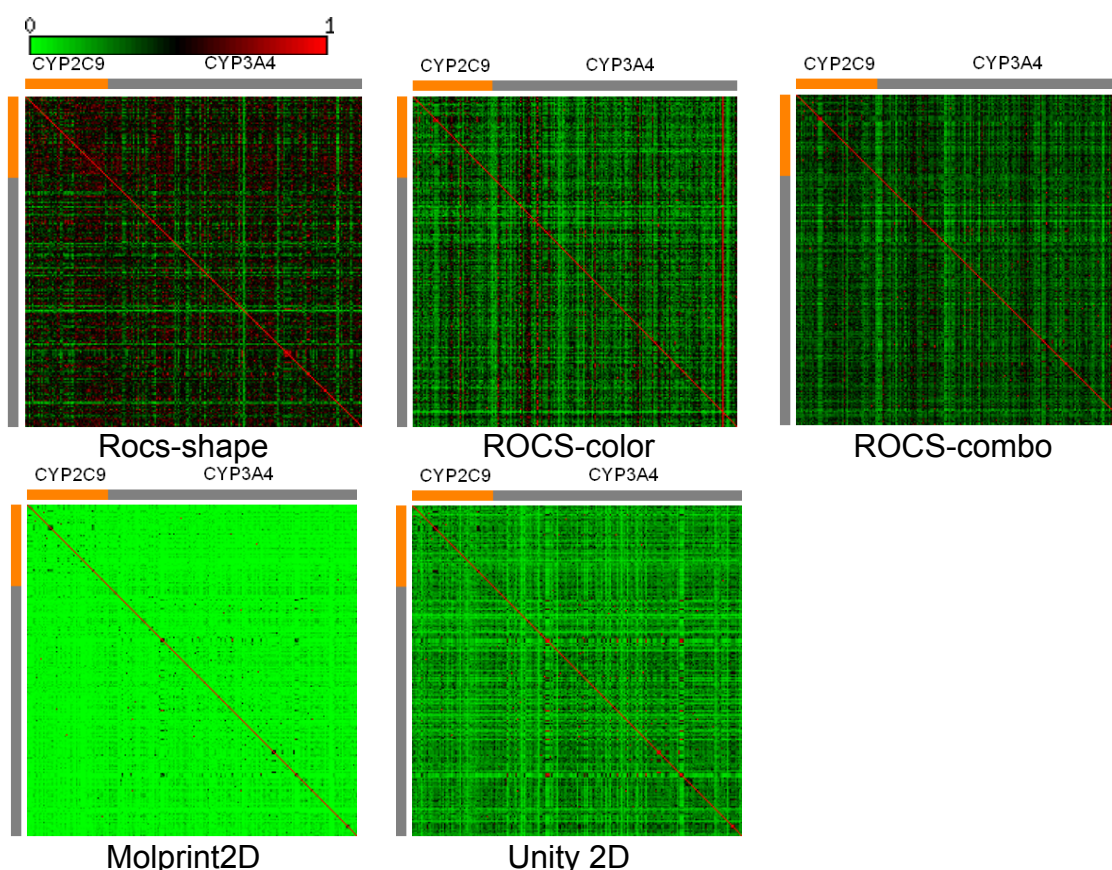


Figura 31: Matrizes de Tanimoto da comparação pareada seletiva da base de dados CYP2C9-CYP3A4.

Nesta Figura, os valores de similaridade foram obtidos utilizando o coeficiente de Tanimoto (Tc). Para a variação de Tc é utilizado um código de cores contínuas do verde ao vermelho conforme o aumento da similaridade (preto indica Tc=0,500). Em cada matriz, a posição dos compostos para um conjunto particular de substratos seletivos está marcado pelas barras de cor.

O Molprint2D é o método onde os compostos aparentam ter as mais diversas estruturas, ilustrado por sua *similaridade por comparação* pareada extremamente baixa na matriz de Tanimoto (Figura 31, células verde claro). Há também uma pequena área no lado superior esquerdo da matriz, onde os coeficientes de Tanimoto são um pouco maiores. Esta região corresponde aos substratos da CYP2C9. Para complementar, os substratos da CYP3A4 também apresentam certa similaridade por pareamento, visualizado como um grande núcleo no lado inferior direito da matriz de Tanimoto (Figura 31). O método de impressões do *Unity 2D* apresentou um comportamento análogo ao demonstrado pelo ROCS, quando nenhum núcleo é observado na matriz de Tanimoto.

4.2.2.2. ANÁLISES AUC-ROC

A Tabela 4 apresenta os resultados das análises ROC-AUC para a base de dados CYP2C9-CYP3A4. A quantidade média de substratos da CYP2C9 utilizados como estruturas de referência foi ineficiente para recuperar os substratos conhecidos nas primeiras posições da lista de pontuação. Isto é evidenciado pelos valores médios de AUC próximos a 0,500 para quase todos os métodos. Apenas o ROCS-combo e o *ROCS-shape* apresentaram uma AUC média entorno de 0,600 para a maior parte dos compostos utilizados como referência. Estes resultados confirmam a CYP2C9 como uma enzima de elevada diversidade química, portanto explicitando sua conhecida promiscuidade.

Tabela 4: Parâmetros estatísticos para a análise ROC-AUC para a base de dados CYP2C9-CYP3A4.

CYP2C9					
	n	Média	Desvio Padrão	Mín	Máx
ROCS-shape	49	0,595	0,067	0,431	0,686
ROCS-color	49	0,543	0,089	0,340	0,716
ROCS-combo	49	0,600	0,085	0,404	0,748
Molprint2D	49	0,521	0,108	0,294	0,674
Unity 2D	49	0,498	0,068	0,377	0,647
CYP3A4					
	n	Média	Desvio Padrão	Mín	Máx
ROCS-shape	182	0,435	0,061	0,270	0,579
ROCS-color	182	0,545	0,075	0,324	0,725
ROCS-combo	182	0,476	0,060	0,263	0,620
Molprint2D	182	0,588	0,083	0,359	0,718
Unity 2D	182	0,573	0,056	0,402	0,737

Algumas estruturas permitem uma discriminação relativamente boa entre substratos e não-substratos da enzima CYP2C9, apesar de sua elevada diversidade química. Entretanto, o ROCS-shape, Molprint2D, e Unity 2D realizaram esta tarefa apenas um pouco melhor frente uma seleção randômica (Figura 32). O ibuprofeno e a nateglinida demonstram-se como as estruturas de referência mais representativas para o conjunto da CYP2C9, pois os maiores valores de AUC foram obtidos com estas duas estruturas utilizando o ROCS-combo (AUC = 0,748) e o ROCS-color (AUC = 0,716), respectivamente (Figura 32). É correto dizer que o substrato nateglinida foi eficiente novamente na separação dos substratos da CYP2C9 dentre os substratos da CYP3A4. Este resultado valida a hipótese deste substrato possuir uma estrutura típica para a maioria dos substratos da CYP2C9, pois a análise prévia (CYP2C9-CYP2D6) também foi uma das provedoras do mais elevado valor de AUC.

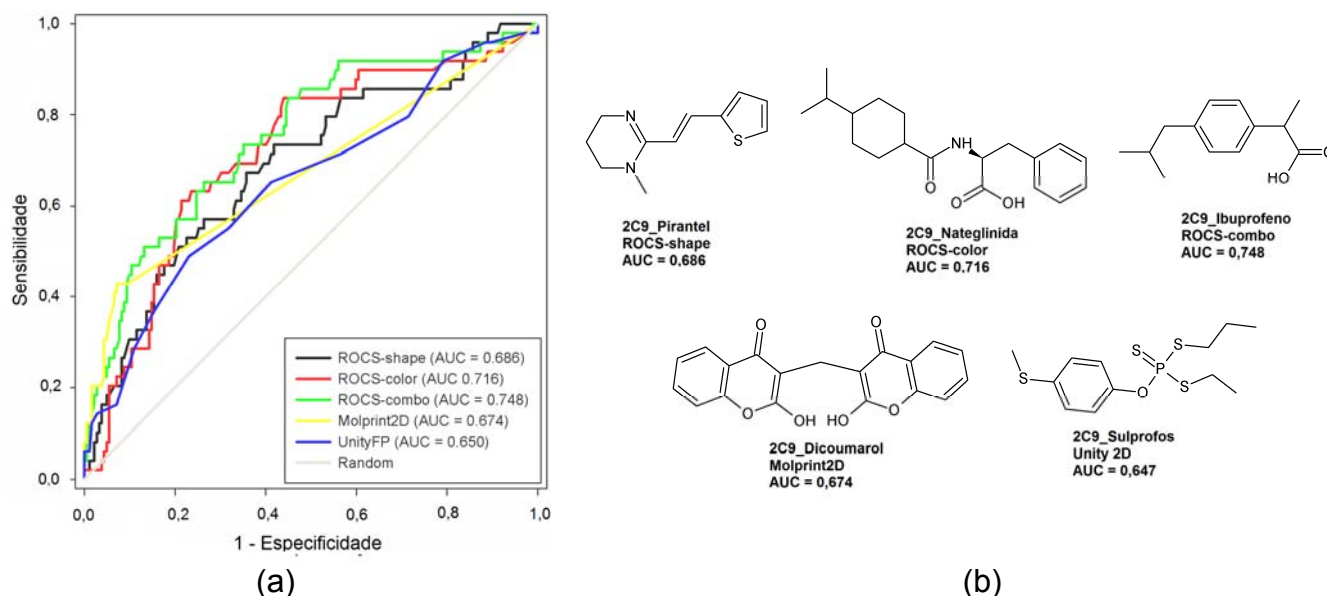


Figura 32: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridade para discriminar os substratos da CYP2C9 dos da CYP3A4; (b) Estruturas químicas dos substratos da CYP2C9 que forneceram o melhor valor de AUC para cada método.

Pelos resultados do ROCS, fica evidente o uso da pontuação referente à complementaridade química e à sobreposição de forma (ROCS-combo) como sendo superiores à pontuação ROCS-shape, a qual utiliza apenas a melhor forma de sobreposição, pois a AUC cresceu de 0,686 (ROCS-shape) para 0,746 (ROCS-combo), respectivamente.

Analisando a Figura 32, pode-se observar o ROCS-combo e o Molprint2D como sendo os métodos de maior sensibilidade com maior especificidade. Ambos mostram uma sensibilidade de 0,428 em uma especificidade de 0,907 ($1-Sp = 0,093$). Entretanto, este resultado não possui efeito pronunciado na identificação de mais substratos para a CYP2C9 no início da lista. Esta afirmação é confirmada pela curva ROC com linha de corte de 20% (46 compostos). O número de substratos recuperados por ambos os métodos neste corte é o mesmo (23 compostos). Contudo, os valores das $AUC_{20\%}$ para estes métodos são sensivelmente diferentes. A $AUC_{20\%}$ revelada por este método é de 0,698 e de 0,585 para o Molprint2D e para o ROCS-combo, respectivamente. A menor área é observada pelo método anterior, pois os substratos estão randomicamente dispersos na lista de pontuação.

A diversidade química dos substratos da CYP3A4 é comparável à dos substratos da CYP2C9. Esta observação é justificada pela análise da Tabela 4 que apresenta valores similares para a AUC média. Também neste caso, um grande número de compostos utilizados como estruturas de referência não proveram uma boa separação entre substratos e não-substratos da enzima CYP3A4, pois a AUC média ficou entorno de 0,500, independente do método utilizado.

Considerando o conjunto da CYP3A4, o ROCS-shape e o ROCS-color foram os métodos menos efetivos na discriminação entre substratos e não-substratos para esta enzima, como pode ser visto pelos baixos valores da AUC média (<0,500).

Com certeza, independente do método empregado, a AUC média para todos eles é apenas ligeiramente maior à média de uma seleção randômica, indicando uma elevada diversidade química para este conjunto de compostos.

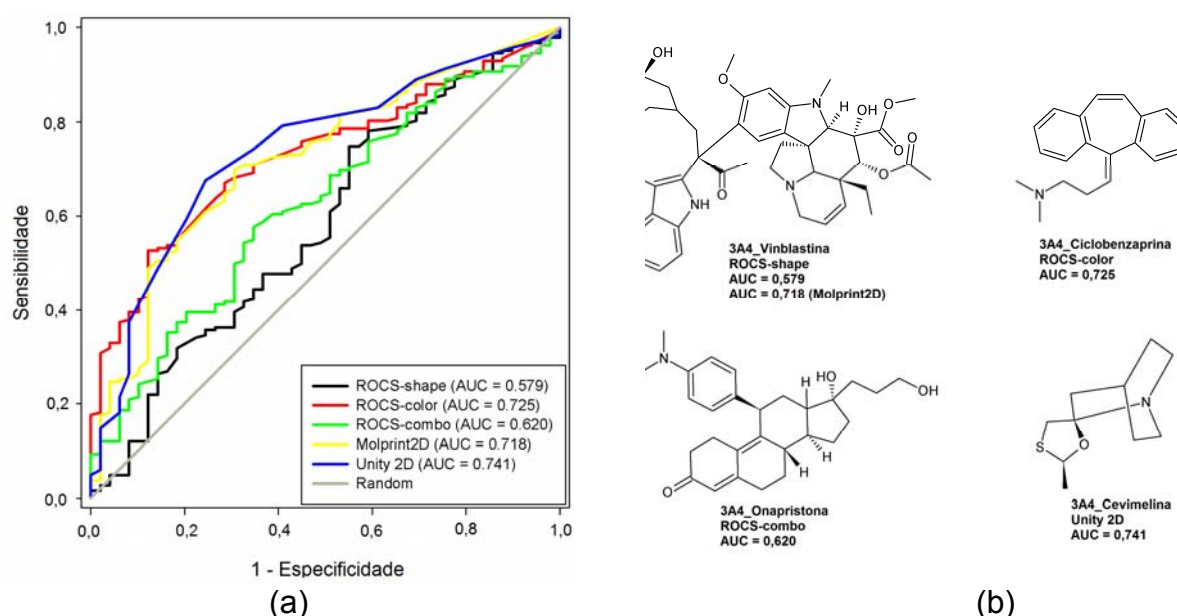


Figura 33: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridade para discriminar substratos da CYP3A4 dos da CYP2C9; (b) Estruturas químicas dos substratos da CYP3A4 que forneceram o melhor valor de AUC para cada método.

Em termos da máxima AUC apresentada por cada método, o ROCS-shape, utilizando a vinblastina como estrutura de referência, efetuou uma fraca discriminação entre os substratos dos não-substratos. A inclusão da tipologia química na pontuação

de forma (ROCS-combo) melhorou apenas um pouco os resultados, utilizando a onapristona como estrutura de referência. A vimblastina, ciclobenzaprina e cevimelina foram as estruturas que forneceram os maiores valores de AUC utilizando o Molprint2D, ROCS-color e Unity 2D, respectivamente (Figura 33).

Um aspecto a ser considerado é sobre os métodos de impressões 2D terem superado os métodos de similaridade 3D. Esta observação foi encontrada diversas vezes ao longo deste trabalho. Inúmeras publicações também mencionam várias classes de alvos, onde os métodos de similaridade 2D atingiram desempenhos consideráveis ou foram superiores aos métodos 3D^{67,68,69}. Uma possível explicação pode ser a respeito da tabela de conexão de uma molécula codificar informações implícitas sobre a estrutura de uma molécula, que são perdidas nos métodos 3D já que ignoram a topologia das ligações a favor da posição dos átomos.

Analisando os resultados da Tabela 4 e da Figura 33, a vimblastina aparenta ser favorecida pelo seu tamanho. Pois esta molécula apresenta a maior massa molecular da base de dados CYP2C9-CYP3A4. Este tipo de erro de sistema do tão empregado coeficiente de Tanimoto (Tc) é sabido há vários anos^{70,71}. Isto ocorre devido ao número de características convertidas a bits diferentes de zero em uma impressão com tamanho molecular e complexidade normalmente elevados. O máximo valor para o Tc que pode ser computado para cada molécula também é influenciado pelo seu tamanho. Assim, como o tamanho pode se referir a qualquer propriedade, como o número de átomos ou volume molecular, isto explica porque a vimblastina foi selecionada pelo ROCS-shape e pelo Molprint2D como a melhor estrutura de referência.

Por outro lado, foi uma surpresa ver uma estrutura pequena (cevimelina) ser capaz de prover a maior AUC dentre todos os substratos da CYP3A4. Uma análise das dez estruturas mais similares, conforme o Unity2D, revelou não haver uma semelhança clara entre a cevimelina e estas estruturas. A única similaridade aparente é que quase

todas estas estruturas possuem um anel heterocíclico de cinco ou seis átomos com o nitrogênio sendo seu heteroátomo. Como todas as demais estruturas possuem um tamanho maior ao da cevimelina, talvez isto possa ser outra influência onde todos os bits computados da estrutura de referência também estejam presentes nos compostos da base de dados. Então, este resultado é favorável a estar associado tanto com a complexidade molecular quanto com a similaridade entre as moléculas. Atualmente, a cevimelina junto com trofosfamida e o clometiazol são um conjunto único em uma análise de subestrutura máxima comum, reforçando a possibilidade de estas estruturas serem atípicas.

O feito do ROCS-color é nitidamente superior aos outros métodos no topo da lista. Na linha de corte de 20% (46 compostos) este método recuperou 45 compostos, com as primeiras 32 posições sendo ocupadas apenas por substratos da CYP3A4. Com a mesma linha de corte, o Molprint2D e o Unity 2D recuperaram 44 e 43 substratos da CYP3A4, respectivamente. Mesmo com o número de substratos encontrados no início da lista sendo similares dentre os métodos, a $AUC_{20\%}$ apresenta valores contrastantes: ROCS-color ($AUC_{20\%} = 0,722$), Molprint2D ($AUC_{20\%} = 0,449$), e Unity 2D ($AUC_{20\%} = 0,543$), respectivamente. Isto é fácil de ver pela análise da Figura 33, onde a linha correspondente ao ROCS-color é a com maior sensibilidade ($Se = 0.374$) e especificidade ($1-Se = 0,061$) no início da lista de pontuação.

4.2.3. CYP2D6-CYP3A4

A comparação baseada puramente na sobreposição de forma (ROCS-shape) dos compostos foi o método onde a média de Tanimoto apresentou os maiores valores para a análise de busca pela similaridade pareada (Figura 34). A inclusão de tipos químicos (ROCS-color) deteriorou os valores da média de Tanimoto, seguindo a tendência observada anteriormente nas duas bases de dados. Adicionalmente, o

ROCS-combo apresentou um valor intermediário para a média e Tanimoto entre estas pontuações.

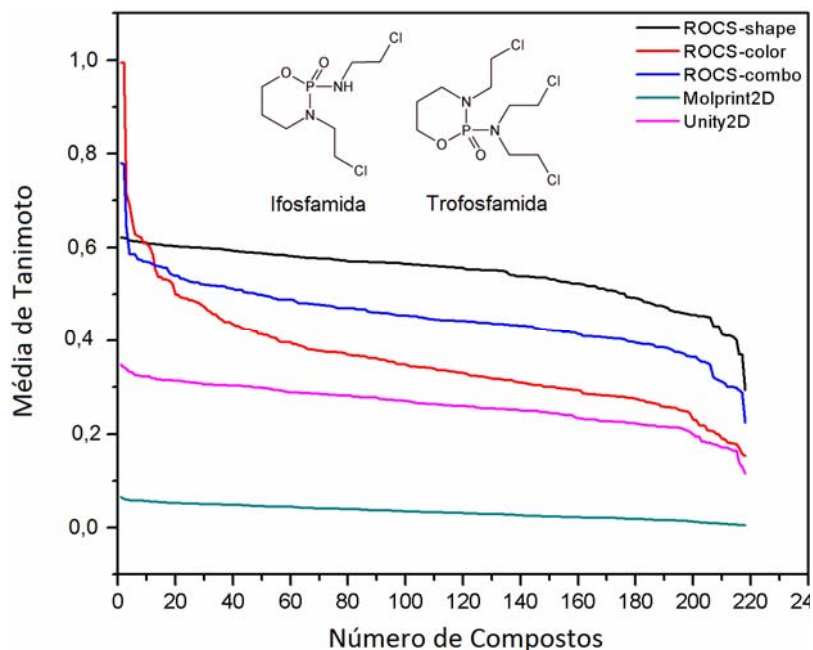


Figura 34: Valor da média de Tanimoto para a comparação sistemática de similaridade pareada para a base de dados CYP2D6-CYP3A4.

Novamente, a trofosfamida e a ifosfamida (substratos da CYP3A4) apresentaram um comportamento atípico quando o campo de força de cor foi usado. Nenhum composto de referência foi capaz de gerar um alinhamento significativo com estes dois substratos, assim como a média de similaridade para estes compostos, as quais apresentaram um valor de 0,147 e 0,148 respectivamente pelo alinhamento deles sobre todos os compostos de referencia. Os valores das médias de Tanimoto apresentados pelas impressões do Unity 2D e do ROCS-color estão presentes na Figura 34. O método que apresentou o menor valor para a média de Tanimoto foi o Molprint2D.

4.2.3.1. DIVERSIDADE DOS COMPOSTOS

Uma comparação sistemática por similaridade de pareamento foi utilizada para avaliar a diversidade química da base de dados da CYP2D6-CYP3A4. Os pontos vermelhos dispersos na matriz de Tanimoto indicam uma alta inter-similaridade dentre

os compostos da base de dados, em acordo com sua sobreposição de forma. Entretanto, isto pode ser visto em um pequeno grupamento no canto esquerdo superior da matriz. Esta região é justamente onde estão localizados os substratos da CYP2D6.

Estes resultados indicam uma pronunciada similaridade entre estes substratos. Após a inclusão da pontuação, indicativa da complementaridade química (ROCS-color), os coeficientes de Tanimoto desceram para valores menores ainda (Figura 35). Apenas algumas áreas no domínio da CYP2D6 mantiveram sua intra-similaridade elevada. Os compostos atípicos, ifosfamida e trofosfamida, estão representados pelas linhas verticais vermelhas na matriz de Tanimoto para o método ROCS-color. A comparação entre forma e a complementaridade química (ROCS-combo) resultou em uma matriz de Tanimoto com uma similaridade total de 0,500 (entre o verde escuro e o preto). Em adição a isto, o agrupamento do canto esquerdo superior pode ser visto claramente nesta matriz.

Os valores de similaridade foram obtidos utilizando-se o coeficiente de Tanimoto (T_c). Para diferentes valores de T_c foi utilizado o código de cores do verde ao vermelho, conforme o aumento da similaridade (preto indica $T_c = 0,500$). Em cada matriz, a posição dos compostos é marcada pelas barras de cor para um conjunto particular de substrato seletivo.

A análise dos métodos do Molprint2D também mostrou um agregado no canto superior esquerdo incluindo todos os substratos da CYP2D6. As áreas remanescentes da matriz apresentaram uma similaridade por comparação pareada extremamente baixa entre as estruturas, com a ausência de qualquer agregado (Figura 35). O método de impressões do Unity2D apresentou uma matriz de Tanimoto similar à vista pelo ROCS-shape, mas com os coeficientes de Tanimoto apresentando valores menores.

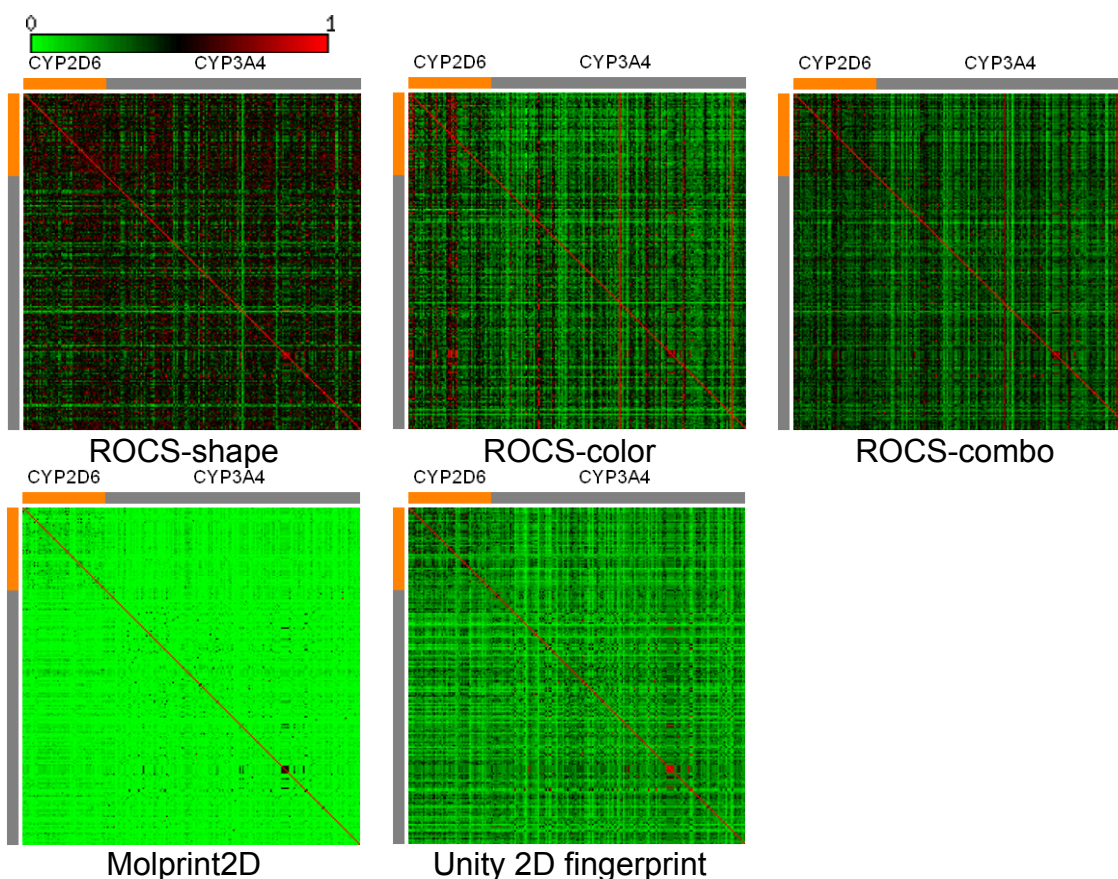


Figura 35: Matrizes de Tanimoto para a comparação seletiva por pareamento da base de dados CYP2D6-CYP3A4.

Nenhum agregado significativo foi observado na matriz de Tanimoto para os substratos da CYP3A4, com exceção de algumas pequenas áreas bem dispersas de alta intra-similaridade. Esta característica corrobora o amplo espectro de inespecificidade de substratos apresentados pela CYP3A4, que metaboliza aproximadamente 50% de todos os fármacos existentes hoje na terapêutica⁷².

4.2.3.2. ANÁLISE ROC-AUC

Nota-se pela Tabela 5 que praticamente todos os métodos foram efetivos em discriminar os substratos dos não-substratos da CYP2D6. Os valores de AUC média foram maiores que 0,500 para todos os métodos. Entretanto, o ROCS-color teve o menor desempenho: sua AUC média é apenas ligeiramente superior a uma seleção

randômica. Novamente, os métodos 2D foram os melhores em separar substratos de não-substratos para as enzimas da CYP2D6, o número médio de compostos utilizados como referência proveram uma AUC média de aproximadamente 0,700.

Tabela 5: Parâmetros estatísticos da análise ROC-AUC para a base de dados CYP2D6-CYP3A4.

CYP2D6					
	n	Média	Desvio Padrão	Min	Máx
ROCS-shape	54	0,632	0,061	0,492	0,726
ROCS-color	54	0,536	0,123	0,248	0,743
ROCS-combo	54	0,628	0,111	0,399	0,841
Molprint2D	54	0,680	0,085	0,375	0,789
Unity 2D	54	0,702	0,090	0,397	0,832
CYP3A4					
	n	Média	Desvio Padrão	Min	Máx
ROCS-shape	164	0,431	0,082	0,253	0,608
ROCS-color	164	0,632	0,093	0,322	0,819
ROCS-combo	164	0,528	0,083	0,243	0,731
Molprint2D	164	0,449	0,112	0,206	0,677
Unity 2D	164	0,507	0,137	0,164	0,776

Dentre as estruturas que discriminaram os substratos dos não-substratos para a enzima CYP2D6, que apresentaram melhor resultado estão representados em curvas ROC na Figura 36. Para todas as estruturas, os valores de AUC são maiores que 0,700. Além disso, o ROCS-combo e o Unity2D foram os métodos a apresentar os maiores valores de AUC, quando empregadas a fenformina e a metanfetamina como estruturas de referência. Os valores da AUC para estas estruturas foram de 0,841 e 0,832, respectivamente, indicando que as maiores pontuações atingiram um total de 8 em 10 frente a uma seleção randômica em selecionar moléculas ativas e inativas. Anteriormente, na análise da base de dados CYP2C9-CYP2D6, a metanfetamina também foi selecionada como estrutura de referência com o maior valor de AUC no método ROCS-color. Além disso, o pindolol, a perexilina e a fenformina apresentaram-se como farmacóforos, anteriormente mencionados, de um grande número de

substratos da CYP2D6. Outro resultado interessante foi obtido por estas estruturas (pindolol, fenformina e perexilina), o que apresentaram os maiores valores de AUC e aparecem entre as dez melhores estruturas de referência para os outros métodos. O pindolol (ROCS-shape, ROCS-combo e Molprint2D) e a metanfetamina (ROCS-color, ROCS-combo e Unity 2D) foram selecionados por três métodos cada um. Finalmente, o Molprint2D foi o único método onde a fenformina não apareceu entre as dez estruturas de referência mais eficientes.

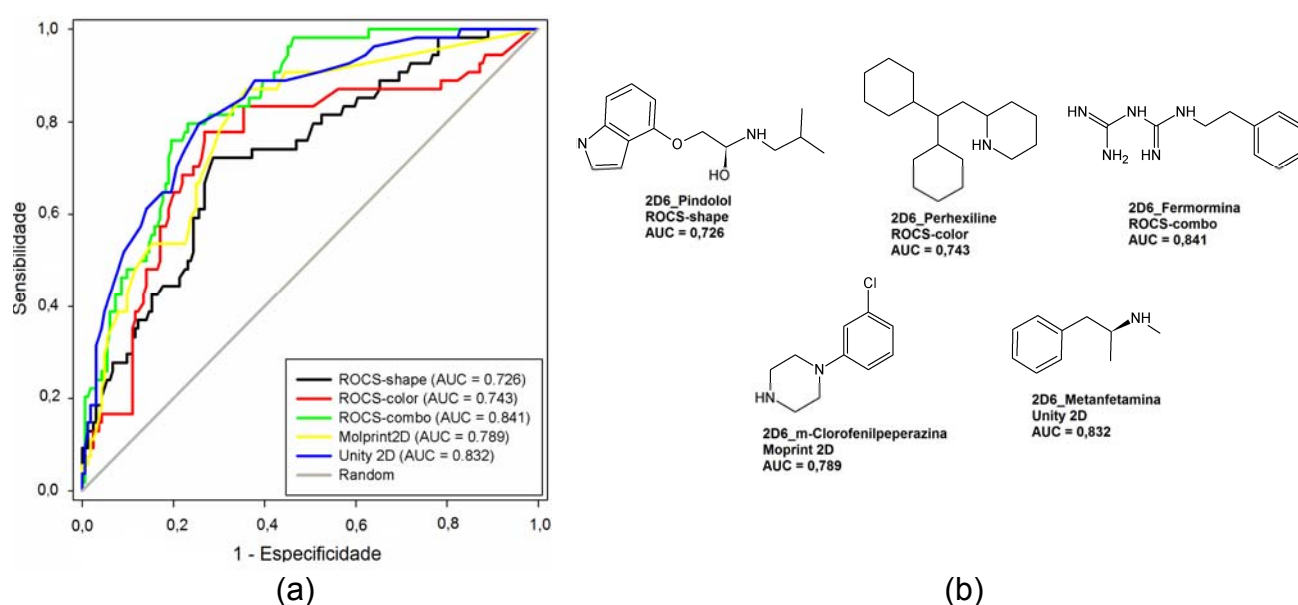


Figura 36: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridade para discriminar substratos da enzima CYP2D6 dos da CYP3A4; (b) Estruturas químicas dos substratos da CYP2D6 que proveram o melhor valor de AUC para cada método.

Com a linha de corte de 20% da base de dados (43 compostos), o número de substratos da CYP2D6 recuperados pelo ROCS-combo, Molprint2D e Unity 2D foi similar: 26, 25 e 28 substratos, respectivamente. Entretanto, nenhum método é tão efetivo para discriminar apenas substratos da CYP2D6 no início da lista. Há muitos substratos dispersos randomicamente dentre os primeiros 43 compostos, uma vez que a $AUC_{20\%}$ para estes métodos é muito pouco superior a uma seleção randômica:

ROCS-combo ($AUC_{20\%} = 0,641$), Molprint2D ($AUC_{20\%} = 0,571$) e Unity 2D ($AUC_{20\%} = 0,373$). A exceção foi no método Unity 2D, que apresentou a menor eficiência na linha de corte especificada.

Pela análise da Tabela 5, é possível observar que os substratos da CYP3A4 possuem uma alta diversidade química. Esta afirmação é sustentada pela AUC média obtida por todos os métodos, pouco melhor que uma seleção randômica ($AUC > 0,500$). O ROCS-shape e o Molprint2D foram os que pior atuaram como pode ser observado pelos seus baixos valores médios de AUC. Este resultado enfatiza positivamente a comparação dos compostos (ao menos para a presente base de dados) baseados puramente em sua melhor sobreposição de forma ineficaz na distinção entre substratos e não-substratos de uma enzima como a CYP3A4, que possui uma vasta promiscuidade de substratos. A inclusão de parâmetros químicos (ROCS-color) melhorou o desempenho do ROCS. Também, o uso do ROCS-combo proveu um valor de AUC média maior ao obtido, caso apenas o ROCS-shape fosse utilizado para pontuar os compostos. Finalmente, o Unity2D também teve um fraco desempenho, com uma AUC média de 0,507 para a média dos compostos.

A máxima AUC apresentada em cada método foi obtida utilizando o tipranavir (ROCS-shape), anastrozol (ROCS-combo), busulfan (ROCS-combo), finasterida (Molprint2D) e fluticasona (Unity 2D), respectivamente (Figura 37). Não obstante, o tipranavir e o busulfan foram selecionados aparentemente devido ao seu tamanho (métrica de Tversk). Para a primeira estrutura, como ela tem um grande volume, é mais favorável a outros compostos com menor volume, pois permite um maior alinhamento sobre ela, resultando em uma maior sobreposição. De fato, os compostos com volume próximo ao do tipranavir foram pontuados primeiro. O busulfan, por outro lado, é um composto simétrico com baixa complexidade e tamanho. A sua elevada AUC, provavelmente devido a outro erro de sistema, não há uma similaridade nítida entre o

busulfan e as dez estruturas melhor pontuadas. A única característica aparente a ser utilizada para explicar porque o busulfan foi o melhor em discriminar substratos de não-substratos da enzima CYP3A4 é a tentativa do ROCS em alinhar o grupo sulfonato deste composto com outros compostos da base de dados, que possuem dois grupos aceitadores situados à mesma distância.

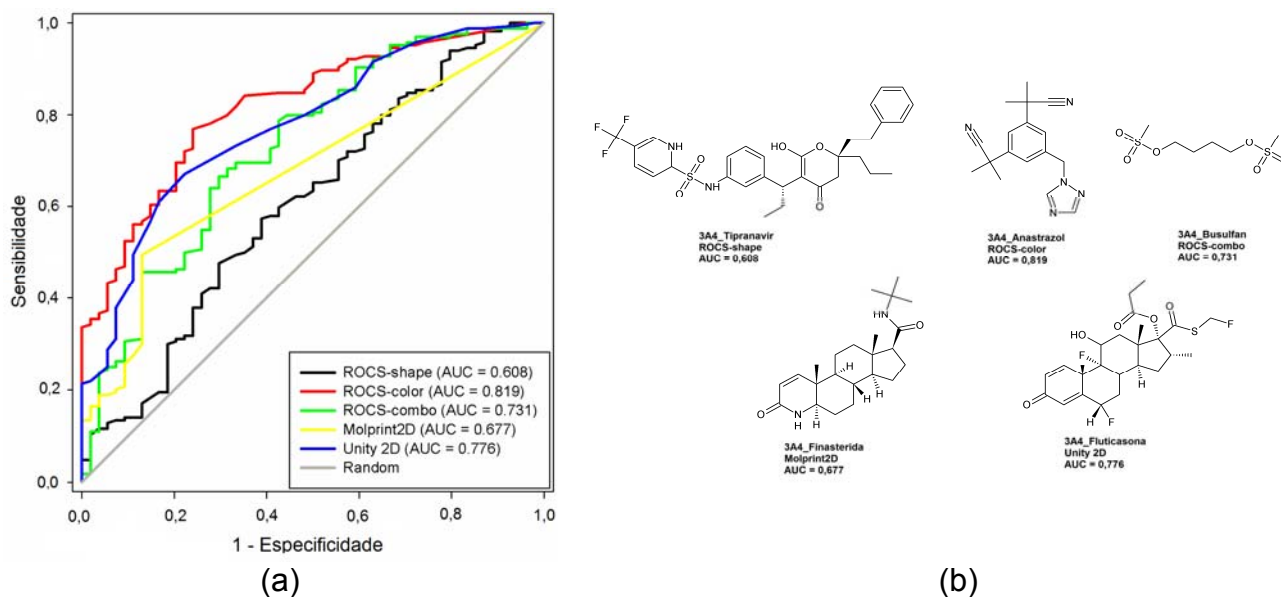


Figura 37: (a) Curvas ROC comparando o desempenho dos métodos de busca por similaridade para discriminar substratos da enzima CYP3A4 dos da CYP2D6; (b) Estruturas químicas dos substratos da CYP3A4 que proveram o melhor valor de AUC para cada método.

Para elevados valores de especificidade, podemos notar que o ROCS-color é quem provê a mais alta sensibilidade dentre todos os métodos (Figura 37). Isto pode ser confirmado pelo fato de que a curva ROC da base de dados com linha de corte de 20% (43 compostos) apresentou a área máxima ($AUC_{20\%} = 1,000$) para o método do ROCS-color. O Molprint2D ($AUC_{20\%} = 0,763$) e o Unity 2D ($AUC_{20\%} = 0,902$) foram os outros dois métodos que apresentaram um excelente resultado. O número de substratos da CYP3A4 recuperados por estes métodos foi: 43 (ROCS-color), 38 (Molprint2D) e 41 (Unity2D), respectivamente. Como pode ser observado, nas primeiras 43 posições, apenas os substratos da CYP3A4 foram identificados pelo ROCS-color.

Pela análise destes resultados, pode-se dizer que apesar de alguns métodos serem mais robustos no processo de seleção de compostos ativos, torna-se impraticável a utilização de apenas um método para distinguir inequivocamente todos os substratos para enzimas altamente promíscuas como é o caso da família do citocromo P450. Cada método conseguiu recuperar alguns compostos ativos que outros métodos dispensariam. Logo, uma boa aproximação seria unir diferentes métodos na intenção de recuperar tantos substratos quanto possíveis para enzimas tão desafiadoras como estas.

4.3. CORRELAÇÃO ENTRE A ESTRUTURA DOS SUBSTRATOS E A SELETIVIDADE DAS ISOFORMAS

Uma das mais desafiadoras características dos citocromos P450 para estudos *in silico* é sua promiscuidade. As isoformas possuem, individualmente, uma larga gama de substratos quimicamente diversos, mas é muito comum a especificidade de substratos se sobreporem dentre as enzimas CYP450, onde mais de uma enzima metaboliza o mesmo substrato. Entretanto, há diversas classes químicas metabolizadas preferencialmente por uma enzima específica, esta característica será enfocada nesta seção.

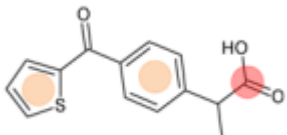
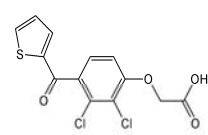
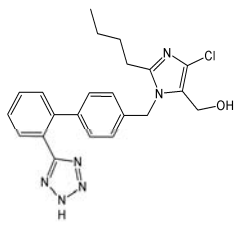
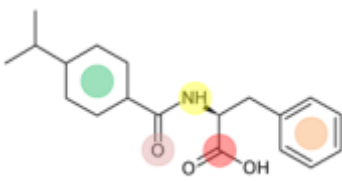
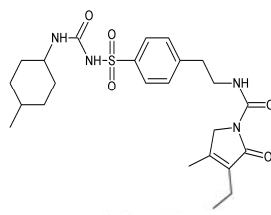
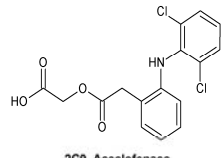
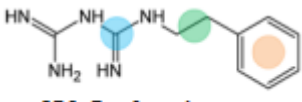
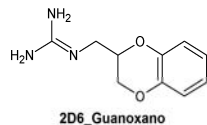
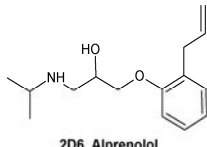
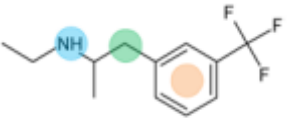
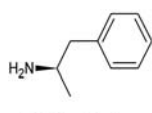
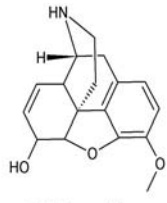
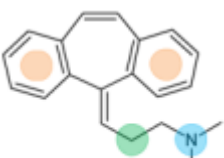
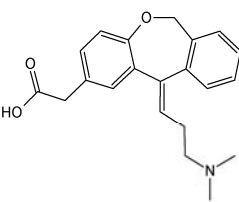
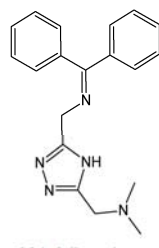
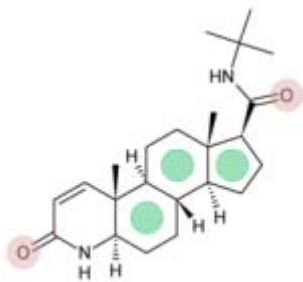
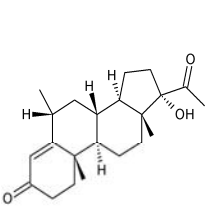
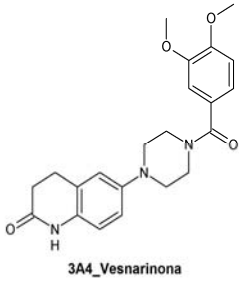
Os substratos da CYP2C9 são em sua maioria fármacos polares, encontrados em sua forma ionizada no pH fisiológico, assim como o ibuprofeno, flurbiprofeno, suprofeno etc. Atualmente, estes três substratos foram selecionados por dois ou mais métodos (três, no caso do flurbiprofeno) dentre as dez melhores estruturas de referência. Estes substratos pertencem à classe dos fármacos anti-inflamatórios não esteroidais (do inglês *Nonsteroidal Antiinflammatory Drugs*, NSAID). O suprofeno é caracterizado pela presença de dois anéis aromáticos e um grupo negativamente

carregado (Tabela 6). De fato, as primeiras vinte estruturas melhores pontuadas pelo ROCS-combo (onde 19 foram substratos da CYP2C9), foram obtidas utilizando o suprofeno como estrutura de referência. Não apenas os fármacos NSAID foram pontuados primeiro, mas também substratos de outras classes apresentaram o farmacóforo mencionado e foram recuperadas dentre as 20 primeiras estruturas.

Entretanto, o substrato suprofeno não foi selecionado como a melhor estrutura de referência por nenhum dos métodos; ele foi selecionado como segunda e terceira melhor estrutura de referência pelo ROCS-combo e ROCS-color, respectivamente, durante a análise da base de dados CYP2C9-CYP2D6. Comparando seu desempenho com a do ácido acetil salicílico (melhor estrutura de referência conforme o ROCS-combo) em 20% da base de dados, o suprofeno demonstra ser melhor em discriminar os substratos dos não substratos para a CYP2C9 no início da base de dados. A $AUC_{20\%}$ para estes substratos, utilizando o ROCS-combo é de: 0,767 (suprofeno) e 0,675 (ácido acetil salicílico), respectivamente. De fato, este resultado está de acordo com um trabalho anterior, onde o flurbiprofeno foi utilizado como estrutura de referência pelo ROCS e foi capaz de predizer eficientemente a correta orientação do sítio ativo dos substratos da CYP2C9.

Outro substrato selecionado da CYP2C9 foi a nateglinida que foi eleita por diversos métodos como uma das dez melhores estruturas de referência. Sua estrutura é caracterizada pela presença de um anel aromático, um grupo negativamente carregado, um grupo aceitador, um grupo doador e um grupo hidrofóbico (Tabela 6). Para as 20 primeiras estruturas, 17 são substratos da CYP2C9 (pontuados pelo método ROCS-color) e elas satisfazem, ao menos, três dos cinco pontos farmacofóricos da nateglinida. Do total da base de dados, aproximadamente 38 substratos da CYP2C9 apresentaram ao menos três grupos farmacofóricos dos cinco apresentados pela nateglinida.

Tabela 6: Pontos farmacofóricos da estrutura de referência que proveram a melhor separação entre substratos e não-substratos para as três CYP450 estudadas aqui. Alguns exemplos também são mostrados abaixo.

Enzima	Composto de Referência	Exemplos		
CYP2C9	 2C9_Suprofeno	 2C9_Ácido tienílico	 2C9_Losartana	
	 2C9_Nateglinida	 2C9_Glimepirida	 2C9_Aceclofenaco	
	CYP2D6	 2D6_Fenformina	 2D6_Guanoxano	 2D6_Alprenolol
		 2D6_Fenfluramina	 2D6_Anfetamina	 2D6_Norcodeína
CYP3A4		 3A4_Ciclobenzaprina	 3A4_Olopatadina	 3A4_Adinazolam
		 3A4_Finasterida	 3A4_Medroxiprogesterona	 3A4_Vesnarinona

* Cada característica química do composto de referência está representada por uma esfera de cor: anel aromático (laranja claro), grupo hidrofóbico (verde claro), grupo carregado negativamente (vermelho claro), grupo carregado positivamente (azul claro) e grupo doador (amarelo claro).

As estruturas de referência mais representativas da CYP2D6 são: fenfluramina e fenformina. O primeiro composto foi selecionado pelo ROCS-combo na análise da base de dados CYP2C9-CYP2D6, esta molécula é um inibidor de recaptção de serotonina. Por outro lado, a fenformina é um fármaco utilizado no tratamento da diabetes e pertence à classe biguanida, foi selecionada pelo ROCS-combo na análise da base de dados da CYP2D6-CYP3A4. Ambas as estruturas possuem um padrão farmacofórico muito similar, com um anel aromático, um grupo hidrofóbico e um grupo positivamente carregado (Tabela 6). Os resultados estão condizentes com trabalhos anteriores que apresentaram como característica usual da maioria dos substratos da CYP2D6 a presença de átomos de nitrogênio com caráter básico e um anel aromático^{73,74} (Tabela 6). Atualmente, 59 substratos da CYP2D6 possuem os mesmos três pontos farmacofóricos como revelados pela fenformina e pela fenfluramina. Em complemento a isto, 5 das 6 melhores estruturas de referência da análise da CYP2C9-CYP2D6 possuem o padrão estrutural mencionado. É interessante notar que mesmo compostos rígidos como a norcodeína que compartilha este simples farmacóforo que está associado com a especificidade de substrato apresentada pela enzima CYP2D6.

A CYP3A4 possui uma grande variedade de especificidade de substrato e metaboliza próximo a 50% dos fármacos presentes hoje na terapêutica. Esta enzima é capaz de metabolizar moléculas altamente lipofílicas de uma grande variedade de diversidades químicas. Apesar da alta diversidade dos substratos da CYP3A4, os resultados fornecidos pelo ROCS-color (na análise da CYP2C9-CYP3A4) para o alinhamento com as moléculas da base de dados com a ciclobenzaprina como estrutura de referência merece uma análise adicional. Uma inspeção mais apurada nesta molécula, que é um anti-depressivo tricíclico, revelou a presença de quatro pontos farmacofóricos: dois anéis aromáticos, uma cadeia carbônica hidrofóbica e um grupo positivamente carregado (Tabela 6). Analisando a lista de pontuação do ROCS-

color, utilizando a ciclobenzaprina como estrutura de referência, notam-se diversas estruturas primeiramente pontuadas, possuindo ao menos três dos pontos farmacofóricos apresentados por este composto. Um total de 44 dos 45 primeiros compostos (20% da base de dados) são substratos da CYP3A4, e todos apresentam ao menos três dos pontos farmacofóricos demonstrados pela ciclobenzaprina. Também foram encontrados 111 substratos da CYP3A4 com três ou mais pontos farmacofóricos apresentados pela ciclobenzaprina. Este é um resultado excepcional, pois representa aproximadamente 48% de todos os substratos da CYP3A4 na base de dados CYP2C9-CYP3A4. Embora algumas destas estruturas possuam uma similaridade visualmente clara com a ciclobenzaprina, muitos deles possuem uma estruturas visualmente distinta da estrutura de referência (Tabela 6), e a similaridade destas estruturas com a ciclobenzaprina está intimamente correlacionada com a presença dos grupos farmacofóricos mencionados anteriormente.

Na análise da CYP2D6-CYP3A4, foram encontradas como estruturas de referência pelos métodos 2D (Molprint2D e Unity 2D) a finasterida e a flutizasona. Ambas possuem alta similaridade estrutural. Analisando suas estruturas, foram observados cinco pontos farmacofóricos: dois grupos aceitadores e três grupos hidrofóbicos (Tabela 6). Apenas, 26 dos 164 substratos da CYP3A4 apresentam o esqueleto de esteróide lipofílico, caracterizado por um lipídio terpenóidico com um esqueleto carbônico de quatro anéis fundidos, geralmente arranjados na forma 6-6-6-5. Uma parte significativa (52 compostos) dos substratos da CYP3A4 revelou ao menos três dos cinco pontos farmacofóricos da finasterida, incluindo estruturas como a vesnarinona, as quais são muito distintas do esqueleto dos esteróides.

As enzimas da família do citocromo P450 são capazes de metabolizar substratos de classes químicas altamente distintas. Isto dificulta os estudos preditivos sobre a especificidade de substratos para estas enzimas. Entretanto, analisando os resultados

discutidos nesta seção, nota-se que mesmo para enzimas altamente promíscuas como as CYP 450 existem algumas estruturas de referência com representatividade suficiente para todos os substratos de uma enzima específica. Em complemento, o uso de diferentes métodos com múltiplas estruturas de referência, tornou-se uma boa estratégia para aumentar a discriminação entre substratos e não substratos das enzimas CYP.

4.4. SOBREPOSIÇÕES DO ROCS

Uma análise adicional deste trabalho foi a avaliação dos alinhamentos realizados pelo método ROCS-combo para substratos no sítio metabólico das CYPs. Para executar esta tarefa, selecionou-se a melhor estrutura de referência de uma enzima específica, e foi analisado o alinhamento dos 20 primeiros substratos pontuados sobre esta estrutura e identificado o sítio primário de oxidação para cada molécula.

O suprofeno foi escolhido como estrutura de referência mais apropriada para analisar o alinhamento dos substratos da CYP2C9. Este substrato experimenta uma hidroxilação no anel tiofênico. A figura abaixo mostra o alinhamento dos 10 primeiros substratos da CYP2C9 pontuados sobre o suprofeno. Todas estas moléculas apresentam seu núcleo de metabolismo contido a uma distância máxima de 4,0 Å do núcleo de metabolismo do suprofeno. Os substratos losartana e indometacina são exceções já que seus núcleos encontram-se a 5,7 e 8,6 Å de distância do núcleo de metabolismo do suprofeno, respectivamente (Figura 38). A evidência mais significativa da capacidade do suprofeno alinhar corretamente os substratos da CYP2C9 vem do fato que dos 20 primeiros substratos pontuados, 16 apresentaram o núcleo de metabolismo dentro de 4,0 Å de distância do núcleo de metabolismo do suprofeno.

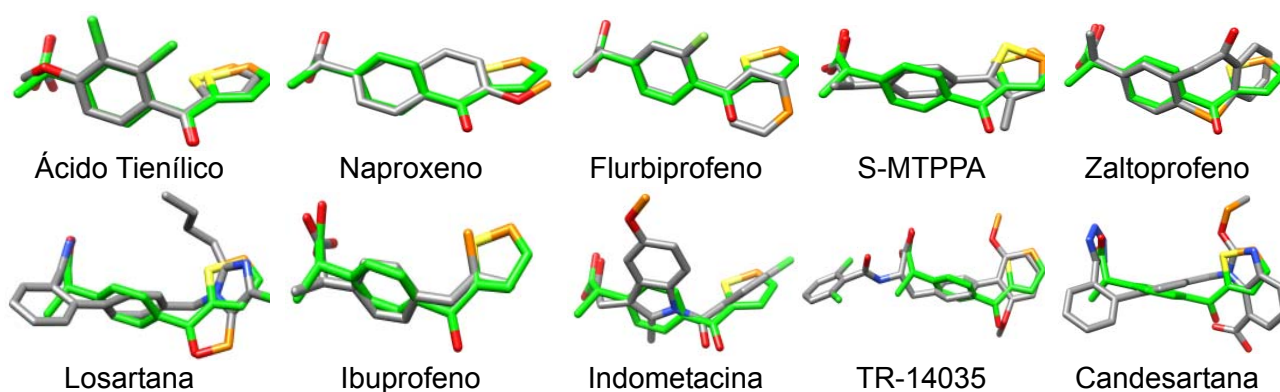


Figura 38: Sobreposição dos 10 primeiros substratos da CYP2C9 com o suprofen, conforme o método do ROCS-combo. Os sítios de metabolismo estão representados em laranja. Os átomos de carbono do suprofen estão coloridos em verde.

A estrutura de referência selecionada para analisar os resultados de sobreposição do ROCS para os substratos da CYP2D6 foi a fenformina. O metabolismo deste substrato é caracterizado por uma 4-hidroilação de seu anel aromático. A sobreposição da fenformina com os 10 primeiros substratos pontuados da CYP2D6 está representada na figura a baixo. Para todas essas sobreposições, a distância entre o núcleo metabólico da estrutura de referência e a molécula alinhada é inferior a 2,0 Å. Houve apenas duas exceções: a debrisoquina (4,6 Å) e o carteolol (3,8 Å). De fato, 16 dos 20 substratos melhores pontuados da CYP2D6 estão a uma distância inferior a 2,0 Å e 8 substratos, metade do conjunto, apresenta seu núcleo de metabolismo exatamente alinhado com o composto de referência.

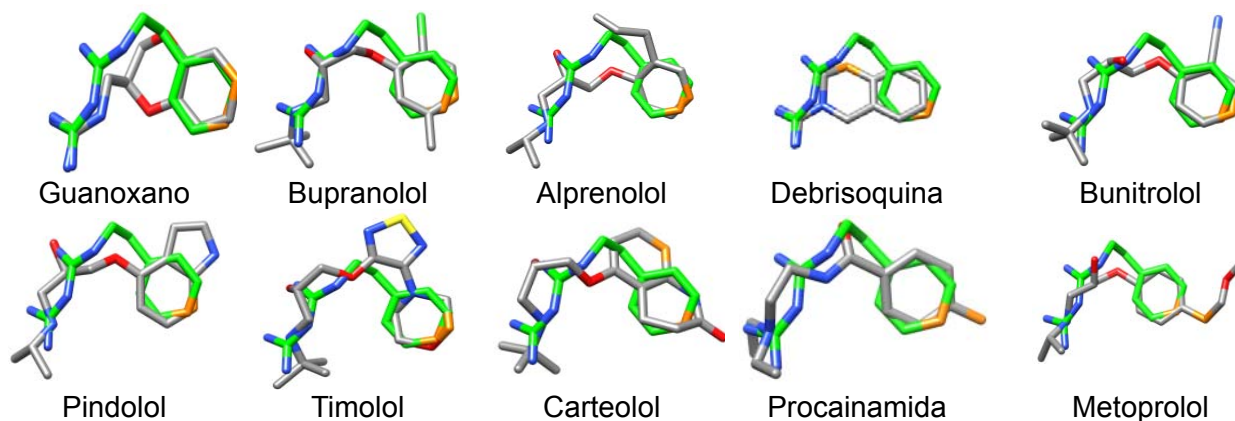


Figura 39: Sobreposição dos 10 primeiros substratos da CYP2D6 com a fenformina, conforme o método do ROCS-combo. Os sítios de metabolismo estão representados em laranja. Os átomos de carbono da fenformina estão coloridos em verde.

Para a análise de alinhamento dos substratos da CYP3A4, a ciclobenzaprina foi selecionada como estrutura de referência. Diferentemente das estruturas anteriores, que experimentam hidroxilação em seu núcleo de metabolismo, a CYP3A4 catalisa a N-desalquilação da ciclobenzaprina. A figura abaixo mostra que para os 10 primeiros substratos pontuados da CYP3A4, oito também passam por uma reação de N-desalquilação. Praticamente todos os 20 melhores substratos para a CYP3A4 são metabolizados seguindo a mesma reação. Apenas seis deles (reboxetina, primaquina, emedastina, doxorubicina, exemestano, galantamina) passam por reação diferente à N-desalquilação por metabolismo. Onze substratos possuem seu núcleo metabólico na posição exata da ciclobenzaprina. Para 13 dos 20 melhores substratos pontuados da CYP3A4, a distância entre seu núcleo de metabolismo e o da ciclobenzaprina é inferior a 2,0 Å.

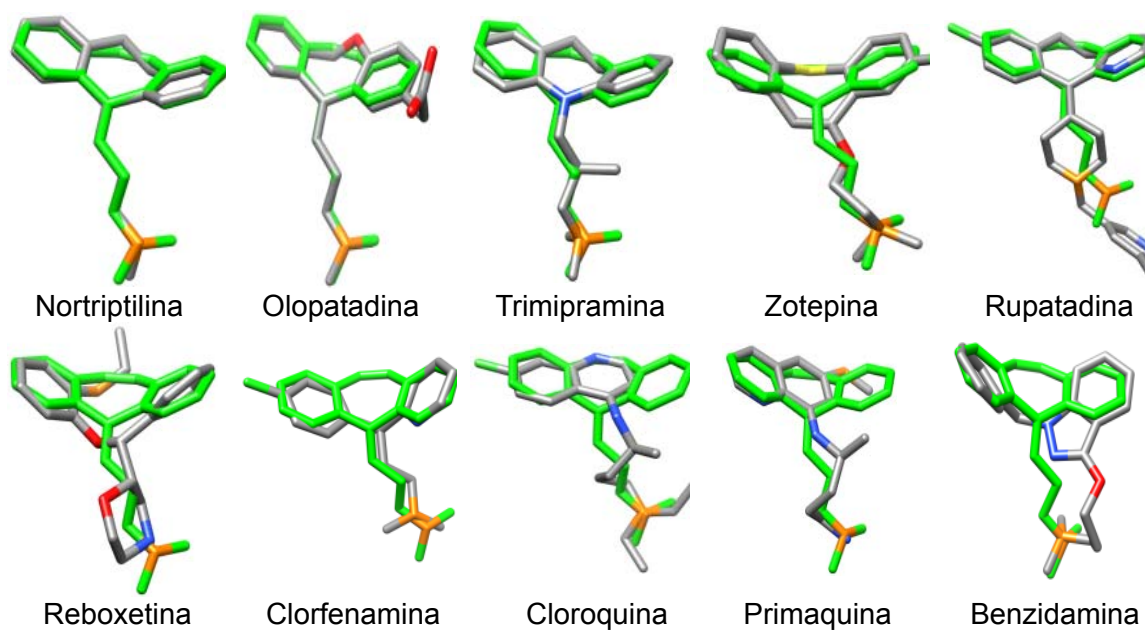


Figura 40: Sobreposição dos 10 primeiros substratos da CYP3A4 com a ciclobenzaprina, conforme o método do ROCS-combo. Os sítios de metabolismo estão representados em laranja. Os átomos de carbono da ciclobenzaprina estão coloridos em verde.

Vale a pena ressaltar que até mesmo estruturas que são aparentemente distintas das estruturas de referência utilizadas para alinhar moléculas com substratos de CYP foram corretamente alinhadas com o ROCS, e o mais notável é o núcleo de metabolismo delas estarem muito próximos (em alguns casos na mesma posição). No geral, estes resultados indicam o uso do ROCS junto com uma seleção apropriada da estrutura de referência, ser de valiosa importância não somente na identificação de substratos de uma enzima específica, mas também no reconhecimento no núcleo de metabolismo destes compostos.

5. CONCLUSÃO

Historicamente, os métodos de busca por similaridade são aplicados para encontrar compostos similares à estrutura de referência, ativa biologicamente. Logo, é esperado que uma molécula de uma base de dados também seja ativa; isto é, caso seja estruturalmente similar à estrutura de referência.

Entretanto, a seletividade junto à potência são duas propriedades essenciais que o composto deve apresentar e, ambas (seletividade e potência) precisam ser otimizadas nos estágios iniciais do processo de descoberta de fármacos para reduzir a taxa de falhas nos estágios finais.

As enzimas CYP1A2, CYP2C9, CYP2C19, CYP2D6 e CYP3A4 são responsáveis pelo metabolismo de mais de 90% de todos os fármacos presentes na terapêutica. Estas enzimas representam um desafio aos estudos computacionais que visam a identificar as características químicas que conferem seletividade ao substrato, pois, elas são altamente promíscuas e também podem apresentar sobreposição com o substrato.

Neste trabalho, o uso de métodos de busca por similaridade foi expandido, demonstrando a possibilidade de serem utilizados com sucesso na distinção entre substratos e não-substratos para estas enzimas.

Uma análise minuciosa foi desenvolvida para avaliar a habilidade dos métodos de busca por similaridade 2D e 3D em identificar substratos de não-substratos para as CYPs 2C9, 2D6 e 3A4.

Os resultados da análise por comparação pareada sugerem um grande intervalo de diversidade química para os substratos destas enzimas, independente do método utilizado, aumentando na seguinte ordem: CYP2D6 < CYP2C9 < CYP3A4.

Seguramente, os resultados da análise da AUC-ROC demonstraram que mesmo para enzimas altamente promíscuas como as CYP450, os métodos de similaridade

empregados neste trabalho forneceram estruturas de referência representativas o suficiente para todos os substratos de uma enzima específica, sendo que, os melhores apresentaram valores de AUC de 0,737 a 0,859.

Para concluir, embora alguns métodos possam ser mais robustos em recuperar substratos de uma isoforma, praticamente, torna-se inviável distinguir todos os substratos de enzimas tão promíscuas como o citocromo P 450, inequivocamente, empregando-se apenas um único método.

O uso de diferentes métodos de busca por similaridade com múltiplas estruturas de referência provê uma poderosa estratégia para aperfeiçoar a discriminação entre substratos e não substratos das enzimas CYP450.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1

NIGSCH, F.; MITCHELL, J. B. O. How to winnow actives from inactives: introducing molecular orthogonal sparse bigrams (MOSBs) and multiclass winnow. **Journal of Chemical Information and Modeling**, v. 48, n. 2, p. 306-318, 2008.

2

SHERIDAN, R. P.; KEARSLEY, S. K. Why do we need so many chemical similarity search methods? **Drug Discovery Today**, v. 7, p. 903-911, 2002.

3

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

4

VAN DE WATERBEEMD, H.; GIFFORD, E. ADMET in silico modelling: towards prediction paradise? **Nature Reviews Drug Discovery**, v. 2, 192-204, 2003,

5

COPELAND, R. A. **Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists**, Hoboken, U.S.A.: John Wiley & Sons, Inc, 2006 v. 46, c. 3, p. 48-80.

6

VOGT, I.; BAJORATH, J. Design and exploration of target-selective chemical space representations. **Journal of Chemical Information and Modeling**, v. 48, p. 1389-1395, 2008.

7

VOGT, I.; AHMED, H. E. A.; AUER, J.; BAJORATH, J. Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. **Molecular Diversity**, v. 12, n. 1, p. 25-40, 2008.

8

STUMPFE, D.; AHMED, H. E. A.; VOGT, I.; BAJORATH, J. Methods for computer-aided chemical biology. part 1: design of a benchmark system for the evaluation of compound selectivity. **Chemical Biology & Drug Design**, v. 70, n. 3, p. 182-194, 2007.

9

BAJORATH, J. Computational analysis of ligand relationships within target families. **Current Opinion in Chemical Biology**, v. 12, n. 3, p. 352-358, 2008.

10

KOEHN, F. E.; CARTER, G. T. The evolving role of natural products in drug discovery. **Nature Reviews Drug Discovery**, v. 4, p. 206, 2004.

11

BAJORATH, J. Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. **Journal of Computer-Aided Molecular Design**, v. 16, p. 431, 2002.

12

TIETZE, L. F.; BELL, H. P.; CHANDRASEKHAR, S. Natural product hybrids as new leads for drug discovery. **Angewandte Chemie International Edition**, v. 42, p. 3996, 2003.

13

OPREA, T. I.; MANNHOLD, R.; KUBINYI, H.; FOLKERS, G. "**Chemoinformatics in Drug Discovery (Methods and Principles in Medicinal Chemistry)**", John Wiley & Sons, 2005.

14

MILLAR, M. A. Chemical database techniques in drug discovery. **Nature Reviews Drug Discovery**, v. 1, p. 220, 2002.

15

XU, J.; HAGLER, A. Chemoinformatics and drug discovery. **Molecules**, v. 7 p. 566, 2002.

16

KERNS, E. H.; DI, L. Pharmaceutical profiling in drug discovery. **Drug Discovery Today**, v. 8, p. 316, 2003.

17

LOMBARDINO, J. G.; LOWE III, J. A. The role of the medicinal chemist in drug discovery — then and now. **Nature Reviews Drug Discovery**, v. 3, p. 853, 2004.

18

MERLOT, C.; DOMINE, D.; CLEVA, C.; CHURCH, D. J. Chemical substructures in drug discovery. **Drug Discovery Today**, v. 8, p. 594, 2003.

19

TESTA, B.; BALMAT, A. L.; LONG A.; JUDSON, F. Predicting drug metabolism – an evaluation of the expert system meteor. **Chemistry & Biodiversity**, v. 2, p. 872, 2005.

20

SUSNOW, R. G.; DIXON, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. **Journal of Chemical Information and Computer Sciences**, v. 43, p. 1308, 2003.

21

TILLEMENT, J. P.; TREMBLAY, D. **Comprehensive Medicinal Chemistry II: ADME-Tox Approaches**, Oxford, U.K.; Elsevier, 2006, v. 5, c. 2, p. 11-30.

22

MOULY, S.; MEUNE, C.; BERGMANN, J. F. Mini-series: I. Basic science. Uncertainty and inaccuracy of predicting CYP-mediated in vivo drug interactions in the ICU from in vitro models: focus on CYP3A4. **Intensive Care Medicine**, v. 35, p. 417-429, 2009.

23

ISIN, E. M.; GUENGERICH, F. P. Multiple sequential steps involved in the binding of inhibitors to cytochrome P450 3A4. **Journal of Biological Chemistry**, v. 9, p. 6863, 2007.

24

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. CYP2D6 cytochrome P450, family 2, subfamily D, polypeptide 6. Disponível em:
<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Retrieve&dopt=full_report&list_uids=1565>. Acessado em: 18 abr. 2008.

25

WIENKERS, L. C.; HEATH, T. G. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. **Nature Reviews Drug Discovery**, v. 4, p. 825-833, 2005.

26

DJORDJEVIC, N.; GHOTBI, R.; BERTILSSON, L.; JANKOVIC, S.; AKLILLU, E. Induction of CYP1A2 by heavy coffee consumption in serbs and swedes. **European Journal of Clinical Pharmacology**, v. 64, p. 381-385, 2008.

27

SYKES, M. J.; MCKINNON, R. A.; MINERS, J. O. Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. **Journal of Medicinal Chemistry**, v. 51, p. 780-791, 2008.

28

LIMA, M. V.; RIBEIRO, G. S.; MESQUITA, E. T.; VICTER, P. R.; VIANNA-JORGE, R. CYP2C9 genotypes and the quality of anticoagulation control with warfarin therapy among brazilian patients. **European Journal of Clinical Pharmacology**, v. 64, p. 9-15, 2008.

29

HUNFELD, N. G.; MATHOT, R. A.; TOUW, D. J.; VAN SCHAİK, R. H.; MULDER, P. G.; FRANCK, P. F.; KUIPERS, E. J.; GEUS, W. P. Effect of CYP2C19 *2 and *17 mutations on pharmacodynamics and kinetics of proton pump inhibitors in caucasians. **British Journal of Clinical Pharmacology**, v. 65, p. 752-760, 2008.

30

HIRT, D.; MENTRÉ, F.; TRAN, A.; REY, E.; AULELEY, S.; SALMON, D.; DUVAL, X.; TRÉLUYER, J. M. Effect of CYP2C19 Polymorphism on nelfinavir to M8 biotransformation in HIV patients. **British Journal of Clinical Pharmacology**, v. 65, p. 548-557, 2008.

31

BUENO, L.; DE PONTI, F.; FRIED, M.; KULLAK-UBLICK, G. A.; KWIATEK, M. A.; POHL, D.; QUIGLEY, E. M. M.; TACK, J.; TALLEY, N. J. Serotonergic and non-serotonergic targets in the pharmacotherapy of visceral hypersensitivity. **Neurogastroenterology & Motility**, v. 19, p. 89-119, 2007.

32

TANAKA, E. Gender-related differences in pharmacokinetics and their clinical significance. **Journal Of Clinical Pharmacology And Therapeutics**, v. 24, p. 339-346, 1999.

33

LILL, M. A.; DOBLER, M.; VEDANI, A. Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. **ChemMedChem**, v. 1, p. 73, 2006.

34

GOODIN, S.; CUNNINGHAM, R. 5-HT₃-receptor antagonists for the treatment of nausea and vomiting: a reappraisal of their side-effect profile. **The Oncologist**, v. 7, p. 424-436, 2002.

35

Molecular Discovery, MetaSite v. 2.7.5 [Programa de Computador]. Pinner, UK, 2007.

36

Unilever Centre for Molecular Informatics, Molprint2D [Programa de Computador]. Cambridge, UK, 2007.

37

Tripos International, Unity 2D fingerprints [Programa de Computador]. Saint Louis, USA, 2007.

38

SCHUFFENHAUER, A.; FLOERSHEIM, P.; ACKLIN, P.; JACOBY, E. Similarity metrics for ligands reflecting the similarity of the target proteins. **Journal of Chemical Information and Computer Sciences**, v. 43, n. 2, p. 391-405, 2003.

39

Openeye Scientific Software, ROCS v. 2.3.1 [Programa de Computador]. Santa Fe, USA, 2007.

40

NICHOLLS, A.; MCCUISH, N. E.; MCCUISH, J. D. Variable selection and model validation of 2D and 3D molecular descriptors. **Journal of Computer Aided Molecular Design**, v. 18, n. 7, 451-474, 2004.

41

NICHOLLS, A.; GRANT, J. A. Molecular shape and electrostatics in the encoding of relevant chemical information. **Journal of Computer Aided Molecular Design**, v. 19, n. 9-10, p. 661-686, 2005.

42

TRIBALLEU, N.; ACHER, F.; BRABET, I.; PIN, J. P.; BERTRAND, H. G. virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4. **Journal of Medicinal Chemistry**, vol. 48, n. 7, p. 2534-2547, 2005.

43

CLEVES, A. E.; JAIN, A. N. Robust ligand-based modeling of the biological targets of known drugs. **Journal of Medicinal Chemistry**, v. 49, n. 10, p. 2921-2938, 2006.

44

KIRCHMAIR, J.; MARKT, P.; DISTINTO, S.; WOLBER, G.; LANGER, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? **Journal of Computer Aided Molecular Design**, v. 22, n. 3-4, p. 213-228, 2008.

45

ISMAIL, I. M.; ANDREW, P. D.; CHOLERTON, J.; ROBERTS, A. D.; DEAR, G. J.; TAYLOR, S.; KOCH, K. M.; SAYNOR, D. A. Characterization of the metabolites of alosetron in experimental animals and human. **Xenobiotica**, v. 35, n. 2, p. 131-154, 2005.

46

<http://www.fda.gov/Cder/drug/infopage/lotronex/lotronex.htm>; acessado em 01/28/2008

47

http://www.fda.gov/medwatch/SAFETY/2005/Jun_PI/Anzemet_PI.pdf; acessado em 01/28/2008

48

DIMMITT, D. C.; CHOO, Y. S.; MARTIN, L. A.; ARUMUGHAM, T.; HAHNE, W. F.; WEIR, S. J. Single- and multiple-dose pharmacokinetics of oral dolasetron and its active metabolites in healthy volunteers: part 2. **Biopharmacology and Drug Disposition**. v. 20, p. 41-48, 1999.

49

GOLDSMITH, B. First choice for radiation-induced nausea and vomiting. **Acta Oncologica Supplement**, v. 15, p. 19-22, 2004.

50

TESTA, B.; KRÄMER, S. D. The Biochemistry of Drug Metabolism – An introduction part 2. redox reactions and their enzymes. **CHEMISTRY & BIODIVERSITY**, v. 4, p. 257-405, 2007.

51

NAKAMURA, H.; HARIYOSHI, N.; OKADA, K.; NAKASA, H.; NAKASAWA, K.; KITADA, M. CYP1A1 is a major enzyme responsible for the metabolism of granisetron in human liver microsomes. **Current Drug and Metabolism**, v. 6, n. 5, p. 469-480, 2005.

52

GOODIN, S.; CUNNINGHAM, R. 5-HT₃-Receptor Antagonists for the Treatment of Nausea and Vomiting: A reappraisal of their side-effect profile. **The Oncologist**, v. 7, p. 424-436, 2002.

53

FISCHER, V.; VICKERS, A. E. M.; HEITZ, F.; MAHADEVAN, S.; BALDECK, J. P.; MINERY, P.; VINES, R. The polymorphic cytochrome p-450_{2d6} is involved in the metabolism of both 5-hydroxytryptamine antagonists, tropisetron and ondansetron. **Drug Metabolism and Disposition**, v. 22, p. 269-274, 1994.

54

STOLTZ, A.; PARISI, S.; SHAH, A.; MACCIOCCHI, A. Pharmacokinetics, metabolism and excretion of intravenous [¹⁴C]-palonosetron in healthy human volunteers. **Biopharmacology and Drug Disposition**, v. 25, p. 329-337, 2004.

55

FIRKUSNY, L.; KROEMER, H. K.; EICHELBAUM, M. In vitro characterization of cytochrome p450 catalysed metabolism of the antiemetic tropisetron. **Biochemical Pharmacology**, v. 49, n. 12, p. 1777-1784, 1995.

56

RENDIC, S. Summary of information on human CYP enzymes: human P450 metabolism data. **Drug Metabolism Reviews**, v. 34, p. 83-448, 2002.

57

TERFLOTH, L.; BIENFAIT, B.; GASTAIGER, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. **Journal Chemical Information and Modeling**, v. 47, p. 1688-1701, 2007.

58

YAP, C. W.; CHEN, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. **Journal Chemical Information and Modeling**, v. 45, p. 982-992, 2005.

59

HAWKINS, P. C. D.; WARREN, G. L.; SKILLMAN, A. G.; NICHOLLS, A. How to do an evaluation: pitfalls and traps. **Journal of Computer Aided Molecular Design**, v. 22, p. 179-190, 2008.

60

PHAM, T. A.; JAIN, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. **Journal of Medicinal Chemistry**, v. 49, p. 5856-5868, 2006.

61

FREITAS, R. F.; BAUAB, R. L.; MONTANARI, C. A. Novel application of 2D and 3D-similarity searches to identify substrates among cytochrome P450 2C9, 2D6 and 3A4. **Journal of Chemical**

Information and Modeling, v. 50, p. 97-109, 2010.

62

JAIN, A. N.; NICHOLLS, A. Recommendations for evaluation of computational methods. **Journal of Computer Aided Molecular Design**, v. 22, p. 133-139, 2008.

63

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

64

HRISTOSOV, D. P.; OPREA, T. I.; GASTAIGER, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. **Journal of Computer Aided Molecular Design**, v. 21, p. 617-640, 2007.

65

JAIN, A. N., NICHOLLS, A. Recommendations for evaluation of computational methods. **Journal of Computer Aided Molecular Design**, v. 22, p. 133, 2008.

66

SYKES, M.; MCKINNON, R. A.; MINERS, J. Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. **Journal of Medicinal Chemistry**, v. 51, n.4, p. 780-791, 2008.

67

SHERIDAN, R. P.; KEARSLEY, S. K. Why do we need so many chemical similarity search methods? **Drug Discovery Today**, v. 7, p. 903-911, 2002.

68

MOFFAT, K.; GILLET, V. J.; WHITLE, M.; BRAVI, G.; LEACH, A. L. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. **Journal Chemical Information and Modeling**, v. 48, n. 4, p. 719-729, 2008.

69

MCGAUGHEY, G. B.; SHERIDAN, R. P.; BAYLY, C. L.; CULBERSON, J. C.; KREATSOULAS, C.; LINDSLEY, S.; MAIOROV, V.; TRUCHON, J. F.; CORNELL, W. D. Comparison of topological, shape, and docking methods in virtual screening. **Journal Chemical Information and Modeling**, v. 47, n. 4, p. 1504-1519, 2007.

70

HOLLIDAY, J. D.; SALIM, N.; WHITTLE, M.; WILLET, P. Analysis and display of the size dependence of chemical similarity coefficients. **Journal Chemical Information and Modeling**, v. 43, n. 3, p. 819-828, 2003.

71

BENDER, A.; GLEN, R. C. Molecular similarity: a key technique in molecular informatics. **Organic & Biomolecular Chemistry**, v. 2, n. 22, p. 3204-3218, 2004.

72

SING, S. B.; SHEN, L. Q.; WALKER, M. J.; SHERIDAN, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. **Journal of Medicinal Chemistry**, v. 46, n. 8, p. 1330-1336, 2003.

73

MARECHAL, J. D., KEMP, C. A., ROBERTS, G. C. K., PAINE, M. J. I., WOLF, C. R., SUTCLIFFE, M. J. Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. **British Journal of Pharmacology**, v. 153, n. S1, p. S82-S89, 2008.

74

GROOT, M. J.; ACKLAND, M. J.; HORNE, V. A.; ALEX, A. A.; JONES, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed n-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. **Journal of Medicinal Chemistry**, v. 42, n.20, p. 4062-4070, 1999.

ANEXO I

J. Chem. Inf. Model. **2010**, *50*, 97–109

Novel Application of 2D and 3D-Similarity Searches To Identify Substrates among Cytochrome P450 2C9, 2D6, and 3A4

R. F. Freitas, R. L. Bauab, and C. A. Montanari*

Grupo de Estudos em Química Medicinal de Produtos Naturais - NEQUIMED-PN, Instituto de Química de São Carlos - Universidade de São Paulo, 13560-970 - São Carlos - SP, Brazil

Received February 26, 2009

Novel Application of 2D and 3D-Similarity Searches To Identify Substrates among Cytochrome P450 2C9, 2D6, and 3A4

R. F. Freitas, R. L. Bauab, and C. A. Montanari*

Grupo de Estudos em Química Medicinal de Produtos Naturais - NEQUIMED-PN, Instituto de Química de São Carlos - Universidade de São Paulo, 13560-970 - São Carlos - SP, Brazil

Received February 26, 2009

Cytochrome P450 (CYP450) is a class of enzymes where the substrate identification is particularly important to know. It would help medicinal chemists to design drugs with lower side effects due to drug–drug interactions and to extensive genetic polymorphism. Herein, we discuss the application of the 2D and 3D-similarity searches in identifying reference structures with higher capacity to retrieve substrates of three important CYP enzymes (CYP2C9, CYP2D6, and CYP3A4). On the basis of the complementarities of multiple reference structures selected by different similarity search methods, we proposed the fusion of their individual Tanimoto scores into a consensus Tanimoto score ($T_{\text{consensus}}$). Using this new score, true positive rates of 63% (CYP2C9) and 81% (CYP2D6) were achieved with false positive rates of 4% for the CYP2C9–CYP2D6 data set. Extended similarity searches were carried out on a validation data set, and the results showed that by using the $T_{\text{consensus}}$ score, not only the area of a ROC graph increased, but also more substrates were recovered at the beginning of a ranked list.

INTRODUCTION

Similarity searching methods are widely used in modern drug discovery virtual screening and are based on the similarity principle that states molecules structurally similar are likely to have similar properties.¹ There are a number of reasons for the rapid propagation of similarity searches, which include the less cost-effective computational time, thus allowing searches in huge databases and a pivotal application when there is little (or none) information about the target and only one or two known actives.²

Since its introduction, similarity search methods have been focused on the screening of databases to identify new compounds with activity similar to that of a known reference ligand.³ However, affinity for a target enzyme is only one criterion used to select or discard a compound during the drug discovery process and constitutes the pharmacodynamic phase of a drug. The other key requirement for a safer drug is a suitable pharmacokinetic profile, which incorporates the study of ADME/Tox (absorption, distribution, metabolism, excretion, along with toxicity) properties. These studies have become an essential task to reduce the attrition rate at the late stages of the drug development process. In 1991, both poor bioavailability and pharmacokinetics were the properties responsible for 40% of all attrition rates in the drug discovery pipeline. This number dramatically dropped to 10% in 2000⁴ after the inclusion of ADME/Tox studies at earlier stages in the pipeline.

Drug metabolism is a crucial pharmacokinetic property where substrate identification is particularly important to know. It is the phase of biochemical transformation of the drug, and it is traditionally divided into phase I and phase II processes. The first involves the modification of a functional

group by oxidation, reduction, or hydrolysis. The second is responsible for the conjugation of the phase I metabolite with an endogenous molecule such as glucuronic acid.⁵ Cytochrome P450 (CYP) enzymes play a key role in phase I metabolism, where the goal is to convert the drug into a polar form by adding an ionizable group, thus making them more water-soluble and more readily to be excreted.

In this regard, CYP inhibition is a major safety problem. Very often, two or more drugs are coadministered to a patient during the treatment of a disease, and the individual compounds may compete to be metabolized by the same enzyme. This leads to unintended effects of drug–drug interactions where a drug inhibits the metabolism, causing an increase in the plasma concentration of a second drug, which can lead to an increase in the toxic side effects.^{6,7} That is, measurement of CYP inhibition during discovery provides early warning of potential safety issues. This means CYP substrate selectivity is sometimes much less of an issue and in fact there may be potential advantages in having a drug that is a substrate for more than one CYP because if there is a problem (inhibition or individual variation) with one CYP another one can take over. In addition, polymorphisms (result of genetic mutations) affect the activity of a number of drug-metabolizing enzymes such as CYP2C19, CYP2D6, and CYP1A2. For example, the polymorphic enzyme CYP2D6 is absent in 5–10% of Caucasians and gives rise to poor metabolizers that have low or lack of activity.⁸ Hence, it would be highly beneficial for the discovery and development of safer drugs to introduce in silico methods to predict, as early as possible, the structural features that imprint substrate identification. Additionally, because the drug molecule can be metabolized to either reactive or chemically stable metabolites, leading, respectively, to drug-induced toxicity or enhanced pharmacology, such methods would be of utmost importance.

* Corresponding author phone: 55-16-3373-9986; fax: 55-16-3373-9985; e-mail: montana@iqsc.usp.br.

Computational work has hitherto experienced the immediate possibility to predict substrate selectivity mainly focused on structure-based methods such as docking using homology models and X-ray structures of the CYP enzymes.^{9,10} Ligand-based approaches such as pharmacophore^{11,12} and QSAR^{13,14} analyses have also been extensively employed. Based upon the need to further the knowledge in this field, the present work introduces a novel approach that expands the scope of similarity searching methods from the classical application for a single target to deal with the challenging task of explaining the substrate identification of relevant CYP enzymes. For a statistically significant database of 596 substrates of the three most important CYP enzymes (CYP2C9, CYP2D6, and CYP3A4), 2D and 3D-similarity searches were used in the determination of the CYP enzyme predominantly responsible for the metabolism of a compound. These enzymes are highly promiscuous, which means that they are able to metabolize a broad range of chemically diverse substrates and many times more than one isoform metabolizes the same substrate, which has previously been mentioned to be sometimes advantageous where selectivity is not necessarily an issue. Herein, we show a successful application of a potentially powerful method to hot spot reference structures that enables the identification and separation among substrates and nonsubstrates for a particular CYP isoform.

METHODS

Data Sets. Drugs that are substrates of CYP2C9, CYP2D6, and CYP3A4 were mainly collected from the previous work of Rendic,¹⁵ Terfloth,¹⁶ and Yap.¹⁷ As we carried out pairwise analysis between these three cytochromes (2C9 versus 2D6, 2C9 versus 3A4, and 2D6 versus 3A4), for each compound, a search in the literature was carried out to confirm if it was entirely metabolized by the CYP enzyme according to the statement in the three papers above. If a compound is metabolized by more than one CYP450, it was classified as being a substrate of the isoform responsible for the main route of its metabolism. To exemplify our strategy to build strong relationships between the desired CYP and the substrate to be metabolized, compounds were hand-selected to allow the outstanding achievement of identifying the CYP for that metabolism. Albeit this is time-consuming, it ensures the high quality of the data. With this protocol, we were able to construct meaningful pairwise substrate sets. For example, in Figure 1, 73 compounds could be used to analyze CYP2C9 over CYP2D6, and 98 for CYP2D6 over CYP2C9 (opposite direction).

The generated data sets are composed of 596 compounds, out of which 417 are unique. The entire set is composed by 121 CYP2C9 substrates (20%), 151 CYP2D6 substrates (25%), and 324 CYP3A4 substrates (55%). From the analysis of Figure 1 and the previous statistics, the data sets CYP2C9–CYP3A4 and CYP2D6–CYP3A4 are very unbalanced. This observation agrees with the fact that more than one-half of all marketed drugs are metabolized by the CYP3A4 enzyme. The whole number of compounds differs from the number of unique compounds because a substrate could be present in more than one pairwise analysis. For example, the drug aceclofenac

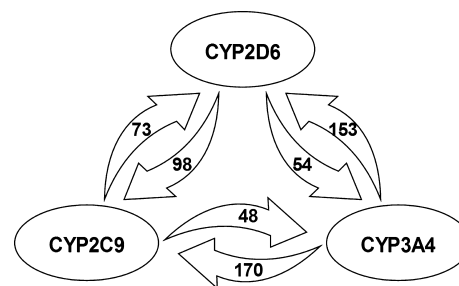


Figure 1. Diagram representing CYPs compound set. Each arrow represents a set, and its direction defines the CYP for that metabolism. For example, in the CYP2C9–CYP2D6 data set, we have 171 compounds. Out of these, 73 are CYP2C9 substrates, and 98 are CYP2D6 substrates.

was considered a substrate of CYP2C9 in both pairwise analysis, CYP2C9–CYP2D6 and CYP2C9–CYP3A4. The reported CYP substrate benchmark will be made freely available to support compound identification analyses using other computational approaches.

Molecular Similarity. On the basis of the similar property principle,¹ which states that structural similar molecules are more likely to have similar properties, we carried out an all-against-all similarity analysis for every compound in our data set to identify the desired CYP substrate. Systematic pairwise similarity calculations were fulfilled to evaluate the structural diversity distribution of each pair of enzymes. The similarity between the compounds was measured on the basis of the Tanimoto coefficient (T_c). Figure 2 shows the workflow for the approach adopted in this work. Each compound was used as reference compound, and a similarity value was calculated to all compounds in the database. This leads to a square similarity Tanimoto matrix with the number of rows and columns equal to the number of compounds (Figure 2).

In the present work, we investigate the performance of 2D and 3D-similarity searches for their ability to distinguish compounds between closely related targets. For 2D-similarity searching, two 2D fingerprints were selected that are calculated from 2D molecular graphs: Molprint2D and Unity 2D fingerprint.¹⁸ In addition, ROCS (rapid overlay of chemical structures) was used to perform 3D-similarity searches. Tanimoto matrices were generated using the publicly available Matrix2png software.¹⁹

Quantification of Performance. The evaluation of the different methods to discriminate between substrates and nonsubstrates was carried out by using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. Plotting a ROC curve consists of the determination of the sensitivity (Se) and the specificity (Sp) at every possible score threshold.²⁹ The first value (Se) describes the ratio of the number of true actives that are selected by the method to the number of all actives in the database:

$$Se = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{TP}{TP + FN}$$

The second value (Sp) describes the ratio of the number of inactives that are discarded by the method to the number of all inactive molecules included in the databases:

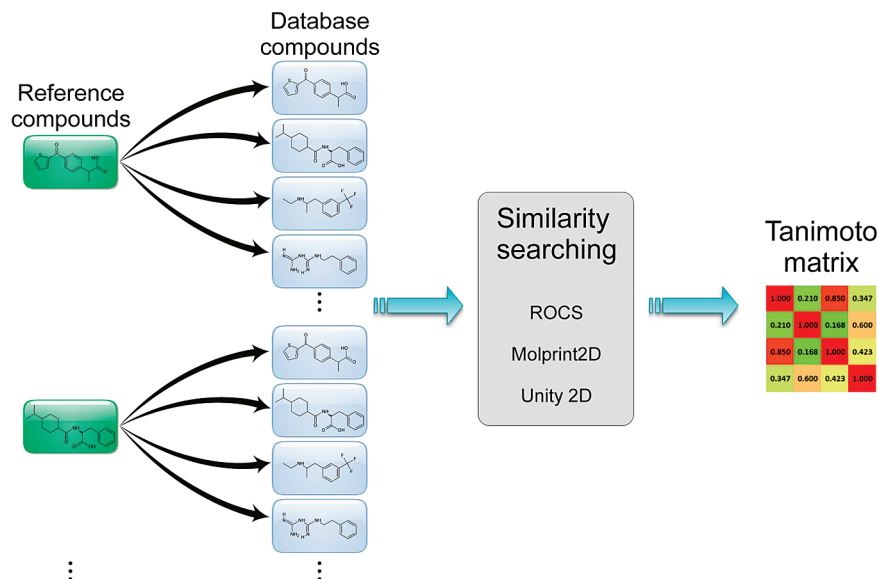


Figure 2. Workflow representation of the pairwise similarity calculations.

$$Se = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{TP}{TP + TN}$$

Both of these parameters can vary between 0 and 1. Usually, the greater is the AUC, the more effective the virtual screening workflow is in discriminating active from inactive compounds.²⁹

RESULTS AND DISCUSSION

The analysis of some physicochemical properties of the compounds studied here shows that they have similar values for almost all properties. The molecules belonging to the CYP2C9 and CYP3A4 sets are the ones that present more related properties, with the molecular weight (MW) being a little large in the CYP3A4 set. Despite that the CYP2D6 set also displays many comparable properties with the other two data sets, the mean molecular weight and mean polar surface area (PSA) have the smallest values of the whole data set, which indicate that the CYP2D6 compounds are smaller than that of the other two enzymes. The constitution of data sets has a pronounced effect on the efficacy of a virtual screening test.²⁰ For instance, in a recent work of Jain et al., the performance of 3D virtual screening methods was comparable to a simple 1D method (molecular weight, $\log P$, number of hydrogen-bond donors, number of hydrogen-bond acceptors, and number of rotatable bonds).²¹ This result indicates that the active compounds are very dissimilar from the inactive set, because a simple 1D method is very efficient in the separation of these two sets. Therefore, the good results found in many publications in the area of retrospective virtual screening are purely due to differences in simple properties between the actives and the inactives. Next, an inactive set should be as similar as possible to the active compounds to obtain a trust indication of the utility of the virtual screening method. Therefore, our virtual library is appropriate to assess the usefulness of the tools employed in the present work because there is no statistical difference within the simple 1D molecular properties (Table 1).

Compound Diversity. The intraset structural diversity of the CYP2C9–CYP2D6 data set was evaluated using sys-

Table 1. Mean and Standard Deviation of Some 1D Properties of the Substrates of Our Database

	CYP2C9		CYP2D6		CYP3A4	
	mean	std. dev.	mean	std. dev.	mean	std. dev.
HBA ^a	5	2	4	2	5	3
HBD ^b	1	1	2	1	2	1
MW ^c	321	88	303	77	369	114
RB ^d	5	3	5	3	5	3
$\log P$ ^e	3	2	3	2	3	2
PSA ^f	69	31	44	23	70	35

^a Hydrogen-bonding acceptor. ^b Hydrogen-bonding donor. ^c Molecular weight. ^d Rotatable bonds. ^e Partition coefficient. ^f Polar surface area.

tematic pairwise similarity comparisons. Because each compound was used as reference compound and a similarity value was calculated to all compounds in the database, the result was a square similarity Tanimoto matrix with the number of rows and columns equal to the number of compounds (Figure 3a–c). As we would expect, there is a diagonal line (red), where the similarity is maximum, which means that the similarity was measured using the same structure, or, in other words, the reference and the database compound are equals. We can observe regions presenting high similarities outside that diagonal, indicating that many substrates are also very similar to other ones.

For the 3D-similarity search (ROCS), the compound structural diversity, quantified by the Tanimoto coefficient, considerably differs depending on the level of information used to score the compounds and on the method used. Analyzing the results of the ROCS method after the inclusion of chemical features (color score) is possible to observe that the CYP2D6 substrates present a high intrasimilarity, as observed by the large right-bottom cluster in the matrix (Figure 3a), meaning that these substrates have a great homogeneity in their structures. With regard to the CYP2C9 substrates, it is possible to verify some disperse points in the matrix presenting a high similarity, but in general no significant cluster is apparent, indicating that the substrates of this isoform display a higher chemical diversity if compared to the substrates from CYP2D6 isoform. The

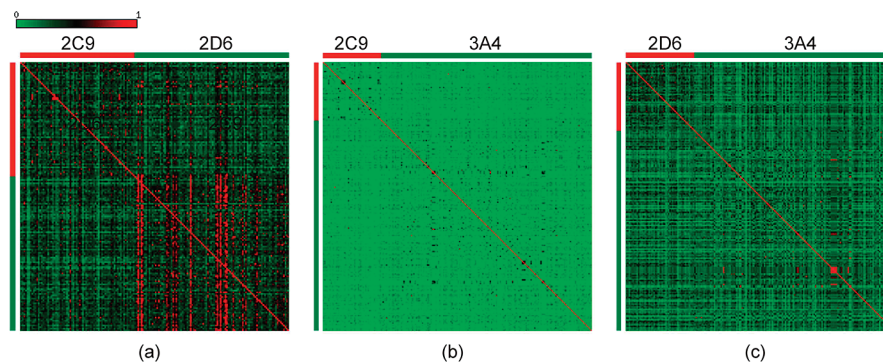


Figure 3. Tanimoto matrices for pairwise comparisons of the three datasets: (a) 2C9–2D6/ROCS-color; (b) 2C9–3A4/MOLPRINT2D; (c) 2D6–3A4/Unity 2D. Similarity values were assessed using the Tanimoto coefficient (T_c). For T_c values, continuous color-coding is used with increasing similarity from green to red (black indicates $T_c = 0.5$). In each matrix, the position of the compounds for a particular substrate selectivity set is marked by color bars.

comparison of molecules on both shape and chemical complementarity was effective to produce a cluster in the Tanimoto matrix for the CYP2D6 substrates (Figure S1).

The similarity searching using fingerprints, Molprint2D and Unity 2D, provided results similar to those with ROCS-color. Using Molprint2D fingerprint, the compounds appear to have the most diverse structures, as observed by their extremely low pairwise similarity in the Tanimoto matrix (Figure S1). However, inside each class, the similarity is higher, and it is possible to see the two clusters, one for the CYP2C9 substrates and the other for the CYP2D6 substrates. The Unity 2D fingerprint shows a behavior similar to the one observed for Molprint2D, with the CYP2D6 substrates presenting high pairwise similarities (Figure S1).

For the CYP2C9–CYP3A4 data set, Molprint2D provided a clear cluster in the diversity matrix at the top left of the matrix, which corresponds to the CYP2C9 substrates (Figure 3b). Using ROCS, independently of the kind of score used, it was not possible to identify a separation between the substrates of the CYP2C9 and CYP3A4 enzymes (Figure S2). Unity 2D shows an analogous behavior to the viewed for ROCS, where no cluster is observed in the Tanimoto matrix (Figure S2).

The systematic pairwise similarity comparison of the CYP2D6–CYP3A4 data set using the Unity 2D fingerprint shows a top-left cluster including all CYP2D6 substrates (Figure 3c). The remaining areas of the matrix presented low pairwise similarity among the structures, with the absence of any cluster. Molprint2D and ROCS show a Tanimoto matrix similar to that viewed for Unity 2D, a top-left cluster corresponding to the CYP2D6 substrates, and an absence of any cluster in the CYP3A4 area (Figure S3).

It is noteworthy that no method was able to produce a significant cluster in the Tanimoto matrix for the CYP3A4 substrates in the data sets where they were present, with the exception of some very disperse small areas of high intrasimilarity. These results demonstrate that its substrates have a huge chemical diversity, and this characteristic is in agreement with the broad range of substrate specificity displayed by CYP3A4 enzyme, which metabolizes nearly 50% of the marketed drugs.²²

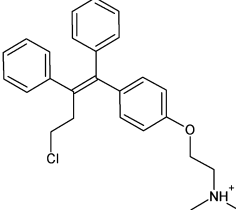
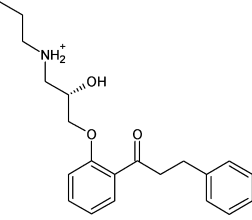
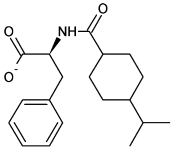
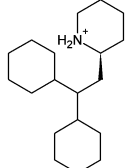
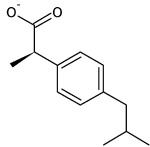
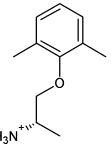
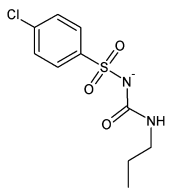
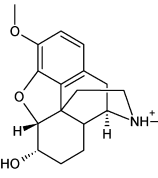
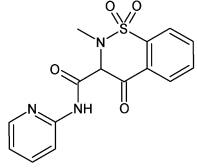
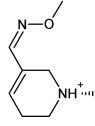
ROC-AUC Analysis. The main objective of the present work was to answer the following question: is it possible to identify reference structures that provide a good separation between the substrates of two CYP isoforms using 2D and

3D-similarity searches? To answer this question, we carried out pairwise similarity searches using each compound as a reference and database compound. Next, the performance of the different similarity searches was assessed by measuring the area under the receiver operator characteristic (ROC) curve for every compound used as a reference. Such graphs show the performance of a given tool when the screening across the entire database is examined. ROC curve has been shown to be a powerful technique for investigating the ability of retrospective virtual screening methods to discriminate between active and inactive compounds.³⁵

Despite the high chemical diversity exhibited by the CYP2C9 substrates, all methods were able to identify reference structures that allowed the identification of substrates and nonsubstrates for this enzyme. Table 2 shows the chemical structures of the substrates that were more efficient to distinguish the substrates of the two isoforms (CYP2C9 and CYP2D6) according to a given similarity search. The one where the reference structure gave the best performance has its AUC value highlighted in bold. For example, the substrate toremifene is the one that has the highest AUC value among all CYP2C9 substrates using the ROCS-shape. Just for comparison, the AUC values using other similarity searches are also displayed. The best performance in the separation of CYP2C9 and CYP2D6 substrates was provided by ROCS-color (0.807), ROCS-combo (0.755), and Unity 2D (0.707), using nateglinide, ibuprofen, and piroxicam, respectively, as reference structures. These AUC values mean that the score of a randomly selected substrate is higher than a randomly selected non-substrate 7 times out of 10. Also, it is clear from the analysis of Table 2 that the results from ROCS are better when using some kind of chemical typing rather than just shape, because the AUC goes from 0.566 (ROCS-shape) to 0.807 (ROCS-color). The substrates toremifene (ROCS-shape) and chlorpropamide (Molprint2D) presented only marginal AUC values.

It is also possible to see from Table 2 the reference structures that were more efficient to separate substrates and nonsubstrates of the CYP2D6 isoform. Using the substrate perhexiline as reference compound, we were able to separate CYP2D6 substrates from CYP2C9 with an impressive AUC value of 0.912. In addition, the AUC values for three reference structures are higher than 0.850.

Table 2. Best Reference Structures and the Respective AUC Values (Bold) That Identified Substrates of the Enzymes CYP2C9 and CYP2D6^a

CYP2C9		CYP2D6	
Structure	Method AUC	Structure	Method AUC
 Toremifene	ROCS-shape 0.566	 Propafenone	ROCS-shape 0.581
	ROCS-color 0.438		ROCS-color 0.667
	ROCS-combo 0.459		ROCS-combo 0.689
	Molprint2D 0.371		Molprint2D 0.625
	Unity 2D 0.250		Unity 2D 0.684
 Nateglinide	ROCS-shape 0.357	 Perhexiline	ROCS-shape 0.398
	ROCS-color 0.807		ROCS-color 0.912
	ROCS-combo 0.639		ROCS-combo 0.822
	Molprint2D 0.561		Molprint2D 0.646
	Unity 2D 0.372		Unity 2D 0.849
 Ibuprofen	ROCS-shape 0.514	 Mexiletine	ROCS-shape 0.550
	ROCS-color 0.721		ROCS-color 0.901
	ROCS-combo 0.755		ROCS-combo 0.893
	Molprint2D 0.547		Molprint2D 0.572
	Unity 2D 0.260		Unity 2D 0.716
 Chlorpropamide	ROCS-shape 0.456	 Dihydrocodeine	ROCS-shape 0.513
	ROCS-color 0.617		ROCS-color 0.719
	ROCS-combo 0.560		ROCS-combo 0.615
	Molprint2D 0.630		Molprint2D 0.795
	Unity 2D 0.512		Unity 2D 0.693
 Piroxicam	ROCS-shape 0.539	 Milameline	ROCS-shape 0.439
	ROCS-color 0.593		ROCS-color 0.844
	ROCS-combo 0.562		ROCS-combo 0.834
	Molprint2D 0.625		Molprint2D 0.624
	Unity 2D 0.707		Unity 2D 0.857

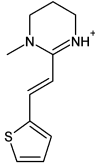
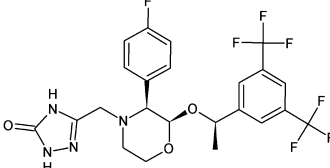
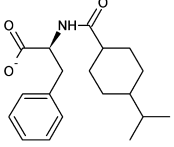
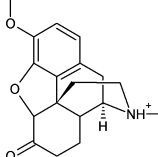
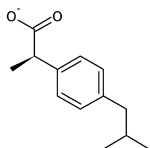
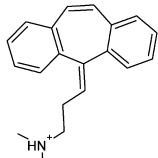
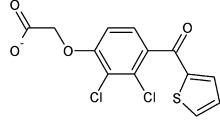
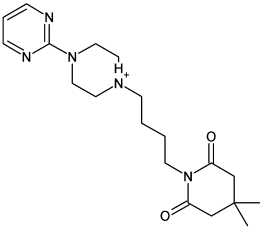
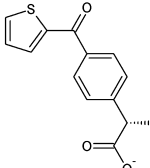
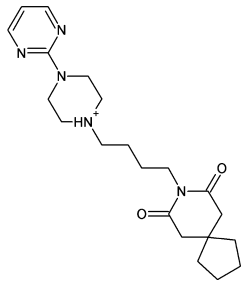
^a For comparison, the AUC values of all methods are also displayed.

Another interesting result shown in Table 2 is that all reference structures of the CYP2D6 enzyme display a protonated nitrogen in their structures. Our results are in agreement with other works where a usual characteristic of the majority of CYP2D6 substrates is the presence of a basic nitrogen atom and an aromatic ring.^{24,25} In addition, CYP2C9 substrates are usually weak acids with multiple aromatic

rings. Looking at Table 2, we see that three reference structures meet this structural pattern.

The reference structures that provided the best separation between the substrates of the CYP2C9 and CYP3A4 enzymes are displayed in Table 3. Direct information that emerges from the analysis of this table is that the substrates nateglinide and ibuprofen were again the best ones to separate CYP2C9

Table 3. Best Reference Structures and the Respective AUC Values (Bold) That Provided a Separation between the Substrates of the Enzymes CYP2C9 and CYP3A4^a

CYP2C9		CYP3A4	
Structure	Method AUC	Structure	Method AUC
 Pyrantel	ROCS-shape 0.684	 Aprepitant	ROCS-shape 0.668
	ROCS-color 0.318		ROCS-color 0.557
	ROCS-combo 0.469		ROCS-combo 0.643
	Molprint2D 0.538		Molprint2D 0.502
	Unity 2D 0.431		Unity 2D 0.623
 Nateglinide	ROCS-shape 0.478	 Hydrocodone	ROCS-shape 0.412
	ROCS-color 0.742		ROCS-color 0.755
	ROCS-combo 0.690		ROCS-combo 0.590
	Molprint2D 0.561		Molprint2D 0.607
	Unity 2D 0.395		Unity 2D 0.618
 Ibuprofen	ROCS-shape 0.650	 Cyclobenzaprine	ROCS-shape 0.424
	ROCS-color 0.640		ROCS-color 0.732
	ROCS-combo 0.739		ROCS-combo 0.683
	Molprint2D 0.623		Molprint2D 0.537
	Unity 2D 0.505		Unity 2D 0.538
 Tielinic acid	ROCS-shape 0.675	 Gepirone	ROCS-shape 0.425
	ROCS-color 0.552		ROCS-color 0.671
	ROCS-combo 0.694		ROCS-combo 0.532
	Molprint2D 0.669		Molprint2D 0.689
	Unity 2D 0.586		Unity 2D 0.696
 Suprofen	ROCS-shape 0.645	 Buspirone	ROCS-shape 0.464
	ROCS-color 0.646		ROCS-color 0.671
	ROCS-combo 0.729		ROCS-combo 0.559
	Molprint2D 0.663		Molprint2D 0.662
	Unity 2D 0.624		Unity 2D 0.707

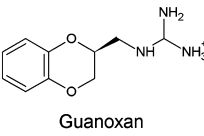
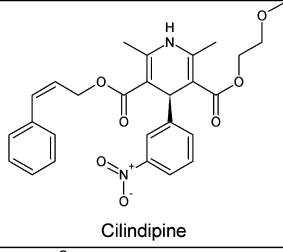
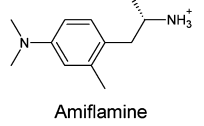
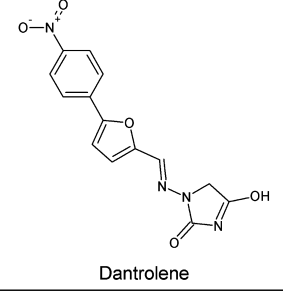
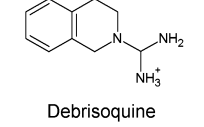
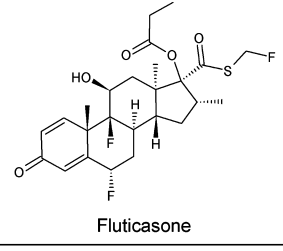
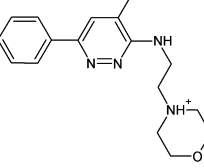
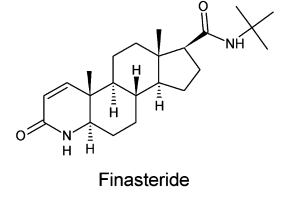
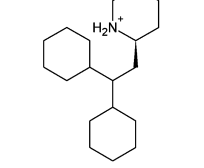
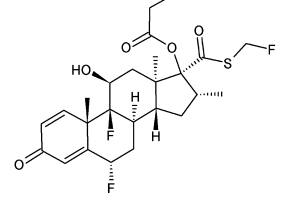
^a For comparison, the AUC values of all methods are also displayed.

substrates from those of the CYP3A4 isoform, indicating that they are typical structures of the CYP2C9 substrates. Another intriguing result is that three of five reference structures of the CYP2C9 enzyme belong to the class of nonsteroidal anti-inflammatory drugs (NSAIDs): ibuprofen, suprofen, and tielinic acid. The substrate nateglinide belongs to the megli-

tinide class of blood glucose-lowering drugs, but as the other three substrates, it also presents the usual structure of the CYP2C9 substrates: a carboxylate group and at least one aromatic ring.

We can see that all reference structures of the CYP3A4 enzyme displayed an AUC value close to 0.700, with

Table 4. Best Reference Structures and the Respective AUC Values (Bold) That Provided a Separation between the Substrates of the Enzymes CYP2D6 and CYP3A4^a

CYP2D6		CYP3A4	
Structure	Method AUC	Structure	Method AUC
 Guanoxan	ROCS-shape 0.741	 Cilindipine	ROCS-shape 0.673
	ROCS-color 0.747		ROCS-color 0.486
	ROCS-combo 0.844		ROCS-combo 0.596
	Molprint2D 0.730		Molprint2D 0.434
	Unity 2D 0.754		Unity 2D 0.666
 Amiflamine	ROCS-shape 0.631	 Dantrolene	ROCS-shape 0.415
	ROCS-color 0.806		ROCS-color 0.822
	ROCS-combo 0.851		ROCS-combo 0.603
	Molprint2D 0.668		Molprint2D 0.539
	Unity 2D 0.772		Unity 2D 0.677
 Debrisoquine	ROCS-shape 0.711	 Fluticasone	ROCS-shape 0.615
	ROCS-color 0.783		ROCS-color 0.668
	ROCS-combo 0.880		ROCS-combo 0.698
	Molprint2D 0.680		Molprint2D 0.573
	Unity 2D 0.826		Unity 2D 0.784
 Minaprine	ROCS-shape 0.568	 Finasteride	ROCS-shape 0.519
	ROCS-color 0.683		ROCS-color 0.608
	ROCS-combo 0.671		ROCS-combo 0.566
	Molprint2D 0.760		Molprint2D 0.685
	Unity 2D 0.615		Unity 2D 0.485
 Perhexiline	ROCS-shape 0.541	 Fluticasone	ROCS-shape 0.615
	ROCS-color 0.769		ROCS-color 0.668
	ROCS-combo 0.761		ROCS-combo 0.698
	Molprint2D 0.592		Molprint2D 0.573
	Unity 2D 0.843		Unity 2D 0.784

^a For comparison, the AUC values of all methods are also displayed.

hydrocodone showing the highest value (0.755). The same observation emerges for the reference structures of the CYP2C9 enzyme. These results reveal the difficulty to separate the substrates of these two isoforms. These can be explained by the fact that these two isoforms are able to metabolize a wide range of chemically diverse substrates. It is quite common that the substrates overlap between these CYP450 enzymes where more than one enzyme metabolizes the same substrate.

From the analysis of Table 4, we can see the reference structures selected by each method to separate the CYP3A4

substrates from those of CYP2D6 enzyme. Of the five reference structures, two (fluticasone and finasteride) present the lipophilic steroid scaffold, characterized by a terpenoid lipid with a carbon skeleton with four fused rings, arranged in a 6–6–6–5 fashion. Actually, the substrate fluticasone was selected by ROCS-combo and Unity 2D fingerprint as the best reference structure. This result agrees with the preference of CYP3A4 to catalyze the oxidation of lipophilic neutral or basic compounds.

The presence of basic nitrogen in the CYP2D6 substrates is evident if we look at Table 4, which shows the

best reference structures to separate the substrates of the CYP2D6 and CYP3A4 enzymes. In addition, three of these structures show AUC values higher than 0.800, indicating that a higher score is assigned 8 times out of 10 to a randomly selected substrate more than that of a randomly selected nonsubstrate. Too, the substrate perhexiline is found to be the best reference structure using ROCS-color in the analysis of the CYP2C9–CYP2D6 data set. It was selected again using Unity 2D fingerprint, validating its structure as an efficient reference structure to identify CYP2D6 substrates.

Before concluding, some further considerations on methods performance and discrimination among substrates and nonsubstrates have to be provided. Clearly, shape score is the poorest of the three available in ROCS (shape, color, and combo score). There is a sensible improvement in ROCS results when we use color and combo score. The two fingerprints used in this work (Molprint2D and Unity 2D) perform as well as ROCS-shape score and in some cases better, but with the advantage that they are less computationally expensive. Several papers have also found that for various target classes, 2D-similarity searches using fingerprints have given comparable performance or even superior results to the 3D methods.^{2,26,27} A possible explanation might be that the connection table of a molecule encodes implicit information about the structure of a molecule that is lost in the 3D-methods, which ignores bond topology in favor of atom positions.² Comparing just the two fingerprints, we see that Unity 2D works better than Molprint2D, because in five out of six reference structures the area under the ROC curve is higher when using the former fingerprint. They only show comparable performance for the CYP2C9–CYP3A4 data set. The other consideration is about the difficulty to discriminate between substrates and nonsubstrates for the three data sets. According to the AUC values achieved by each reference structure, we can see that the most challenging is the CYP2C9–CYP3A4 data set, because it is the one that is constituted by substrates of the two most chemically diverse substrates. At the other extreme, we have the CYP2C9–CYP2D6 data set. Identifying CYP2D6 substrates among the ones from CYP2C9 is not a difficult task because both enzymes are often represented by distinctly compounds that occupy different chemical spaces: CYP2C9 substrates are usually weak acids, and the CYP2D6 substrates have a basic nitrogen atom in its structures. Finally, the CYP2D6–CYP3A4 is at an intermediate position.

Consensus Similarity. In the light of the comparable efficiency of 2D and 3D-similarity searches, we have noted that despite that some method could be more robust to identify substrates on the whole, it is practically unfeasible for only one method to distinguish unambiguously all of the substrates of enzymes highly promiscuous like the cytochrome P450. Clearly, each method can find some structures that all other methods would miss. Therefore, we think the union of different similarity searching methods and multiple reference structures is a good strategy to retrieve as many substrates as possible for challenging enzymes like the CYPs. Actually, similar strategies were proposed in the works of Muresan et al.²⁸ and Willett et al.²⁹ The former carried out a multifingerprint selection by merging the compounds coming from a ranked list of selected targets corresponding to all fingerprints. They claimed that a multifingerprint

approach could be an efficient tool to balance the strengths and weaknesses of various fingerprints.²⁸ In the work of Willett et al., they merged individual fingerprints of a set of 10 reference structures into a single combined fingerprint. This resulted in an increase of over two-thirds in the numbers of actives retrieved.²⁹ In another recent work, Gasteiger et al. found that the use of similarity search with subsequent data fusion produced up to 16% better BEDROC scores than the novelty detection with self-organizing maps.³⁰

To investigate our hypothesis, we took the best three reference structures for each data set and combined their Tanimoto score in a new score called consensus Tanimoto score ($T_{\text{consensus}}$). The scores of each method were scaled to a number between 0 and 1 using the following formula:

$$T_{\text{scaled}} = (T - T_{\text{min}})/(T_{\text{max}} - T_{\text{min}}) \quad (1)$$

Here, T_{scaled} is the scaled Tanimoto values of the i_{score} (for example, the Tanimoto values of the ROCS-color), T is the raw value of the i_{score} , T_{min} is the smallest Tanimoto value for a given method, and T_{max} is the largest Tanimoto value for a given method. To obtain the consensus Tanimoto score, $T_{\text{consensus}}$, we carried out the sum of the three T_{scaled} selected, and this value was divided by 3 according to eq 2.

$$T_{\text{consensus}} = \sum_{i_{\text{score}}} T_{\text{scaled}}/3 \quad (2)$$

With this scaling process, all of the Tanimoto scores can be simultaneously employed in consensus similarity analysis, regardless of the differences between them, such as, for example, the ROCS-combo score, which is a combination of shape and color score, with a maximum value equal to 2.

By the analysis of Figure 4, we can see the use of the $T_{\text{consensus}}$ score improves the AUC values for almost all data sets. Only one exception was observed in the case of the CYP2D6–CYP3A4 data set (Figure 4c), where the performance of the new $T_{\text{consensus}}$ score (AUC = 0.878) was similar to that using just the compound debrisoquine as a reference structure (AUC = 0.880).

ROC curves have many advantages over the more traditional graphical method (enrichment curves). First, they do not depend on the ratio of the screened set of molecules. Second, they provide the entire spectrum of sensitivity/specificity pairs, the only one that supplies a complete picture of test accuracy reporting the dual aspect of any test, the ability to select active compounds and discard inactive ones.²⁹ However, ROC curves have an important drawback: they are not able to deal with the early recognition problem.^{31,32} For instance, consider the case where two ROC curves have the same AUC of exactly 0.5. In the first curve, one-half of the actives are retrieved at the very beginning of the rank-ordered list and the other one-half at the end, while in the second curve, all of the actives are retrieved in the middle of the list. Clearly, the first curve, in terms of the early recognition problem, is better than the second curve. Because the retrieval of active structures on the top ranked list is the main aim of a virtual screening test, saving us the trouble of actually screening all of the compounds, it is entirely reasonable to prefer good “early” behavior.^{30,31}

On the basis of the previous paragraph, we consider the observation that in five out six curves, our strategy was

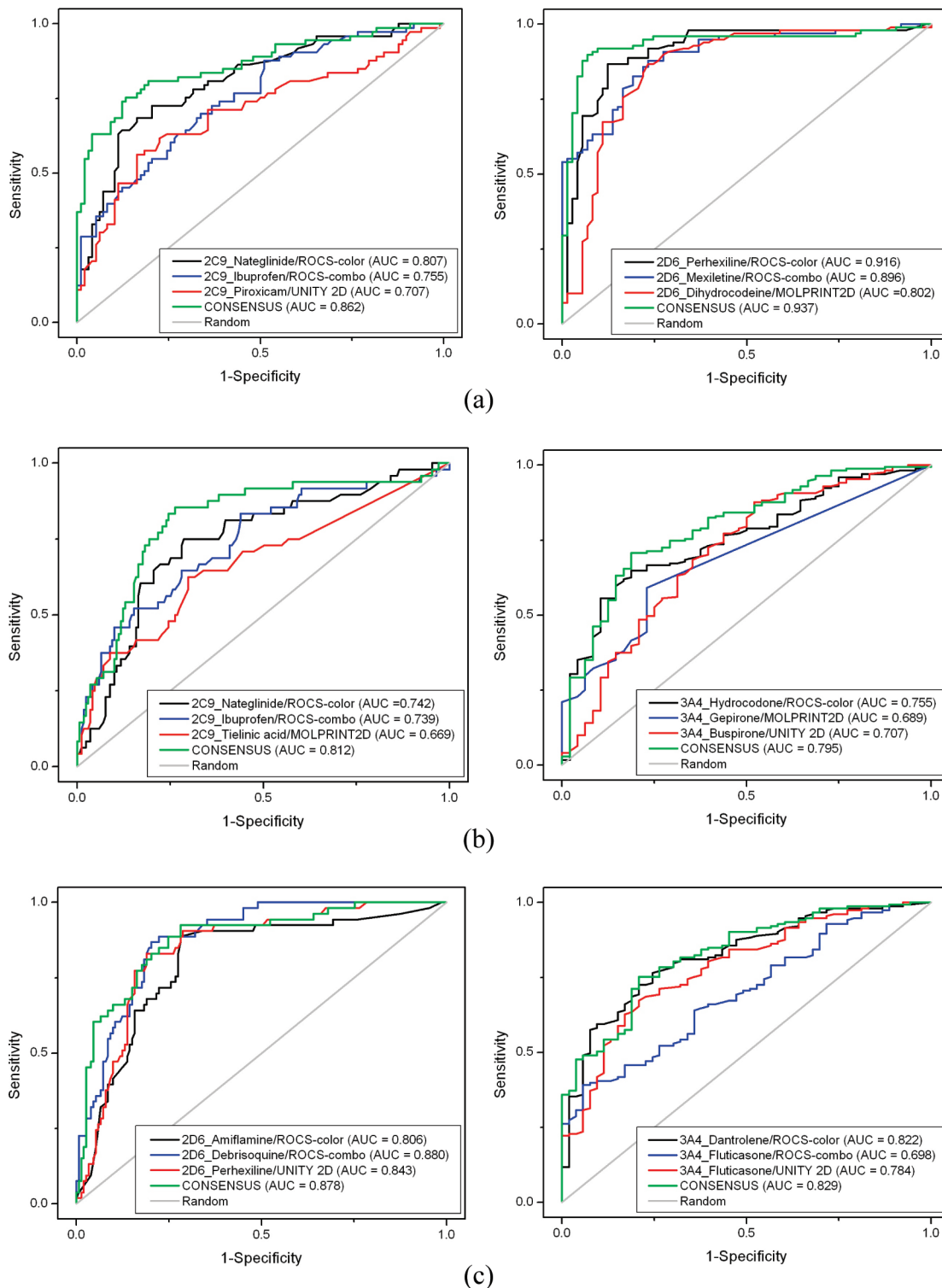


Figure 4. ROC curves comparing the performance between isolated reference structures with the consensus of three reference structures: (a) CYP2C9–CYP2D6; (b) CYP2C9–CYP3A4; and (c) CYP2D6–CYP3A4.

responsible for the identification of more substrates at the beginning of the ROC curve, even more important than the improvement of the area under the curve. In the CYP2C9–CYP2D6 data set, using the $T_{\text{consensus}}$ score, a high sensitivity (63% and 81% for CYP2C9 and CYP2D6, respectively) was achieved at a low 1-specificity of only 4% (Figure 4a). The same sensitivity in the CYP2C9–CYP3A4 data set was achieved at a higher 1-specificity (14–15%) (Figure 4b). For the CYP2D6–CYP3A4 data set, a sensitivity and 1-specific-

ity of 60% and 4.6%, respectively, were obtained using the $T_{\text{consensus}}$ score for the separation of CYP2D6 substrates (Figure 4c). The only exception mentioned above was found in the separation of CYP3A4 substrates that showed a sensitivity and 1-specificity equal to 60% and 19%, respectively. These results reinforce the challenging task of identifying CYP3A4 substrates using similarity search methods, mainly because of its large and varied “menu” of structures.

Table 5. Results from a Similarity Search Carried Out in an External Set of CYP Substrates Using the Best Reference Structures and the $T_{\text{consensus}}$ Score (Consensus)

CYP2C9–CYP2D6 ($n = 53$)					
CYP2C9			CYP2D6		
reference structure	method	AUC	reference structure	method	AUC
nateglinide	ROCS-color	0.865	perhexiline	ROCS-color	0.944
ibuprofen	ROCS-combo	0.762	mexiletine	ROCS-combo	0.917
piroxicam	Unity 2D	0.760	milameline	Unity 2D	0.790
consensus		0.913	consensus		0.955
CYP2C9–CYP3A4 ($n = 70$)					
CYP2C9			CYP3A4		
reference structure	method	AUC	reference structure	method	AUC
nateglinide	ROCS-color	0.559	hydrocodone	ROCS-color	0.609
ibuprofen	ROCS-combo	0.797	gepirone	Molprint2D	0.565
tielinic acid	Molprint2D	0.560	bupirone	Unity 2D	0.620
consensus		0.781	consensus		0.652
CYP2D6–CYP3A4 ($n = 79$)					
CYP2D6			CYP3A4		
reference structure	method	AUC	reference structure	method	AUC
amiflamine	ROCS-color	0.858	dantrolene	ROCS-color	0.714
debrisoquine	ROCS-combo	0.861	fluticasone	ROCS-combo	0.664
perhexiline	Unity 2D	0.830	fluticasone	Unity 2D	0.756
consensus		0.929	consensus		0.814

Validation. Our strategy was further validated with an external data set of 99 substrates (21 2C9 substrates, 31 2D6 substrates, and 47 3A4 substrates). These compounds are distinct from those used in the model development and were collected from various sources.^{33,34}

Analyzing Table 5, we can see that the reference structures previously selected were able to correctly classify the new substrates with a higher efficiency. In addition, the new $T_{\text{consensus}}$ score allowed us to achieve impressive AUC values ranging from 0.652 to 0.955. This performance was similar to, and in some cases better than, the AUC values displayed in Figure 4. The poorest result, but not totally unacceptable, comes from the CYP3A4 substrates. It is noticeable that this bad result is tightly associated with the diversity of the substrates of this enzyme. The larger is the chemical diversity of a set of compounds, the more difficult it is to find a reference structure able to retrieve the majority of the compounds. Nevertheless, the use of the $T_{\text{consensus}}$ score gave rise to a moderate improvement in the AUC values when compared to the search using a single reference structure.

For the validation data set, as observed during the model development, the use of the $T_{\text{consensus}}$ score not only increased the area of a ROC graph, but also increased the recovery of the substrates at the beginning of the ranked list. Using the $T_{\text{consensus}}$ score, the sensitivity and the 1-specificity were 81.8% and 0%, respectively, for the identification of the CYP2C9 substrates in CYP2C9–CYP2D6 data set (Figure 5). A similar result was obtained in the identification of the CYP2D6 substrates in the same data set where the sensitivity and the 1-specificity were 84.4% and 4.7%, respectively (Figure 5). Thus, the selection of appropriate reference structures and the combination of their individual Tanimoto measures coming from different similarity search methods in a consensus score is a robust strategy to explore better the chemical space described by the set of reference

structures, and in this way increase the number of CYP substrates identified.

CONCLUSION

In this work, we carried out pairwise similarity searches to assess the ability of 2D and 3D-similarity searches to identify reference structures with a high capacity to discriminate substrates and nonsubstrates for three pairs of data sets: CYP2C9–CYP2D6, CYP2C9–CYP3A4, and CYP2D6–CYP3A4. A detailed ROC-AUC analysis demonstrates that even for enzymes highly promiscuous like CYP450, the similarity search methods used in the present work were capable of identifying reference structures that are sufficiently representative of the whole substrates for a specific enzyme, with the best ones showing AUC values that range from 0.749 to 0.916.

The effectiveness of using single structures in similarity searching prompted us to use the combination of the Tanimoto scores of the best three reference structures into a new consensus Tanimoto score, $T_{\text{consensus}}$. The results of this analysis showed that by using multiple reference structures we achieved an improvement in both the area under the curve and the recall performance at the beginning of the list.

Our strategy was validated with an external data set of substrates using the best single reference structures and their combination. The performance, quantified by the AUC values, to discriminate between substrates and nonsubstrates was at the same level or superior to the results obtained during the development of the model.

Herein, we prefer not to establish threshold values for the Tanimoto coefficient to be used to predict the predominant isoform, because this might be the user's choice and resource. For instance, a conservative attitude (privileging specificity over sensitivity by requiring a high score) allows the majority

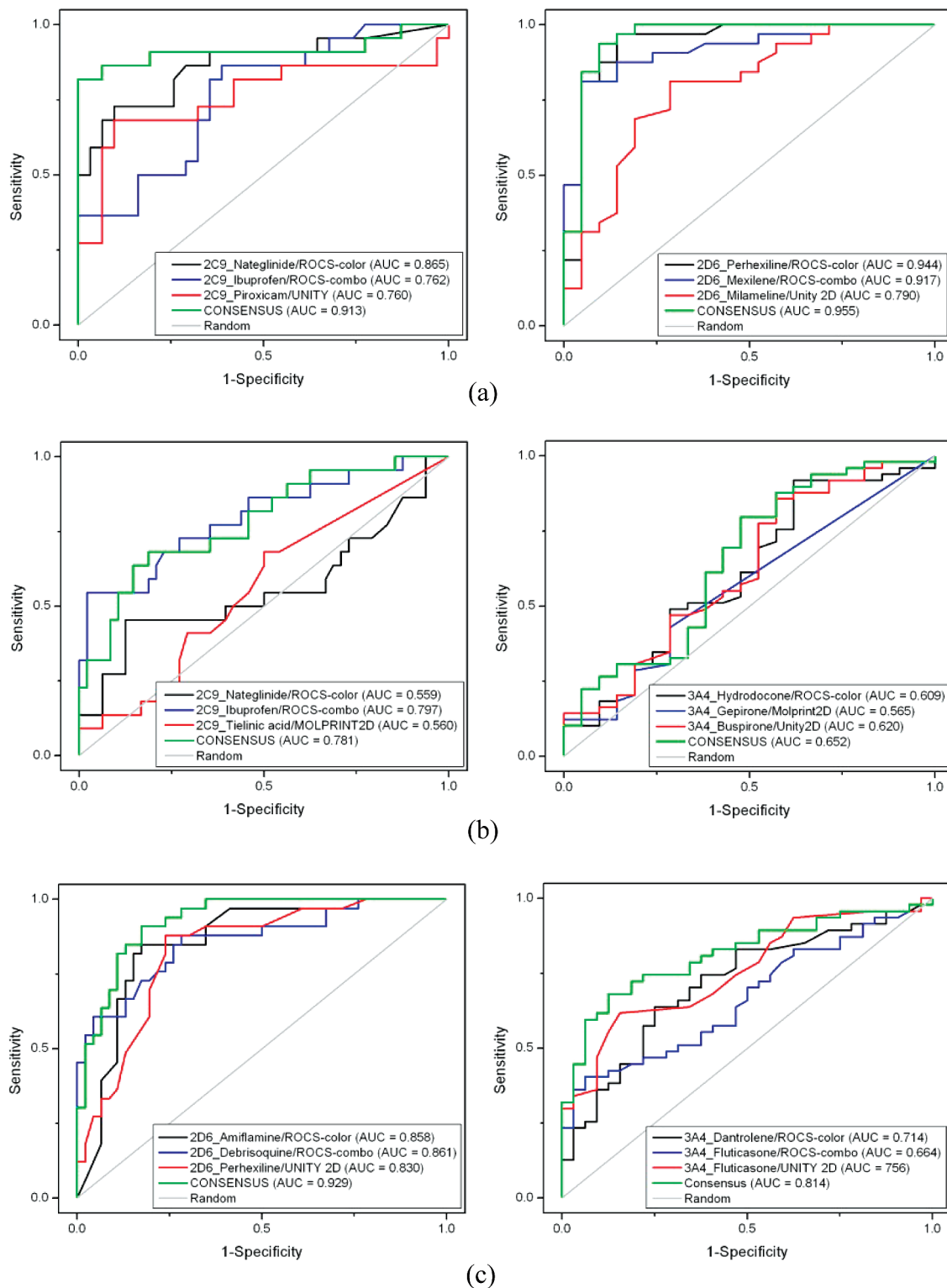


Figure 5. ROC curves for the validation data set: (a) CYP2C9-CYP2D6; (b) CYP2C9-CYP3A4; and (c) CYP2D6-CYP3A4.

of inactives (nonsubstrates) to be left aside. The liberal strategy (privileging sensitivity over specificity by requiring a high score) is a way to account for the uncertainty of models (see Triballeau et al.²⁹).

Overall, we believe that the approach presented here, which uses different similarity search methods with multiple reference structures, will provide a powerful strategy in the identification of potential substrates for the CYP450 enzymes. We can also foresee its application on chemogenomics,

which entails the identification of compounds to be efficient in gene families.

EXPERIMENTAL SECTION

Data Set Preparation. All compounds were subjected to a common preparation procedure, which included the following: addition of hydrogen atoms, determination of the most probable protonation state at pH 7.4, and generation

of up to three low energy conformers via the LigPrep modules of the Schrödinger software suite.³⁵

Before running ROCS, the program OMEGA v.2.1 was used to convert all compounds to 3D multiconformer structures.³⁶ The algorithm implemented in OMEGA dissects the molecules into fragments, reassembles and regenerates many possible combinations, and then submits each conformer to a simplified energy evaluation. Next, all conformers below an energy threshold are compared, and those within a certain rms distance are clustered into one single representation. Default parameters were used with the following exceptions: (1) *buildff* (force field used for model construction and torsional search), this parameter was set as *mmff94s* (default = *mmff94s_NoEstat*); (2) *maxconfs* (sets the maximum number of conformations to be generated), this parameter was set to 500 (default = 400); and (3) *rms* (sets the minimum root mean square (rms) distance below which two conformers are duplicates), this parameter was set to 0.6 (default = 0.8).

ROCS Calculations. A single low-energy conformation of each substrate in the database was used as the reference structure for ROCS.^{27,28} In default operation, ROCS compares molecules purely on the basis of their best shape overlap, quantitated by their shape Tanimoto. It was quickly found that adding to the shape Tanimoto the score for the appropriate overlap of groups with comparable properties (donor, acceptor, hydrophobes, cation, anion, and ring), the so-called color score, and then ranking on this summed score improved virtual screening performance considerably. In this mode, ROCS optimizes the molecular overlay to maximize both the shape overlap and the color overlap obtained by aligning groups with the same properties that are contained in the color force field file. This overlay is then subsequently scored using the sum of shape Tanimoto for the overlay and the color score (the so-called combo score). The resulting database of pairwise CYP450 substrates, generated by OMEGA, was initially screened and scored using the ROCS algorithm to generate and score the 3D overlays of the database molecules. ROCS measures the shape similarity of two compounds by using the Tanimoto coefficient, which can vary from 0.0 to 1.0, with 1.0 representing an exact match.

Molprint2D Calculations. First, a hydrogen-depleted molecular structure for every substrate was constructed, and all heavy atoms had its Sybyl atom types assigned. Second, an all atom environment fingerprint was calculated, using distances from 0 to 2 bonds (from 0 to 3 bonds, in the case of the CYP2C9–CYP2D6 analysis), using the *mol22aefp.pl* pearl script. Next, the similarity of the compounds in the database was assessed to each reference structure using the *tanimoto.pl* pearl script.

Unity 2D Calculations. A sybyl database was constructed for each pairwise database. A single (optimized) conformation of each molecule, in the sybyl mol2 format, was put in the corresponding sybyl database. Next, a tripos molecular spreadsheet was created for each one of the three pairwise databases. We then used the sybyl script *tan_columns* to create a Tanimoto distance array using Unity 2D fingerprints. The Unity 2D fingerprints column is created on the fly by this script.

ROC-AUC Analysis. In a database consisting of substrates of two CYP isoform, when a set was assigned as

substrate (set as 1 in the spreadsheet) of an enzyme, the other was identified as nonsubstrate (set as 0 in the spreadsheet) for the same enzyme and vice versa. For instance, to perform the ROC-AUC analysis in relation to the CYP2C9, the substrates of the CYP2D6 were assigned as nonsubstrate for the former enzyme. The opposite strategy was adopted when the analysis was with regard to the CYP2D6. The ROC graph and the calculation of the AUC values were performed using the ROC module of the SigmaPlot software.³⁷

ACKNOWLEDGMENT

We are grateful to the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for financial support and scholarships.

Supporting Information Available: Excel spreadsheet with AUC values of all cytochrome P450 substrates and the names of the substrates used in the external validation data set; Word document with all Tanimoto matrices for the selective pairwise comparisons and the ROC curves for the validation data set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Nigsch, F.; Mitchell, J. B. O. How to winnow actives from inactives: introducing molecular orthogonal sparse bigrams (MOSBs) and multiclass winnow. *J. Chem. Inf. Model.* **2008**, *48*, 306–318.
- (2) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.
- (3) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- (4) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates. *J. Med. Chem.* **2004**, *3*, 711–716.
- (5) Tillement, J. P.; Tremblay, D. Clinical Pharmacokinetic Criteria for Drug Research. In *Comprehensive Medicinal Chemistry II: ADME-Tox Approaches*, 1st ed.; Trigg, D. J., Taylor, J. B., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Vol. 5, pp 11–30.
- (6) Wienkers, L. C.; Heath, T. G. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825–833.
- (7) Lill, M. A.; Dobler, M.; Vedani, A. Prediction of small-molecule binding to cytochrome P450 3A4: Flexible docking combined with multidimensional QSAR. *ChemMedChem* **2006**, *1*, 73–81.
- (8) Iyer, R.; Zhang, D.; Zhang, D. Role of Drug Metabolism in Drug Development. In *Drug Metabolism in Drug Design and Development: Basic Concepts and Practice*, 1st ed.; Zhang, D., Zhu, M., Humphreys, W. G., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2008; Vol. 1, pp 261–285.
- (9) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting Drug Metabolism: A site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313–2324.
- (10) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (11) Bonn, B.; Masimirembwa, C. M.; Aristei, Y.; Zamora, I. The molecular basis of CYP2D6-mediated N-dealkylation: balance between metabolic clearance routes and enzyme inhibition. *Drug Metab. Dispos.* **2008**, *36*, 2199–2210.
- (12) Schuster, D.; Laggner, C.; Steindl, T. M.; Langer, T. Development and validation of an in silico P450 profiler based on pharmacophore models. *Curr. Drug Discovery Technol.* **2006**, *3*, 1–48.
- (13) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173–3184.
- (14) Ekins, S.; Bravi, G.; Wikel, J. H.; Wrighton, S. A. Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 424–443.
- (15) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

- (16) Terfloth, L.; Bienfait, B.; Gastaiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.
- (17) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.
- (18) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- (19) Pavlidis, P.; Nobel, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **2003**, *19*, 295–296.
- (20) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.
- (21) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
- (22) Sing, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J. Med. Chem.* **2003**, *46*, 1330–1336.
- (23) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. J. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
- (24) Maréchal, J. D.; Kemp, C. A.; Roberts, G. C. K.; Paine, M. J. I.; Wolf, C. R.; Sutcliffe, M. J. Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. *Br. J. Pharmacol.* **2008**, *153*, S82–S89.
- (25) Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2d6 catalyzed n-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (26) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. L. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719–729.
- (27) Mcgaughey, G. B.; Sheridan, R. P.; Bayly, C. L.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- (28) Kojec, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201–1213.
- (29) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- (30) Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 617–640.
- (31) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
- (32) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (33) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic site prediction and virtual screening of Cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49*, 2417–2430.
- (34) Cytochrome P450 Drug Interaction Table. <http://medicine.iupui.edu/clinpharm/ddis/table.asp>; accessed June 10, 2009.
- (35) *LigPrep, version 2.0*; Schrödinger, LLC: New York, NY, 2005.
- (36) Böstrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of omega with respect to retrieving bioactive conformations. *J. Mol. Graphics Model.* **2003**, *21*, 449–462.
- (37) *SigmaPlot, version 10.0*; Systat Software Inc.: San Jose, CA, 2008.

CI900074T