Sérgio Amorim de Alencar

Utilização de ferramentas computacionais para o estudo do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas

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

Orientador: Prof. Julio Cesar Dias Lopes

Belo Horizonte
Instituto de Ciências Biológicas
Universidade Federal de Minas Gerais
2010

Alencar, Sérgio Amorim de
Utilização de ferramentas computacionais para o estudo do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas. [manuscrito] / Sérgio Amorim de Alencar. - 2010.
113 f. : il. ; 29,5 cm.

Orientador: Julio Cesar Dias Lopes.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Farmacogenética - Teses. 2. Modelagem molecular – Teses. 3. Proteínas – Teses. 4. Bioinformática – Teses. 5. Proteínas – estrutura – Teses. 6. Polimorfismo de um único nucleotídeo. 7. Receptor IGF tipo 1. I. Lopes, Julio Cesar Dias. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 577.112:004

Universidade Federal de Minas Gerais -
Programa de Doutorado em Bioinformática/UFMG
Av. Antônio Carlos, 6627 – Pampulha
31270-901 – Belo Horizonte – MG
e-mail: bioinfo@icb.ufmg.br (31)3409-2554

ATA DA DEFESA DA TESE DE DOUTORADO DE SÉRGIO AMORIM DE ALENCAR. Aos vinte e cinco dias do mês de Junho de 2010 às 13h30min, reuniu-se no Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais a Comissão Examinadora da tese de doutorado, indicada durante a reunião de treze de maio de 2010 do Colegiado do Curso, para julgar, em exame final, o trabalho intitulado "Utilização de ferramentas computacionais para o estudo do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas" requisito final para a obtenção do grau de Doutor em Ciências, Área de Concentração: Bioinformática. Abrindo a sessão o Presidente da Comissão, Prof. Júlio César Dias Lopes da Universidade Federal de Minas Gerais, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores, com a respectiva defesa do candidato. Logo após a Comissão se reuniu sem a presença do candidato e do público para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações: Dr. Emmanuel Dias-Neto do Centro de Pesquisas do Hospital AC Camargo – São Paulo, SP, aprovado; Dr. Walter Filgueira de Azevedo Júnior da Pontifícia Universidade Católica do Rio Grande do Sul – PUC-RS, Porto Alegre, RS, aprovado; Drª Raquel Melo Minardi da Universidade Federal de Minas Gerais, aprovado; Dr. Eduardo Martin Tarazona Santos da Universidade Federal de Minas Gerais, aprovado; Dr. Júlio César Dias Lopes, orientador, da Universidade Federal de Minas Gerais, aprovado. Pelas indicações o candidato foi considerado APROVADO. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar o Presidente da Comissão encerrou a reunião e lavrou a presente ata que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, aos vinte e cinco dias de Junho de 2010.

Dr. Emmanuel Dias-Neto – Centro de Pesquisas do Hospital AC Camargo/SP

Dr. Walter Filgueira de Azevedo Júnior – PUC-RS

Drª Raquel Melo Minardi – UFMG

Dr. Eduardo Martin Tarazona Santos - UFMG

Dr. Júlio César Dias Lopes – orientador – UFMG

"Ao lado da música e da arte, a ciência é a maior, mais bela e mais iluminadora das conquistas do espírito humano."

Karl Popper

Aos meus pais

AGRADECIMENTOS

- Aos meus pais, Danton e Maria, meus irmãos Marcos e Emerson, à Cláudia, Júlia e Bianca, muito obrigado por todo carinho, e por estarem sempre ao meu lado
- à FAPEMIG pelo financiamento do projeto, e à coordenação do Programa de Pós-Graduação em Bioinformática pelo auxílio a congressos e cursos
- A todos os funcionários que passaram pela secretaria da Bioinformática, em especial ao Carlos
- aos colegas do laboratório 288: Andrelly, Eduardo, Ramon, Henrique, Julio, e Bernardo
- Ao Prof. Julio Lopes, pela orientação do meu doutorado
- Aos colegas da Bioinformática: Rodrigo, Adhemar, Cris, Caio, Valdete, Calouro, Cécile, Bráulio, Wagner, Deive, Priscila

RESUMO

Os polimorfismos de base única (SNPs) são a forma mais comum de variação na sequência de DNA entre humanos, e têm o potencial de afetar a função gênica, principalmente quando estão localizados em regiões codificadoras ou regulatórias. Dentre os diferentes tipos de SNPs, acredita-se que os SNPs não-sinônimos (nsSNPs) têm o maior impacto na função protéica, sendo frequentemente associados a doenças, alterações na resposta a fármacos, e a reações adversas. A motivação deste trabalho é o fato de que uma abordagem computacional pode ter grande utilidade na avaliação preliminar do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas em humanos, possibilitando assim a priorização de nsSNPs candidatos para estudos experimentais. Com este propósito, fizemos a modelagem de nsSNPs nas correspondentes estruturas protéicas nativas como codificadas pelos genes, buscando determinar o impacto causado por estas variações utilizando diferentes métodos computacionais, tais como o *docking* molecular e a otimização de estruturas protéicas. Um banco de dados foi montado, relacionando os resultados das análises computacionais feitas com informações já existentes, tais como de doenças, vias metabólicas, alvos terapêuticos, fármacos, enzimas metabolizadoras de fármacos, e anotações de sequências protéicas, possibilitando a integração de resultados obtidos por diferentes métodos utilizados no estudo do impacto de nsSNPs na função protéica.

ABSTRACT

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation between humans, and have the potential to affect gene function, especially when they are located in coding or regulatory regions. Among the many types of SNPs, non-synonymous SNPs (nsSNPs) are believed to have the greatest impact on protein function, often being associated to diseases, changes in drug response, and adverse drug reactions. The motivation of this work was the fact that a computational approach could be highly useful in the preliminary evaluation of the functional and structural impact of nsSNPs in protein encoding genes in humans, hence enabling the prioritization of candidate nsSNPs for experimental studies. For this purpose, nsSNP modeling was carried out in their corresponding native protein structures as coded by their genes, aiming to determine the impact caused by these variations using different computational methods, such as molecular docking and protein structure optimization. A database was built, relating results data from the computational analysis carried out with information which already exist, such as disease, metabolic pathways, drug targets, drugs, drug metabolizing enzymes, and protein sequence annotations, enabling the integration of results obtained by different methods used in the study of the impact of nsSNPs on protein function.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADT	AutoDock Tools
BLAST	Basic Local Alignment Search Tool
CNVs	Variação no número de cópias
dbSNP	Single Nucleotide Polymorphism Database
DeepView	Swiss-PdbViewer
DNA	Deoxyribonucleic acid
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
IGF1R	Insulin-Like Growth Factor 1 Receptor
INDEL	Inserção e Deleção
LMC	Leucemia Mielóide Crônica
NCBI	National Center for Biotechnology Information
NCS	NEQUIM Contact System
nsSNP	Non-synonynoums Single Nucleotide Polymorphism
PDB	Protein Data Bank
PGH	Projeto Genoma Humano
PolyPhen	Polymorphism Phenotyping
PSIC	Position-Specific Independent Counts
PSSM	Position-Specific Scoring Matrix
RDBMS	Relational Database Management Systems
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
SIFT	Sorting Intolerant from Tolerant
SNP	Single Nucleotide Polymorphism
sSNP	Synonynoums Single Nucleotide Polymorphism

SUMÁRIO

1. INTRODUÇÃO

1.1 Conceitos Básicos

1.1.1 DNA e RNA

O DNA e o RNA (ácido desoxirribonucléico e ácido ribonucléico, respectivamente) são substâncias químicas envolvidas na transmissão de caracteres hereditários, regulação da expressão gênica e síntese de proteínas em humanos.

O DNA é uma molécula formada por duas cadeias (ou fitas) na forma de uma dupla hélice (Fig. 1). Cada fita consiste de um arranjo linear de unidades químicas básicas chamadas nucleotídeos, que consistem de uma molécula de açúcar (desoxirribose) e uma de fosfato ligadas a uma das quatro bases nitrogenadas – Adenina (A), Guanina (G), Citosina (C), e Timina (T). Uma fita simples de DNA pode ter qualquer sequência dessas quatro letras [Griffiths et al., 1998].

Dada a sequência de letras de uma fita de DNA, podemos saber qual sequência de nucleotídeos a outra fita deve ter, pois os nucleotídeos formam pares complementares (A sempre forma par com T, e G sempre forma par com C). Isto é um fator essencial na replicação do DNA durante a divisão celular, onde cada fita serve de molde para a geração de uma nova fita [Watson e Crick, 1953].

O RNA é formado apenas por uma cadeia de ribonucleotídeos que, por sua vez, são compostos por uma molécula de açúcar (ribose), um grupo fosfato, e uma das quatro bases nitrogenadas (uracila, no entanto, ao invés de timina). Os principais tipos de RNA são os RNAs mensageiros (RNAm), os transportadores (RNAt), os ribossomais (RNAr), os microRNAs (RNAmi), e os RNAs nucleares pequenos (RNAsn) [Griffiths et al., 1998; Bartel, 2009].

1.1.2 Código genético e síntese protéica

A informação genética, armazenada nos cromossomos e transmitida às células filhas através da replicação do DNA, é expressa através da transcrição em RNA e, no caso de RNAm, tradução subsequente em cadeias polipeptídicas. Este fluxo de informação do DNA ao RNA e à proteína é denominado de “dogma central” da biologia molecular. O processo de síntese protéica requer um código genético, através do qual as informações contidas em janelas abertas de leitura (ORFs) nos

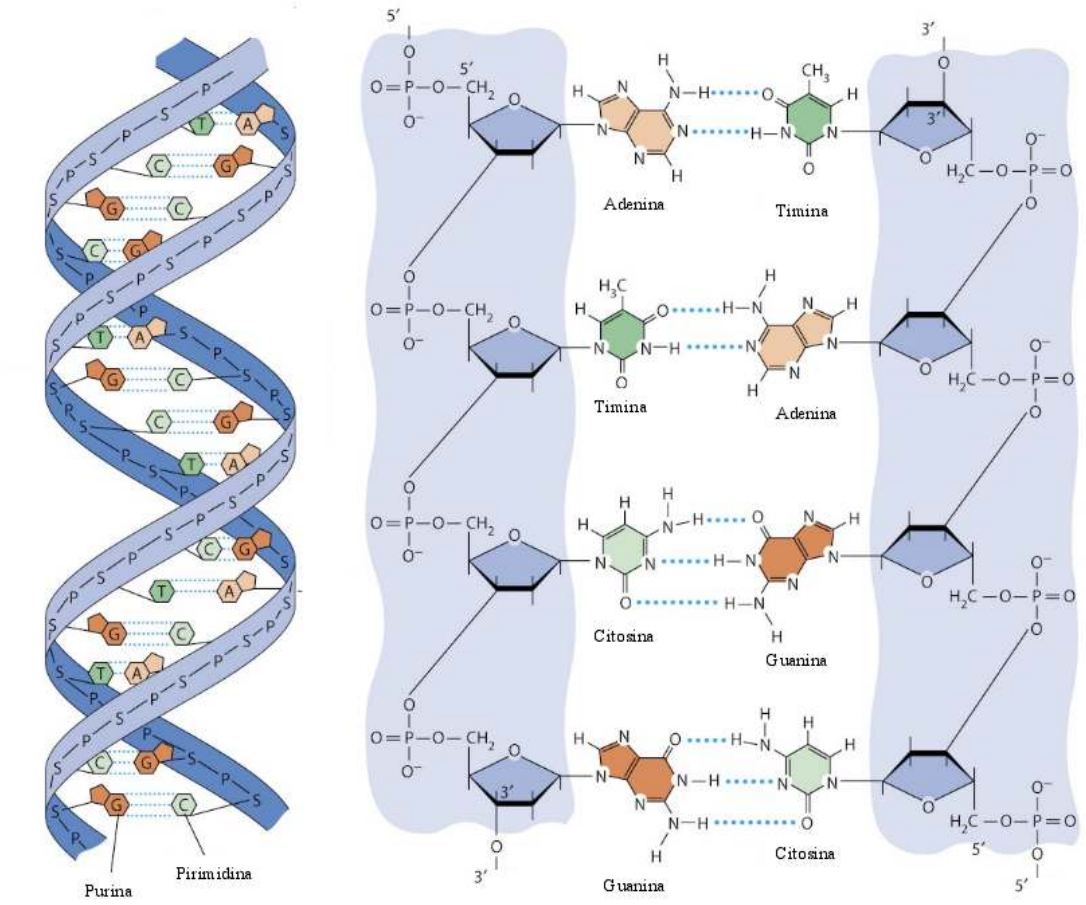


Figura 1. Esquerda: Um modelo simplificado mostrando a estrutura helicoidal do DNA. Direita: A dupla hélice do DNA em forma plana, para mostrar os filamentos com a sequência desoxirribose e os degraus de pares de base. Cada par de bases tem uma purina (adenina (A) ou guanina (G)), e uma pirimidina (timina (T) ou citosina (C)) conectadas por ligações de hidrogênio (pontilhados).

genes são expressas para produzir uma sequência específica de aminoácidos pelo processo de tradução. A ligação molecular entre estes dois tipos relacionados de informação (o código de DNA dos genes e o código de aminoácidos das proteínas) é o RNA [Griffiths et al., 1998].

O código genético consiste em códons, cada um composto por uma trinca de bases nitrogenadas (tripletes) (Fig. 2). Dos 64 códons possíveis, três indicam o término da região de tradução do gene, e são conhecidos como códons finalizadores (ou sem sentido): UAA, o UGA e o UAG. Os outros 61 especificam aminoácidos. Como existem apenas 20 aminoácidos essenciais (Fig. 3), isto significa que a maioria dos aminoácidos pode ser especificada por mais de um códon. Por exemplo, a leucina e a

arginina são especificadas por seis códons. Apenas a metionina e o triptofano são cada um deles especificado por um único códon. O código genético é, portanto, dito "redundante" (ou degenerado). Embora um determinado aminoácido possa ser especificado por mais de um códon, cada códon só pode designar um aminoácido [Griffiths et al., 1998].

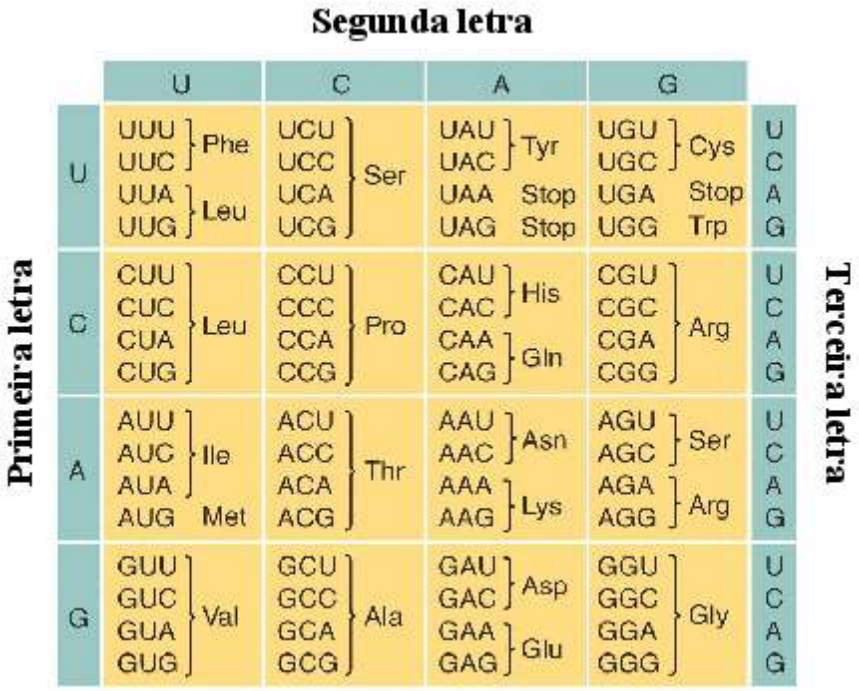


Figura 2. A informação genética é estocada no DNA por meio do código genético, no qual a sequência de bases adjacentes determina a sequência de aminoácidos no polipeptídeo codificado.

1.1.3 Forças interatômicas não-covalentes nas proteínas

Todas as proteínas que compõem o nosso organismo são constituídas por sequências de resíduos de aminoácidos ligados covalentemente. Estes resíduos possuem grupos capazes de formar interações não-covalentes entre si, e com outras moléculas. Estas interações não são tão fortes quanto as ligações covalentes, mas são muito importantes, sendo altamente responsáveis pelo enovelamento e estabilidade correta das estruturas protéicas [Stryer, 1999]. As forças interatômicas não-covalentes podem ser classificadas em vários tipos, dentre eles as forças de Van der Waals, as ligações de hidrogênio, as ligações iônicas, e as interações hidrofóbicas, que serão descritas a seguir.

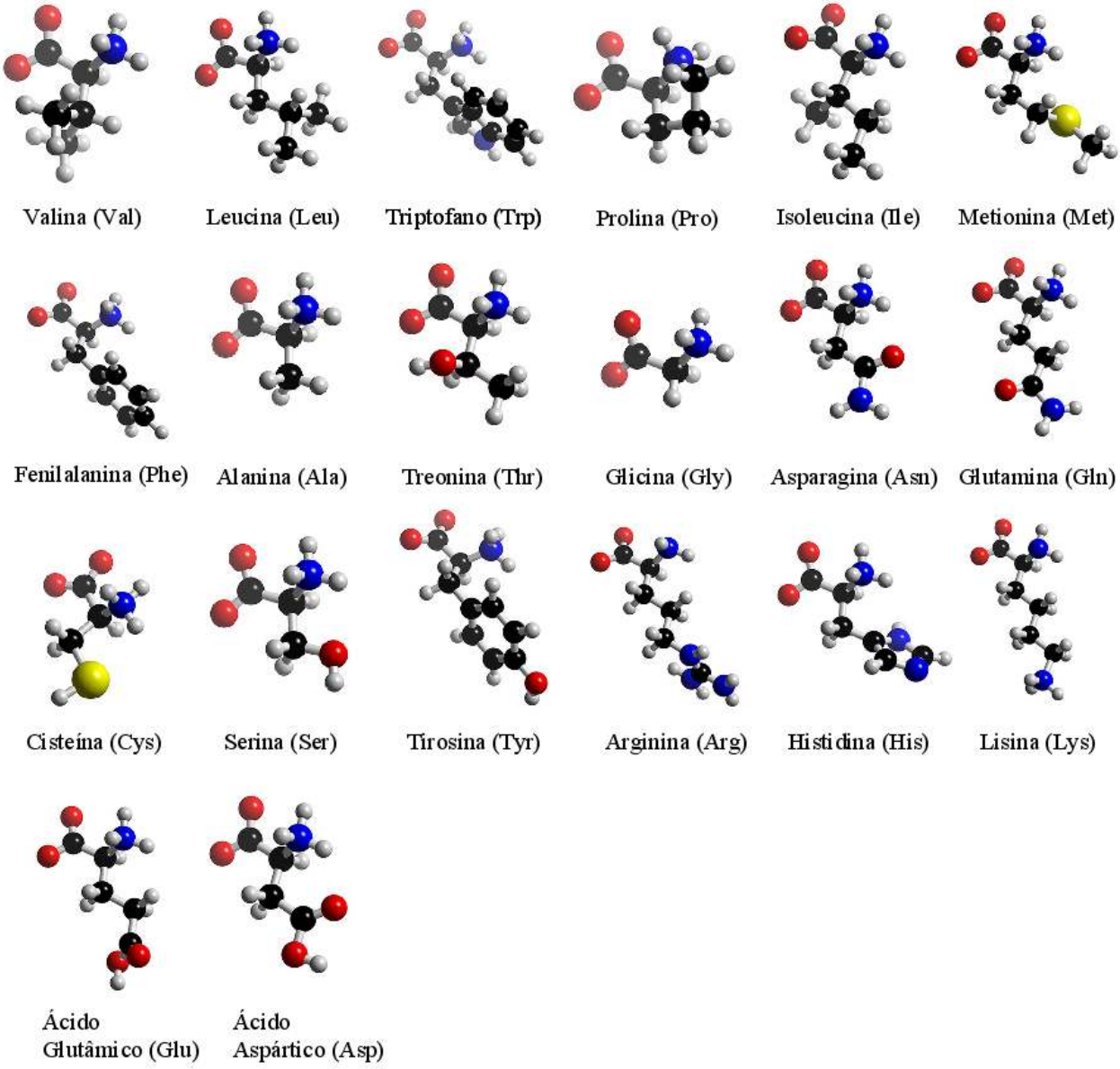


Figura 3. Os vinte aminoácidos essenciais que compõem as proteínas em humanos.

Forças de van der Waals

As forças de van der Waals podem ser divididas em três tipos, conforme a natureza das partículas. Em primeiro, certas moléculas, embora eletricamente neutras, podem possuir um dipolo elétrico permanente. Devido a alguma distorção na distribuição da carga elétrica, um lado da molécula é ligeiramente mais "positivo" e o outro é ligeiramente mais "negativo". A tendência é que estas moléculas se alinhem, e interajam umas com as outras, por atração eletrostática entre os dipolos opostos. Esta interação é chamada de dipolo-dipolo.

Em segundo, a presença de moléculas que têm dipolos permanentes pode distorcer a distribuição de carga elétrica em outras moléculas vizinhas, mesmo as que não possuem dipolos (apolares), através de uma polarização induzida. Esta interação é chamada de dipolo-dipolo induzido.

Em terceiro, mesmo em moléculas que não possuem momento de dipolo permanente existe uma força de atração. Nestas moléculas, em um determinado instante, o centro de carga negativa dos elétrons e carga positiva do núcleo atômico pode não coincidir. Esta flutuação eletrônica pode transformar as moléculas apolares em dipolos tempo-dependentes, podendo induzir a polarização das moléculas adjacentes, resultando em forças atrativas. Estas forças são conhecidas como forças de dispersão (ou forças de London), e estão presentes em todas as moléculas apolares e, algumas vezes, mesmo entres moléculas polares [Israelchvili, 1992].

Ligações de hidrogênio

Algumas moléculas exibem um tipo especial de interação dipolo-dipolo chamada de ligação de hidrogênio, que é a mais intensa de todas as forças intermoleculares, e que constitui uma das forças de estabilização mais importantes na estrutura das proteínas. Estas interações surgem quando dois grupos polares de tipos específicos interagem. Um deve ser um doador de hidrogênio, um grupo químico em que um átomo de hidrogênio é covalentemente ligado a um átomo bastante eletronegativo, como o oxigênio. A ligação entre o hidrogênio e o átomo eletronegativo é polarizada, fornecendo ao hidrogênio uma carga elétrica parcialmente positiva e ao átomo eletronegativo uma carga parcialmente negativa. O outro grupo deve ser um aceptor de hidrogênio, um átomo eletronegativo com uma carga parcialmente negativa. O hidrogênio positivamente polarizado no primeiro grupo é atraído para o segundo grupo negativamente polarizado [Israelchvili, 1992].

Ligações iônicas (interações eletrostáticas)

Outro tipo de força extremamente importante são as ligações iônicas (ou interações eletrostáticas). Estas interações ocorrem devido ao fato de que grupos carregados positivamente nas cadeias laterais dos resíduos de aminoácidos podem interagir com grupos carregados negativamente. Cerca de dois terços dos resíduos de aminoácidos com cargas nas proteínas formam pares iônicos [Israelchvili, 1992].

Efeito hidrofóbico

O efeito hidrofóbico é bastante importante para o enovelamento e a estabilidade da estrutura enovelada das proteínas. Este efeito resulta da tendência das cadeias laterais hidrofóbicas (eg. alanina, isoleucina, leucina, fenilalanina, e valina) de serem atraídas umas pelas outras para se agruparem em áreas específicas e definidas para minimizar seus contatos com a água. Quando circundados por moléculas de água, os grupos hidrofóbicos são induzidos a se unir para ocupar o menor volume possível. Assim, as moléculas de água altamente ordenadas são liberadas do interior da proteína, aumentando a desordem do sistema (entropia). O aumento da entropia é termodinamicamente favorável e dirige o enovelamento protéico [Israelchvili, 1992].

1.1.4 O mecanismo de ação dos fármacos

Para que os fármacos façam de fato efeito na fisiologia do organismo, eles precisam interagir com áreas-alvo específicas, também denominadas alvos (ou receptores) terapêuticos. As moléculas dos fármacos formam ligações químicas (geralmente interações atômicas não-covalentes – ver seção 1.1.3) com os receptores e a força dessas ligações é determinante para a afinidade do receptor pelo fármaco. Portanto, em suas conformações ativas, as moléculas do fármaco e do receptor exibem complementaridade geométrica e química, as quais são essenciais para o sucesso do tratamento [Schellack, 2005].

A formação do complexo entre um fármaco e um receptor biológico pode ser vista como a soma de várias contribuições energéticas que, por sua vez, podem ser favoráveis ou desfavoráveis à interação [Böhm, 1994]. A formação de tal complexo é favorecida pela diminuição na Energia Livre de Gibbs (ΔG) do sistema [Perrot, 1998], que se relaciona com a constante de equilíbrio do processo de formação do complexo pela seguinte relação:

$$\Delta G = - RT \ln K_{eq}$$

A Energia Livre, contudo, não é facilmente avaliável, uma vez que ela envolve o componente entrópico, ΔS, para o qual os modelos estabelecidos são complicados e nem sempre precisos:

$$\Delta G = \Delta H - T \Delta S$$

A energia envolvida na formação do complexo (ΔE) entre um fármaco e seu receptor pode ser avaliada quantitativamente como a soma de várias contribuições:

$$\Delta E = \Delta E_{elet} + \Delta E_{pol/dis} + \Delta E_{lig.H} + \Delta E_{tc} + \Delta E_{hf} + \Delta E_{vdw}$$

onde ΔE_{elet} representa a contribuição das ligações de caráter eletrostático (íon-íon, íon-dipolo, ou dipolo-dipolo), $\Delta E_{pol/dis}$ refere-se aos efeitos de polarização e dispersão, $\Delta E_{lig.H}$ às ligações hidrogênio, ΔE_{tc} é a energia referentes à formação dos complexos de transferência de carga, ΔE_{hf} às interações hidrofóbicas e ΔE_{vdw} representa a energia das forças de van der Waals e de dispersão de London.

1.2 O Genoma Humano

Genoma é o nome dado ao conjunto de todo o DNA de todos os cromossomos de um gameta humano (óvulo ou espermatozóide), sendo constituído de 3,4 bilhões de bases. A sequência de bases de sua porção não-repetitiva, constituída de ~2,8 bilhões de bases, já foi completamente elucidada, com a conclusão do Projeto Genoma Humano (PGH) em 2003 [The Human Genome, 2001 e 2001b; Leite, 2003]. Um dos grandes legados do PGH foi a disponibilização dos dados obtidos para toda a comunidade científica através da construção de bancos de dados públicos, como o *National Center for Biotechnology Information* (NCBI) (http://www.ncbi.nlm.nih.gov/), o que possibilitou o desenvolvimento do presente trabalho.

O mapa atual do genoma humano tem uma precisão de aproximadamente 99,96% [Borém e Santos, 2008]. Estima-se que o genoma humano possui ~24000 genes codificadores de proteínas, um número significativamente menor do que se pensava inicialmente (50 a 140 mil genes) [International Human Genome Sequencing Consortium, 2004]. De fato, as regiões de DNA codificadoras representam uma pequena porção (~1,5%) do genoma total. Apesar de que estimativas mostrem que mais da metade do genoma humano consiste de sequências repetitivas não-codificadoras [Wolfsberg et al., 2001], estas sequências de DNA que não codificam proteínas podem codificar moléculas de RNA funcionais envolvidas na regulação da expressão gênica [Lander et al., 2001; Birney et al., 2007]. Além disso, algumas sequências não-codificadoras de DNA têm um papel estrutural nos cromossomos, que é o caso dos centrômeros e telômeros, que são regiões de baixa frequência gênica, mas que são

importantes para a estabilidade dos cromossomos [Pidoux et al., 2005].

Os dados do PGH revelaram, ainda, que cada ser humano, independentemente das suas diferenças aparentes, possui alta similaridade no seu material genético com o de outro indivíduo qualquer, sendo que as diferenças genéticas ocorrem devido à existência de polimorfismos genéticos no genoma humano.

1.3 Polimorfismos Genéticos

A mutação é um processo de mudança genética na estrutura do genoma geralmente causado por um erro na duplicação do DNA, podendo ter consequências deletérias, benéficas ou neutras para o organismo. Diferentes versões de uma certa sequência de DNA em um determinado local cromossômico (*locus*) são chamados de alelos. Qualquer *locus* no qual existam alelos múltiplos como componentes estáveis da população (na qual estão presentes em uma frequência maior do que 1%) é geralmente definido como polimórfico [Lewin, 2001; Nussbaum et al., 2002; Kirk et al., 2002].

As formas mais comuns de polimorfismos genéticos em humanos são inserções, deleções, inversões, duplicações, polimorfismos de base única (SNP – *Single Nucleotide Polymorphisms*) (~1% de todo o genoma), variações no número de sequências repetidas (VNTR – *Variable Number of Tandem Repeats*), variações no número de cópias (CNVs) (~5% de todo o genoma), microsatélites e minisatélites [Wright, 2003; Tuzun et al., 2005; Feuk et al., 2005].

Os SNPs caracterizados pela substituição de uma base nucleotídica por outra na sequência do DNA (para diferenciá-los das inserções e deleções de base única, ou *indels*, que também são caracterizados como SNPs) podem surgir por dois processos: incorporação incorreta de base durante a replicação de DNA, e modificação química *in situ* de uma base. Como os mecanismos celulares de correção de bases não emparelhadas são extremamente eficazes, é necessário entender como estes eventos de substituição progridem de uma substituição na sequência, que é prontamente editada de volta para a base correta, para se tornar alélica [Phillips, 2007].

O primeiro processo é um evento extremamente raro em DNA genômico, devido ao alto grau de fidelidade de replicação da enzima DNA Polimerase e a um sistema elaborado de edição de bases incorporadas incorretamente [Nachman e Crowell, 2000]. Consequentemente, o processo de modificação *in situ* deve explicar o aparecimento da maioria dos SNPs, o que pode ser visto nas regiões

de DNA que sofrem a metilação (como as regiões CpG), onde a citosina metilada pode sofrer a desaminação para formar uma timina estável. Isto pode explicar o fato de que a grande maioria dos SNPs compreendem substituições C-T ou A-G [Phillips, 2007].

Um SNP pode ser sinônimo ou não-sinônimo: no primeiro caso (também conhecido como sSNP), o aminoácido codificado pelo códon que contém o SNP é o mesmo que aquele codificado pelo códon sem o SNP; e no segundo caso (também conhecido como nsSNP), o códon modificado codifica um resíduo de aminoácido diferente daquele codificado pelo códon sem o SNP (Fig. 4). As variações mais frequentes dos SNPs são substituições entre bases nucleotídicas de mesma característica estrutural (A/G ou G/A e C/T ou T/C), que são chamadas de transições. As outras substituições são conhecidas como transversões [Kiewitz e Tummler, 2002].

Atualmente, existem cerca de 24 milhões de registros de SNPs humanos depositados na base de dados pública dbSNP (build 129) (http://www.ncbi.nlm.nih.gov/SNP). Considerando apenas os registros não-redundantes, estima-se que existam mais de 1,4 milhões de SNPs no dbSNP, dos quais mais de 90,000 são não-sinônimos [Ryan et al., 2009].

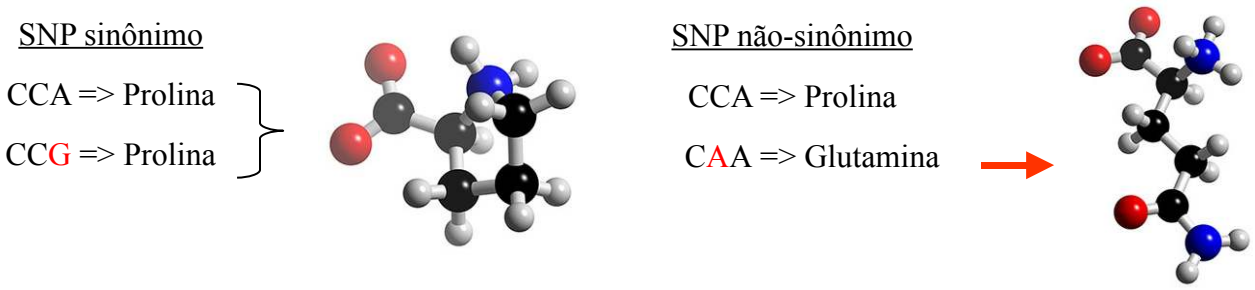


Figura 4. De acordo com o código genético, um certo aminoácido pode ser determinado por mais de um códon. Na figura à esquerda, a terceira base de um dos códons que codifica o aminoácido Prolina (CCA) foi substituída, criando outro códon que também codifica o aminoácido Prolina. SNPs deste tipo, que não provocam mudanças na sequência de aminoácidos da cadeia polipeptídica são chamados SNPs sinônimos (sSNPs). Na figura à direita, a substituição da segunda base de um dos códons que codifica o aminoácido Prolina (CCA) criou um códon que codifica o aminoácido Glutamina. SNPs deste tipo são conhecidos como SNPs não sinônimos (nsSNPs).

A raridade dos nsSNPs pode ser uma consequência de pressões seletivas, pois os nsSNPs são observados menos frequentemente na população humana do que esperado pela taxa de mutação, o que é evidência de que estão sob forte seleção purificadora. Especificamente, se uma mutação aleatória ocorresse em uma região codificadora do gene, ela deveria levar a uma mudança de aminoácidos 2/3 do tempo, mas nsSNPs compreendem apenas a metade dos SNPs codificadores no genoma humano [Cargill et al., 1999].

1.4 Efeito dos nsSNPs na estrutura, função, e interação protéica

Os efeitos causados por substituições de resíduos de aminoácidos decorrentes de nsSNPs em genes codificadores de proteínas podem ser agrupados em quatro categorias distintas, apesar de que estes efeitos possam ser mutuamente dependentes:

(a) *Enovelamento protéico, estabilidade, flexibilidade e agregação*

O enovelamento protéico é um processo complexo que converte uma cadeia linear polipeptídica em uma estrutura tridimensional. Durante este processo a proteína "experimenta" uma variedade de estados intermediários seguindo o gradiente de energia [Dill et al., 1993; Dill et al., 2007]. A mudança de um resíduo de aminoácido chave poderia tornar alguns destes estados intermediários inacessíveis, ou perturbar a paisagem de energia potencial (*energy landscape*), afetando a cinética de enovelamento da proteína.

O efeito mais evidente causado por um nsSNP é na estabilidade protéica [Koukouritaki et al., 2007; Ode et al., 2007; De Cristofaro et al., 2006]. A explicação física disto pode variar desde restrições geométricas (substituição de uma cadeia lateral pequena para uma volumosa no interior da proteína), a efeitos físico-químicos (substituição de um resíduo hidrofóbico para um polar), e o rompimento de ligações de hidrogênio [Shirley et al., 1992].

Também é possível que um nsSNP não afete a estabilidade da proteína, mas que cause uma alteração na flexibilidade da proteína. É sabido que a capacidade das proteínas submeterem-se a mudanças conformacionais é essencial para suas funções [Tang e Dill, 1998; Song et al., 2005]. Uma mutação que torna a proteína muito rígida ou que afeta conformações alostéricas, pode afetar significativamente a função protéica [Song et al., 2005]. Por outro lado, uma mutação que desestabiliza

e torna a proteína muito flexível, poderia levar à agregação e a formação de fibrilas [Board et al., 1990].

(b) *Sítios funcionais e cinética de reações*

A substituição de um resíduo de aminoácido catalítico certamente afeta a função protéica [Yamada et al., 2006], e o nsSNP causador desta substituição é definido como deletério [Stevanin et al., 2004]. No entanto, como existem poucos resíduos de aminoácido catalíticos, a probabilidade de tal nsSNP ocorrer é baixa [Sunyaev et al., 2000]. Mas a reação pode ser afetada pela substituição de resíduos de aminoácidos localizados próximo a grupos catalíticos [Takamiya et al., 2002]. A substituição de tal resíduo pode não cessar completamente a reação, mas poderia alterar sua cinética [Koukouritaki et al., 2007].

(c) *Expressão protéica e localização subcelular*

Mesmo que um nsSNP não cause algum dos efeitos descritos acima, ainda assim este nsSNP poderia afetar a função protéica. A substituição de um resíduo de aminoácido na estrutura de um peptídeo sinalizador poderia resultar em uma localização subcelular deste peptídeo diferente daquela da proteína nativa que interage com o peptídeo [Tiede et al., 2006; Krumbholz et al., 2006]. Isto poderia causar uma grande redução na concentração da proteína no compartimento onde ela evoluiu para funcionar. Além disso, a presença desta proteína em um compartimento "não-desejado" poderia afetar o funcionamento de outras proteínas que ali atuam [Hanemann et al., 2000].

(d) *Interações proteína-ligante, proteína-proteína, proteína-DNA, e proteína-membrana*

Um nsSNP localizado em uma interface, ou dentro de um sítio de ligação, poderia afetar dramaticamente a ligação entre moléculas que interagem (tais como proteína-ligante, proteína-proteína, proteína-DNA, ou proteína-membrana) [Ung et al., 2006]. Isto poderia ser causado simplesmente por um efeito geométrico, como por exemplo no caso de uma cadeia lateral volumosa ser introduzida em um *pocket* de ligação estreito, podendo bloquear a entrada de um ligante no sítio ativo [van Wijk et al., 2003]. A substituição de um resíduo de aminoácido que leva a uma alteração na geometria do sítio ativo poderia afetar o reconhecimento do ligante e reduzir, ou alterar a especificidade [Rignall et al., 2002; Hardt e Laine, 2004]. Quase todas as substituições de resíduos de aminoácidos localizadas na interface de ligação afetam a ligação entre as moléculas que interagem [Ortiz et al., 1999]. A afinidade de ligação

poderia diminuir ou aumentar por causa da substituição, o que levaria a uma alteração da afinidade obtida com a proteína nativa, podendo afetar outros processos celulares [Jones et al., 2007].

Além disso, mecanismos reguladores como a ligação proteína-DNA também podem ser afetados pela presença de nsSNPs na interface destas duas moléculas [Venkatesan et al., 2007; Elles e Uhlenbeck, 2008; Wright e Lim, 2007], assim como a transdução de sinais poderia ser afetada pela presença de nsSNPs na interface proteína-membrana [Kwa et al., 2008], e também o processo de adesão celular [Kariya et al., 2003].

1.5 Influência dos nsSNPs no desenvolvimento de doenças genéticas

As doenças genéticas podem estar associadas aos nsSNPs, devido à possibilidade destes afetarem a estrutura e a função das proteínas expressas, como visto na seção anterior. No entanto, apesar de podermos encontrar na literatura um grande número de nsSNPs associados a doenças, fica cada vez mais evidente que a correlação entre genótipo e fenótipo não é direta [Hartman et al., 2001]. Assim, para muitas doenças, apenas um subconjunto de todos os nsSNPs conhecidos seguramente predizem um fenótipo [Dipple e McCabe, 2000], sendo que este pode ter outras causas, tais como CNVs, mutações pontuais, ou variações genéticas que resultam em mudanças na expressão gênica [Feuk et al., 2006].

Os nsSNPs podem contribuir para o desenvolvimento de doenças monogênicas ou doenças complexas. Doenças monogênicas seguem um padrão simples de herança Mendeliana, em que um gene pode ser o principal responsável pela patogênese, e algum (ou alguns) outro gene modificador herdado independentemente pode influenciar o fenótipo. Estas doenças são geralmente raras, mas graves.

Um exemplo clássico de doença monogênica é a anemia falciforme, a primeira doença molecular descoberta. Primeiramente estudada por Sir John Kendrew há mais de 50 anos, a anemia falciforme resulta da substituição de um único nucleotídeo que altera de ácido glutâmico para valina (GAG → GTG; Glu6Val) o códon do sexto aminoácido da globina-β. Este resíduo de aminoácido está localizado na interface entre cadeias alpha e beta, e a substituição Glu6Val reduz significativamente a solubilidade da forma desoxigenada da hemoglobina [Stryer, 1995].

Por outro lado, as doenças complexas são doenças comuns (como a hipertensão, apoplexia, doenças coronárias, câncer, etc.), geneticamente complexas, onde alelos de vários genes contribuem

para o desenvolvimento da doença. Em doenças complexas, o predomínio de algum gene específico não é perceptível, e a interação entre dois ou mais pares de alelos herdados independentemente, provavelmente influenciados por genes modificadores adicionais, resulta na doença. Além disso, nem sempre os alelos causam a doença, sendo necessária a interação com o ambiente para que a doença se desenvolva (Fig. 5) [Dipple e McCabe, 2000b].

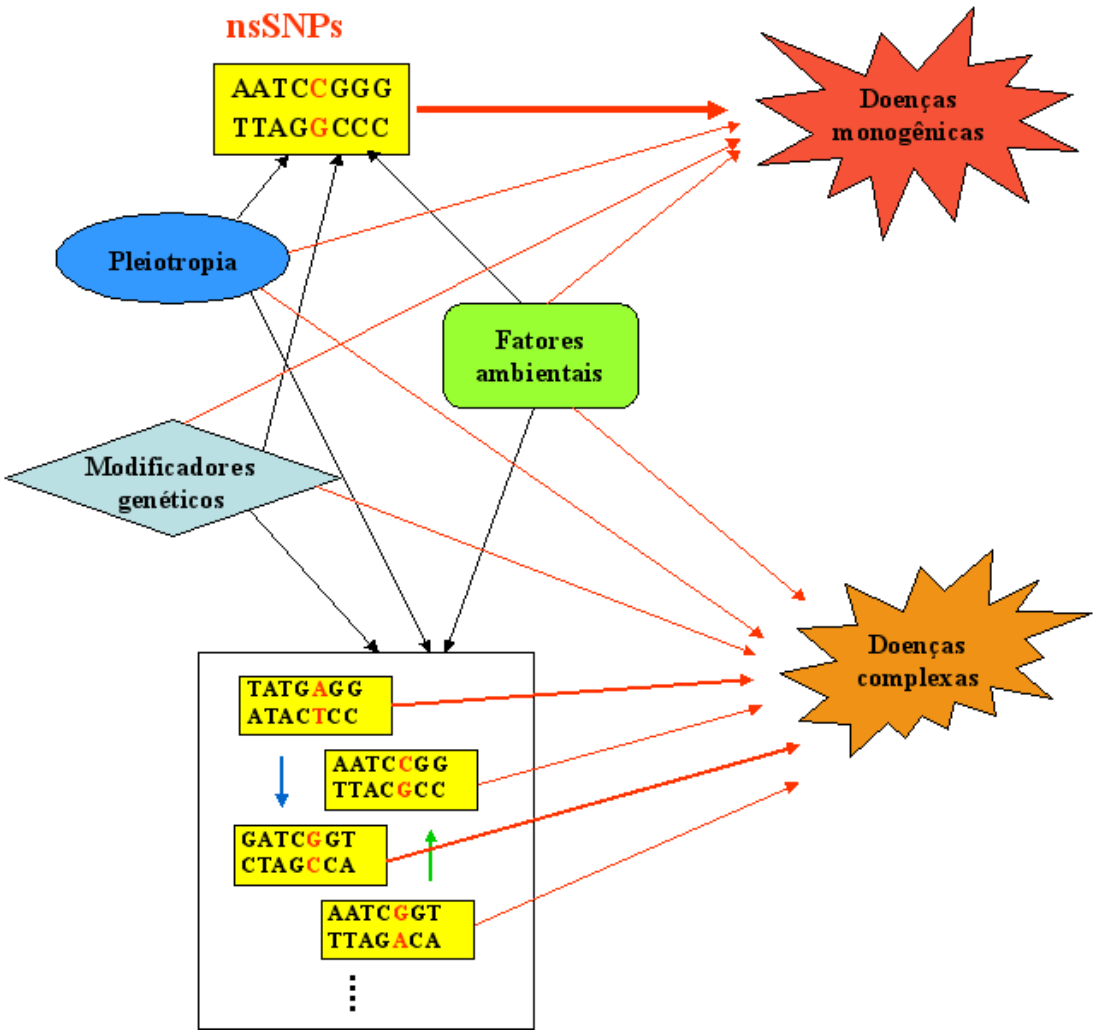


Figura 5. Representação esquemática de doenças monogênicas e complexas causadas por nsSNPs: (a) A maioria das doenças monogênicas são determinadas por mutações em um único *locus*. (b) As doenças complexas podem ser causadas por diversos nsSNPs, os quais podem afetar outros nsSNPs, potencializando (setas verdes) ou inibindo (setas azuis) suas ações. O fenômeno de pleiotropia (quando um único gene influencia múltiplos traços fenotípicos), os modificadores gênicos, e os fatores ambientais também influenciam os genes e o desenvolvimento das doenças monogênicas e complexas.

Existe atualmente uma grande expectativa de que o conhecimento sobre os nsSNPs presentes no genoma de indivíduos da população humana irá possibilitar a avaliação da susceptibilidade de desenvolvimento de doenças e, consequentemente, a escolha do melhor tratamento terapêutico. No entanto, o grande desafio para que isto seja um dia possível é compreender como e quando os nsSNPs podem causar doenças [Sunyaev et al., 2000; Kann, 2007; Torkamani e Schork, 2007].

1.6 A importância dos nsSNPs para pesquisas em Farmacogenética

Um mesmo fármaco pode ter efeitos diversos em pessoas diferentes. Fruto do sequenciamento do genoma humano, a Farmacogenética (ou Farmacogenômica) é uma área que busca estudar a relação desta diversidade com a influência de fatores genéticos no grau de eficiência dos fármacos [Kalow, 1962; Hedgecoe, 2003]. A partir do mapeamento genético de populações, sequenciamento de DNA, análise da expressão gênica e testes clínicos de fármacos, pode-se conhecer as relações entre genes e processos de metabolização, podendo chegar a novos fármacos ou à prescrição daqueles que atendam especificidades genéticas de determinados grupos de pacientes, obtendo assim mais eficácia e menores reações adversas [Kalow et al., 2005].

As diferenças quanto às respostas terapêuticas entre os indivíduos geralmente estão associadas a polimorfismos genéticos presentes em genes que afetam a farmacocinética ou a farmacodinâmica [Chowbay et al., 2005]. Um número considerável de evidências sugere que nsSNPs em genes que codificam receptores, transportadores, ou enzimas metabolizadoras de fármacos, ou envolvidas na biossíntese e reparo do DNA, poderiam determinar a eficácia dos fármacos e sua toxicidade [Ingelman, 2001]. Além dos nsSNPs, outros polimorfismos com consequências farmacogenéticas podem ocorrer, como alterações na região promotora (segmento do DNA em que atuam fatores que estimulam a expressão do gene), defeitos no processo de recomposição (*splicing*) da cadeia do DNA ou duplicações, multiplicações e amplificações de genes, entre outros [Kurtz, 2004].

Portanto variações estruturais nos alvos terapêuticos decorrentes de nsSNPs presentes em genes codificadores destes alvos podem afetar a interação com o fármaco. Assim, quando o fármaco se liga a uma região de interação (sítio ativo ou sítio de ligação) que apresenta variação estrutural decorrente de nsSNPs, diferentes respostas podem ocorrer, dependendo do impacto desta variação na interação com o fármaco (Fig. 6). E, além das variações presentes no sítio de ligação do alvo terapêutico, existem

também as variações presentes em outras regiões da proteína, mais distantes do sítio ativo, mas que também podem afetar a afinidade do fármaco através de mudanças conformacionais que modificam a estabilidade do complexo [Weinshilboum, 2003].

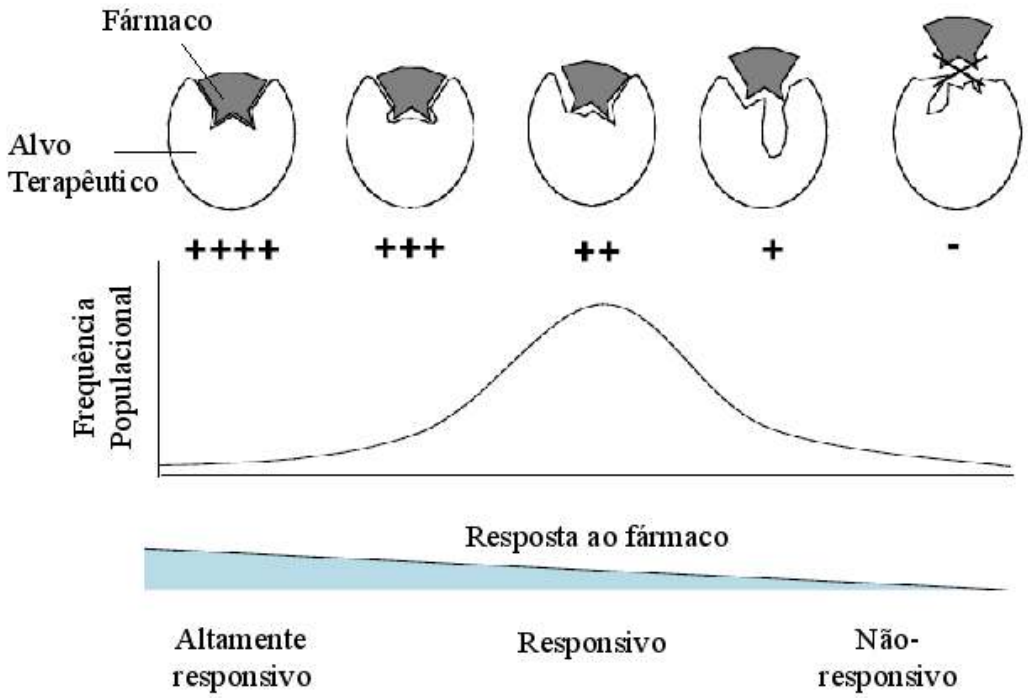


Figura 6. Variações genéticas nos genes codificadores das moléculas receptoras podem afetar a interação com o fármaco. Portanto, quando o fármaco se liga a uma região de interação no alvo terapêutico que apresenta variação estrutural, uma variedade de respostas pode ocorrer, dependendo do impacto da variação na interação do fármaco. Na população, em um extremo estão os pacientes altamente responsivos, e no outro extremo, os pacientes não-responsivos.

1.7 Limitações em estudos de associação de nsSNPs a fenótipos

Como vimos nas seções anteriores, os nsSNPs podem estar associados a doenças e a variações na resposta aos fármacos. Existem várias abordagens para se identificar nsSNPs associados a determinados fenótipos. Após a suspeita de que um determinado fenótipo tem uma causa genética, pode-se fazer uma triagem (*scanning*) no genoma de indivíduos da população, buscando-se variações

genéticas em associação com este fenótipo. No caso dos nsSNPs, este tipo de busca pode ser bastante exaustivo, devido ao grande número de nsSNPs que necessitariam ser submetidos ao processo de triagem [Risch, 2000; Lai et al., 1998]. Através do método de clonagem posicional para encontrar genes candidatos, pode-se reduzir o número de nsSNPs estudados para apenas aqueles localizados em genes que apresentam uma alta probabilidade de estarem associados ao fenótipo [Risch, 2000; Emahazion et al., 2001]. No entanto, mesmo este método pode resultar em uma busca por centenas, ou milhares de SNPs, principalmente se um grande número de genes candidatos for considerado.

Em geral, os estudos de associação genética testam se uma variante genética específica é mais comum entre indivíduos afetados do que em indivíduos controle, ou seja, busca-se determinar a frequência da variação entre os dois grupos. Nestes estudos, os indivíduos controle são recrutados de populações que compartilham semelhanças étnicas ou geográficas com os indivíduos afetados. Caso exista uma diferença estatisticamente significante da frequência desta variação para os dois grupos, ela poderá estar associada ao fenótipo. Portanto, nestes estudos, o efeito de uma dada variação para um fenótipo pode ser visto apenas como uma diferença de frequência desta variação entre indivíduos que apresentam o fenótipo e indivíduos controle [Ramensky et al., 2002]. Assim, nem todos os nsSNPs associados a determinados fenótipos são funcionais, podendo estar em desequilíbrio de ligação com as mutações funcionais.

Mesmo se a associação de um dado nsSNP a um determinado fenótipo for demonstrada inequivocamente, não é evidente que o nsSNP identificado tenha uma relação causal com o fenótipo, ou que a associação estatística não seja o resultado de associação com as mutações funcionais [Johnson e Todd, 2000]. Além disso, diferentemente das mutações penetrantes que causam doenças hereditárias Mendelianas, os nsSNPs associados a fenótipos de doenças humanas complexas ou à resposta a fármacos, não são uma condição necessária e suficiente para definir o fenótipo, pois seus efeitos dependem de muitos outros componentes genéticos e ambientais, como vimos na seção 1.5.

1.8 A utilização da Bioinformática e Quimioinformática para priorizar nsSNPs em estudos de associação a fenótipos

Em estudos de associação de nsSNPs a fenótipos, existem várias recomendações para aumentar as chances de encontrar associações verdadeiras e replicáveis, dentre elas considerar o conhecimento

prévio da probabilidade de que um dado nsSNP esteja associado ao fenótipo, e aumentar esta probabilidade através da priorização dos nsSNPs de acordo com sua importância funcional, por meio de evidências independentes do impacto funcional e estrutural destas variações [Emahazion et al., 2001; Schork et al., 2000]. O conhecimento do significado funcional dos nsSNPs é chave para a compreensão da base biológica da associação a um determinado fenótipo.

Métodos experimentais, tais como a mutagênese sítio-dirigida, são frequentemente aplicados em estudos de especificidade funcional [Wu et al., 1999], estabilidade estrutural [Matthews, 1995], cinética e mecanismo de enovelamento protéico [Ladurner e Fersht, 1997], oligomerização [Chattopadhyay et al., 2006], e estabilidade de complexos protéicos [Otzen e Fersht, 1999]. No entanto, apesar de estes métodos fornecerem a mais forte evidência para o impacto funcional e estrutural causado por nsSNPs, a avaliação experimental da funcionalidade de cada nsSNP existente no genoma humano seria inviável e altamente dispendiosa. Consequentemente, um dos maiores desafios em estudos de associação tem sido identificar nsSNPs funcionais de forma eficiente.

No entanto, a disponibilidade atual de sequências e estruturas protéicas em vários bancos de dados públicos permite o uso de ferramentas computacionais de Quimioinformática e Bioinformática Estrutural para a avaliação das características estruturais, interações moleculares, propriedades dinâmicas e de solvatação de complexos formados, e outros aspectos relevantes ao estudo do impacto causado pela substituição de resíduos de aminoácidos nas proteínas [Laskowski e Thornton, 2008].

Como resíduos de aminoácidos conservados tendem a ser importantes funcionalmente, ou críticos para a manutenção da integridade estrutural, as propriedades evolutivas de resíduos de aminoácidos mutantes podem ser fatores determinantes do seu impacto na função protéica [Ng e Henikoff, 2001]. Vários estudos mostram que o impacto causado pela substituição de resíduos de aminoácidos na estrutura protéica pode ser predita pela análise de alinhamento múltiplo de sequências [Sunyaev et al., 2000; Chasman et al., 2001; Ng e Henikoff, 2001; Ferrer-Costa et al., 2002].

A utilização de métodos computacionais também pode contribuir para aumentar a eficiência da predição do impacto causado por nsSNPs em alvos terapêuticos e enzimas metabolizadoras de fármacos [Kapetanovic, 2008]. A etapa de interação com o receptor (relacionada à potência) é a fase da ação terapêutica mais bem descrita por modelos teóricos, e será abordada neste trabalho através da utilização do método computacional de *docking* molecular.

Assim, ferramentas computacionais de predição de impacto de nsSNPs podem ser usadas para avaliar se uma associação relatada pode, de fato, ter um impacto funcional e, portanto, uma menor

probabilidade de representar um falso-positivo ou falso-negativo, resultando em uma conclusão mais confiável sobre o possível impacto de um nsSNP na função protéica.

Uma das grandes vantagens da utilização de uma abordagem computacional é o fato de poder auxiliar na predição preliminar do impacto de um enorme número de nsSNPs em um curto espaço de tempo e com baixo custo. Além disso, esta abordagem pode elucidar os mecanismos que afetam a função gênica, algo que é frequentemente apenas especulado por estudos experimentais.

2. JUSTIFICATIVA E RELEVÂNCIA

O desenvolvimento de tecnologias de sequenciamento nos últimos anos permitiu um grande avanço na geração de dados genômicos e proteômicos. No entanto, a enorme quantidade de dados atualmente disponíveis requer o auxílio de métodos computacionais que possibilitem a análise do significado biológico representado nestes dados, algo que no momento seria impraticável utilizando apenas métodos experimentais. Além disso, existem dados biológicos disponíveis em diversas bases de dados públicas que poderiam ser integrados, possibilitando novas perspectivas de estudo.

Devido à enorme quantidade de dados sobre nsSNPs identificados no genoma humano, torna-se necessária a utilização de ferramentas computacionais para a avaliação preliminar do impacto funcional e estrutural destas variações na função protéica. Além disso, a integração e filtragem de dados provenientes de uma variedade de fontes relevantes a estudos de associação permite a priorização de nsSNPs para estudos de validação experimental de forma rápida e econômica.

No presente trabalho, propomos a utilização de várias ferramentas computacionais para o estudo do impacto causado por substituições de resíduos de aminoácidos na função protéica decorrentes de nsSNPs, dentre elas ferramentas de modelagem, otimização estrutural e *docking* molecular. Através da construção de um banco de dados público, este trabalho também descreve a importância de se relacionar resultados obtidos nestas análises computacionais com informações já existentes sobre doenças, vias metabólicas, alvos terapêuticos, fármacos, enzimas metabolizadoras de fármacos, e anotações de sequências protéicas, possibilitando diversos tipos de busca, dependendo do interesse de pesquisa.

Este estudo fornecerá dados preliminares que poderão ser usados para auxiliar na predição do impacto causado por nsSNPs em genes codificadores de alvos terapêuticos, e também na escolha de nsSNPs para estudos experimentais sobre possíveis associações com doenças humanas.

3. OBJETIVOS

O objetivo do presente trabalho foi utilizar uma abordagem computacional para a predição preliminar do impacto funcional e estrutural causado por nsSNPs em genes codificadores de proteínas em humanos, relacionando resultados de análises obtidos por softwares de *docking* molecular, modelagem molecular, e de análises do impacto de substituições de resíduos de aminoácidos, com informações já existentes sobre doenças, vias metabólicas, enzimas, anotações, genes, fármacos e alvos terapêuticos.

A integração destas informações foi disponibilizada através da construção de um banco de dados relacional, o TargetSNPdb, que pode ser acessado no *site: http://nequim.qui.ufmg.br/targetsnp/*

4.1 Análise do efeito funcional de nsSNPs usando um método baseado em homologia de sequências (SIFT)

A ferramenta computacional SIFT (*Sorting Intolerant From Tolerant*) tem como função determinar o efeito funcional causado por substituições de resíduos de aminoácidos nas proteínas. O algoritmo utilizado pela ferramenta é baseado na premissa de que a evolução protéica está correlacionada com a função protéica. Portanto, o alinhamento de proteínas de uma mesma família deve mostrar a conservação de resíduos de aminoácidos localizados em posições importantes para a função protéica [Ng et al., 2001].

Através da *homepage* do programa SIFT (http://sift.jcvi.org/), pode-se submeter uma sequência protéica de interesse, e as posições e substituições de resíduos de aminoácidos que serão avaliadas pelo algoritmo de predição. Inicialmente, a sequência protéica de entrada é usada pela ferramenta PSI-BLAST (Position-Specific Iterated BLAST) para se recuperar sequências protéicas similares, assim como um alinhamento múltiplo de todas estas sequências. Em seguida, o algoritmo utiliza as sequências resultantes desta primeira busca que obtiveram um *score* de similaridade acima do limite de 90% para criar uma matriz de valores posição-específica (*position-specific scoring matrix*, PSSM) baseada no alinhamento destas sequências. Esta matriz tenderá a fornecer *scores* mais altos para regiões conservadas dentro deste conjunto de sequências estudadas e *scores* baixos para regiões pouco conservadas.

Usando a matriz de valores posição-específica gerada, o algoritmo calcula as probabilidades normalizadas para todas as substituições de resíduos de aminoácidos possíveis em cada posição do alinhamento. As substituições que apresentam um valor de tolerância menor do que 0,05 são preditas intolerantes ou deletérias, enquanto aquelas que apresentam um valor maior do que 0,05 são preditas tolerantes [Ng et al., 2001; Ng et al., 2006].

Em estudos em que o programa SIFT foi utilizado para analisar nsSNPs presentes em genes associados a doenças humanas, foi demonstrado uma precisão de predição entre 65 e 92% [Ollila et al., 2006; Balasubramanian et al., 2005; Bao et al., 2005., Raevaara et al., 2005; Xi et al., 2004].

Usando dados experimentais de um estudo de mutagênese com a proteína protease HIV-1 (336 mutações) [Loeb et al., 1989], o efeito fenotípico causado por cada mutação nesta proteína também foi comparado com as predições computacionais. A precisão de predição do método SIFT para todas as mutações estudadas na proteína foi de 78% [Ng et al., 2001], sendo que as taxas de "falso-negativos" e "falso-positivos" registrada foram iguais a 31% e 20%, respectivamente [Ng et al., 2006].

4.2 Análise do efeito funcional de nsSNPs usando um método baseado em homologia de estruturas (PolyPhen)

Polyphen (*Polymorphism Phenotyping*) é uma ferramenta usada na predição do impacto funcional e estrutural de substituições de resíduos de aminoácidos em proteínas (http://genetics.bwh.harvard.edu/pph/). As predições são feitas pelo PolyPhen usando-se três fontes de dados: anotações de sequências obtidas no banco de dados SwissProt (http://expasy.org/sprot/), alinhamento múltiplo de sequências usando o software BLAST, e informações estruturais (Fig. 7). A disponibilidade destas três fontes de dados indica a mais alta confiabilidade na predição [Ramensky et al., 2002].

Semelhantemente ao programa SIFT, para uma dada sequência protéica de interesse, o primeiro passo do algoritmo do programa PolyPhen é a busca e o alinhamento múltiplo de sequências homólogas usando a ferramenta BLAST. Em seguida, o alinhamento múltiplo resultante é usado pelo software PSIC (*Position-Specific Independent Counts*) para calcular uma matriz de "perfil de *scores*". Os elementos desta matriz são razões logarítmicas entre a probabilidade de um dado resíduo de aminoácido ocorrer em uma posição específica e a probabilidade deste resíduo de aminoácido ocorrer em qualquer posição da proteína (frequência *background*). PolyPhen calcula o valor absoluto da diferença entre os "perfis de *scores*" dos resíduos de aminoácidos variantes na posição variante de interesse. Valores altos desta diferença podem indicar que a substituição estudada é raramente (ou nunca) observada na família protéica [Sunyaev et al., 1999].

Utilizando informações anotadas nas bases de dados SWALL e SwissProt, a posição do resíduo de aminoácido variante é também mapeada na estrutura protéica correspondente à sua sequência protéica primária, com o objetivo de avaliar se a substituição do resíduo de aminoácido poderia afetar o núcleo hidrofóbico da proteína, acessibilidade a solvente, interações eletrostáticas, interações com

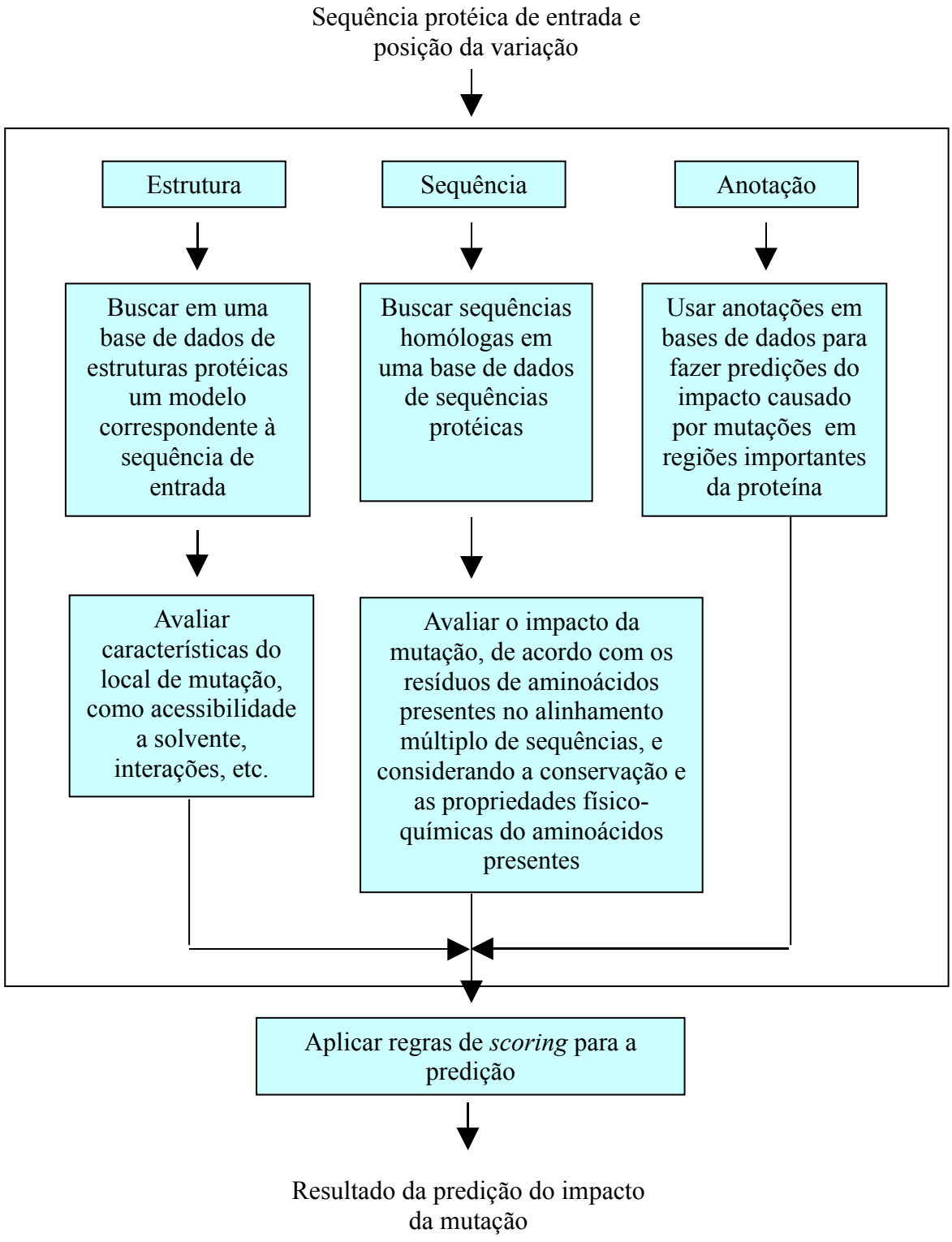


Figura 7. Fluxograma mostrando a sequência de passos utilizados pelo programa PolyPhen na predição do impacto de uma mutação pontual.

ligantes, ou outras características importantes da proteína. Caso não existam estruturas protéicas correspondentes à sequência protéica primária usada como entrada, PolyPhen utiliza proteínas homólogas que possuem estruturas elucidadas [Ramensky et al., 2002].

A precisão de predição do programa PolyPhen foi avaliada em 80% [Ramensky et al., 2002], com uma taxa de "falsos-negativos" e "falsos-positivos" de 31% e 9%, respectivamente [Ng et al., 2006].

4.3 Modelagem molecular de cadeias laterais de resíduos de aminoácidos

A correta modelagem molecular das conformações das cadeias laterais de resíduos de aminoácidos é importante para se compreender vários aspectos da estrutura e função protéica, como a interação com outras moléculas e a estabilidade termodinâmica. Isto implica que a predição da conformação das cadeias laterais é útil apenas se for altamente precisa, o que a torna um problema desafiador.

Apesar de que a modelagem molecular de uma única cadeia lateral de um resíduo de aminoácido em um dado ambiente atômico pareça ser um dos problemas mais simples de predição de estrutura protéica, este problema ainda não está totalmente resolvido [Fiser, 2004]. Uma pequena mudança de uma cadeia lateral de um único resíduo de aminoácido pode conduzir a uma mudança conformacional ou perda significativa de função protéica [Wu et al., 1999].

Duas simplificações são frequentemente usadas na modelagem da conformação de cadeias laterais. Primeiro, a substituição de resíduos de aminoácidos frequentemente deixa a cadeia principal inalterada [Chothia et al., 1986]. Portanto, muitos algoritmos fixam a cadeia principal durante a busca pelas melhores conformações da cadeia lateral. Segundo, foi observado que a maioria das cadeias laterais em estruturas cristalográficas de alta resolução pode ser representada por um número limitado de confôrmeros que obedecem a restrições estereoquímicas e energéticas [Janin et al., 1978].

Esta observação motivou Ponder e Richards a desenvolver a primeira biblioteca de rotâmeros de cadeias laterais para 17 tipos de resíduos de aminoácidos que possuem graus de liberdade em ângulos diedros nas suas cadeias laterais [Ponder e Richards, 1987]. A biblioteca foi baseada em 10

estruturas protéicas de alta resolução determinadas experimentalmente por cristalografia de raios X. Métodos mais recentes e eficientes também são baseados em bibliotecas de rotâmeros, embora alguns destes métodos tenham expandido radicalmente o tamanho da biblioteca, chegando a conter aproximadamente 50000 estados de rotâmeros [Xiang et al., 2001; Canutescu et al., 2003; Peterson et al., 2004].

Pelo método de busca em bibliotecas de rotâmeros, cada rotâmero é avaliado usando-se uma função de energia (ou função de *score*). Apesar de que as funções de energia utilizadas pelos primeiros métodos de modelagem de cadeias laterais eram geralmente simplificadas [Dunbrack et al., 1993], estas abordagens eram justificadas pelos seus desempenhos.

Em contraste, surgiram também métodos baseados no procedimento de minimização de energia (ou otimização) da estrutura protéica. As várias abordagens incluem simulação de Monte Carlo [Eisenmenger et al., 1993; Jain et al., 2006], anelamento simulado [Lee e Levitt, 1991], uma combinação de Monte Carlo com anelamento simulado [Holm e Sander, 1992], o teorema da eliminação *dead-end* [Lasters e Desmet, 1993; Looger e Hellinga, 2001], algoritmos genéticos [Tuffery et al., 1991], redes neurais com anelamento simulado [Hwang e Liao, 1995], otimização do campo médio [Koehl e Delarue, 1994], e buscas combinatoriais [Dunbrack et al., 1993; Bower et al., 1997; Petrella et al., 1998].

4.3.1 Comparação de diferentes métodos de modelagem molecular de estruturas protéicas mutantes

Neste trabalho, foi feita a comparação da precisão de modelagem molecular de cadeias laterais entre quatro métodos frequentemente utilizados para este fim: Swiss-Pdb Viewer (DeepView), MODELLER, SCWRL3 e 4. Um total de 212 pares de estruturas protéicas que diferem por um único resíduo de aminoácido, e que foram resolvidas por cristalografia de raios X em uma resolução igual ou menor do que 2,0 Å foram obtidos do PDB. Usando os métodos descritos a seguir, foi feita a substituição de resíduos de aminoácidos de um membro de cada par de estruturas protéicas, de forma que o resíduo modelado pudesse ser comparado com o resíduo nativo da proteína cristalizada.

Swiss-Pdb Viewer (DeepView)

O programa Swiss-Pdb Viewer permite fazer a substituição das cadeias laterais de resíduos de

aminoácidos através de uma busca em uma biblioteca de rotâmeros. A cadeia lateral original é substituída por um rotâmero da cadeia lateral do resíduo de aminoácido variante de interesse, sendo que este possui o mais baixo *score* resultante de cálculo usando a seguinte fórmula (http://spdbv.vital-it.ch/mutation_guide.html):

Score = (4 x Número de colisões com os átomos N, Ca e C da cadeia principal) +
(3 x Número de colisões com os átomos O da cadeia principal) +
(2 x Número de colisões com átomos da cadeia lateral) –
(Número de ligações de hidrogênio) –
(4 x Número de pontes dissulfeto)

Apesar de que o processo de busca e seleção de um rotâmero é extremamente rápido, o programa não está disponível como linha de comando, o que impossibilita sua utilização em estudos de larga escala.

MODELLER

Usando o script mutate_model.py do programa MODELLER [Sali e Blundell, 1993] (http://salilab.org/modeller), podemos fazer a substituição de um resíduo de aminoácido em uma posição de interesse. Em seguida a conformação da cadeia lateral do resíduo de aminoácido variante é otimizada pelo método de gradiente conjugado, e em seguida é feito o refinamento usando-se dinâmica molecular, considerando todos os átomos do aminoácido variante, incluindo átomos da cadeia principal. Como o programa está disponível como linha de comando, todo o processo de criação de estruturas variantes pode ser automatizado, possibilitando o processamento de um grande número de estruturas.

A função de *scoring* utilizada pelo MODELLER para avaliar as conformações geradas considera a energia interna que descreve aspectos conformacionais através de termos do campo de força CHARMM, restringindo o comprimento das ligações covalentes, dos ângulos diedros, e da planaridade das ligações peptídicas. Esta função usa o potencial de Lennard-Jones para termos de interação de átomo não-ligados e combina restrições espaciais derivadas por homologia em ângulos diedros do modelo com preferências estatísticas observadas em diversas estruturas representativas [Sali e Blundell, 1993].

SCWRL3 e SCWRL4

A modelagem de cadeias laterais de resíduos de aminoácidos é feita pelo programa SCWRL3,

utilizando uma biblioteca de rotâmeros dependente da cadeia principal [Dunbrack e Cohen, 1997], uma função de energia simples baseada na frequência de rotâmeros na biblioteca e em um termo de energia conformacional repulsiva, e um gráfico de decomposição para solucionar o problema de empacotamento combinatorial [Canutescu et al., 2003]. A função de energia da versão 4 do SCWRL foi aperfeiçoada usando-se uma biblioteca de rotâmeros nova, que utiliza estimativas de densidade de Kernel e regressões de Kernel para fornecer frequências de rotâmeros, e ângulos diedros [Krivov et al. 2009].

4.4 Minimização de Energia

Uma vez que todos os átomos da estrutura protéica são conectados por ligações com comprimentos rigidamente fixos, a movimentação de um átomo em uma parte da estrutura protéica possui efeitos de longo alcance em seus vizinhos. Portanto a movimentação de uma parte da proteína para uma melhor configuração, que pode ocorrer como decorrência da substituição de um resíduo de aminoácido, pode causar a movimentação de outra parte da proteína para uma configuração desfavorável [Gibas e Jambeck, 2002].

As estruturas protéicas podem se adaptar a mutações pontuais através do rearranjo espacial do ambiente localizado ao redor do resíduo de aminoácido mutante. Em alguns casos, ocorre apenas uma leve mudança na conformação da cadeia principal, mas em ambientes menos empacotados, também é possível que a mutação não cause qualquer alteração ou distorção da cadeia principal [Feyfant et al., 2007].

Em todo caso, uma vez obtida uma estrutura protéica modelada, a conformação em questão pode não ser – e frequentemente não é – aquela correspondente a um mínimo local de energia. Através do método computacional de minimização de energia, é feita uma série iterativa de pequenas mudanças nas posições dos átomos da proteína, visando obter uma estrutura de mínimo de energia local.

Dentre vários métodos utilizados na minimização de energia está o método do declive máximo (*steepest descent*), que é empregado quando se está partindo de uma situação muito energética e se deseja chegar às imediações de um mínimo local tanto quanto possível. Esta etapa pode ser seguida de uma minimização refinada (usando-se, por exemplo, o método do gradiente conjugado), que se beneficia de informações do passo anterior: se a "história" da minimização que está sendo seguida leva

a uma conformação de menor energia, a história é mantida; caso contrário, muda-se a direção do cálculo [Cramer, 2004; Young, 2001].

4.5 Método de predição de afinidade de ligação (AutoDock 4.0)

Ao processo de se posicionar o ligante em várias orientações no sítio ativo do receptor e, usualmente, em diferentes conformações, com o intuito de se obter a melhor interação, chama-se pela designação em inglês *docking*, que pode-se traduzir como "docagem" ou "ancoragem". Este procedimento permite o estabelecimento de uma classificação entre os compostos de maior e de menor afinidade a um determinado receptor (Fig. 8). Existem vários programas de *docking*, dentre eles o DOCK [Ewing et al., 2001], AutoDock [Morris et al., 1998], GOLD [Jones et al., 1997], FlexX [Kramer et al., 1999], SURFLEX [Jain, 2003], que realizam esta ordenação de forma automática. No presente trabalho, utilizamos o programa AutoDock 4.0, que é amplamente utilizado em estudos de ancoragem de pequenas moléculas em macromoléculas protéicas, além de ser gratuito.

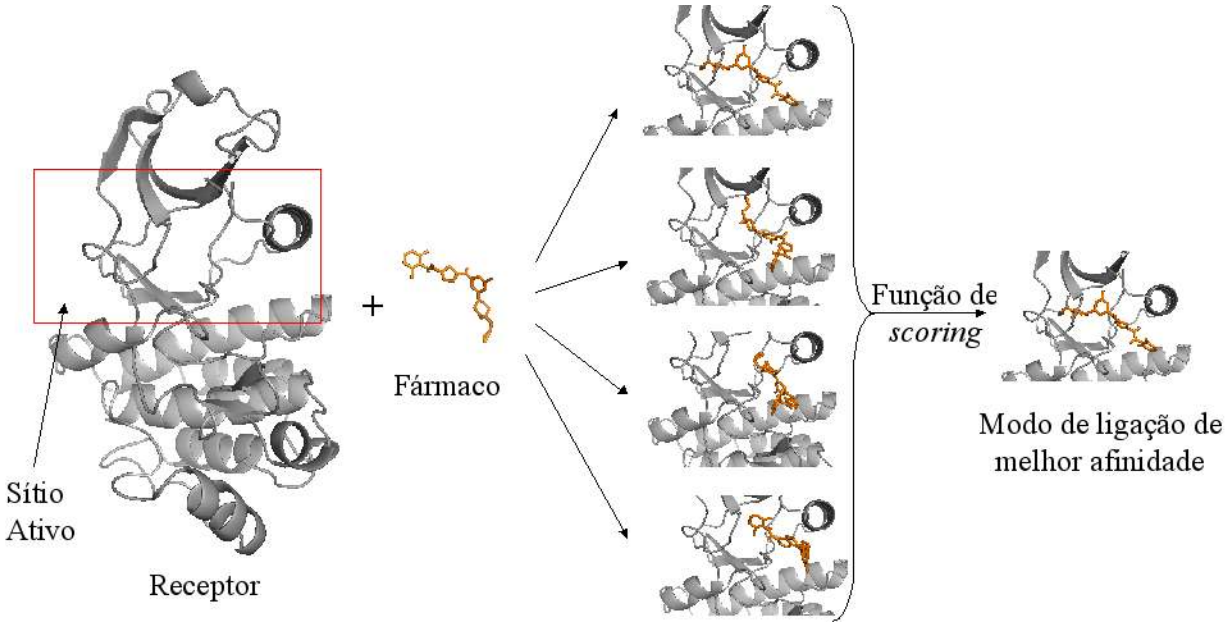


Figura 8. Através do método de *docking* molecular, é possível fazer a busca de um fármaco que seja capaz de ajustar ao sítio ativo de um receptor tanto geometricamente quanto quimicamente. A simulação compreende dois procedimentos: a busca conformacional por diferentes modos de ligação do

ligante no sítio ativo do receptor, e a avaliação da afinidade de cada um deste modos de ligação usando uma função de *scoring*. A figura acima mostra a interação do fármaco Imatinib no interior do sítio ativo do domínio ABL da tirosina quinase BCR-ABL (PDB id: 2hyy).

No AutoDock 4.0, o primeiro passo do *docking* molecular de um ligante no sítio ativo de uma proteína é a criação de mapas de potenciais de afinidade atômicos para cada átomo da molécula do ligante usando uma biblioteca de sondas pré-definidas (Fig. 9). Para realizar este procedimento, a região do sítio ativo é selecionada no interior de uma grade tridimensional de pontos posicionados regularmente. Uma sonda de um átomo do ligante é posicionada em cada ponto da grade e a energia de interação entre este átomo (em cada ponto da grade) e os átomos da proteína é calculada. Uma grade de afinidade é calculada para cada tipo de átomo do ligante (tipicamente carbono, oxigênio, nitrogênio e hidrogênio). O tempo de cálculo das grades de afinidade é proporcional apenas ao número de átomos do ligante, e é independente do número de átomos da proteína.

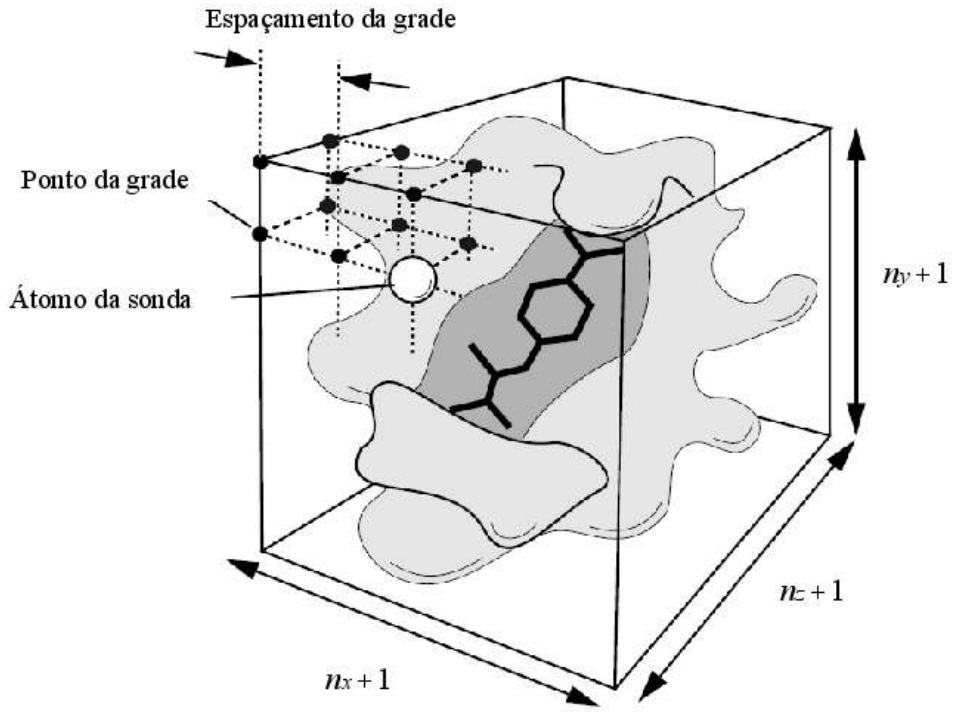


Figura 9. O processo de busca conformacional do ligante pode ser acelerado através da criação de mapas de potenciais de afinidade atômica para cada átomo da molécula do ligante (fonte: Morris et al., 2001).

A simulação de *docking* molecular foi feita usando o algoritmo genético de busca Lamarckiano disponível no programa AutoDock 4.0 (Fig. 10), que é um método mais eficiente e robusto do que os métodos de simulação de Monte Carlo [Smith et al., 2000]. Com a proteína estática durante a simulação, a molécula do ligante faz uma busca aleatória pelo espaço determinado para busca. Em cada passo da simulação, uma pequena modificação aleatória é feita: translação do centro de gravidade, orientação, e rotação ao redor de cada um dos ângulos diedros internos flexíveis. Esta modificação resulta em uma nova configuração, e a energia deste modo de ligação é avaliada usando a grade de afinidade pré-calculada. O valor desta nova energia é comparado ao valor obtido no passo anterior. Se o valor da nova energia é menor, a nova configuração é aceita.

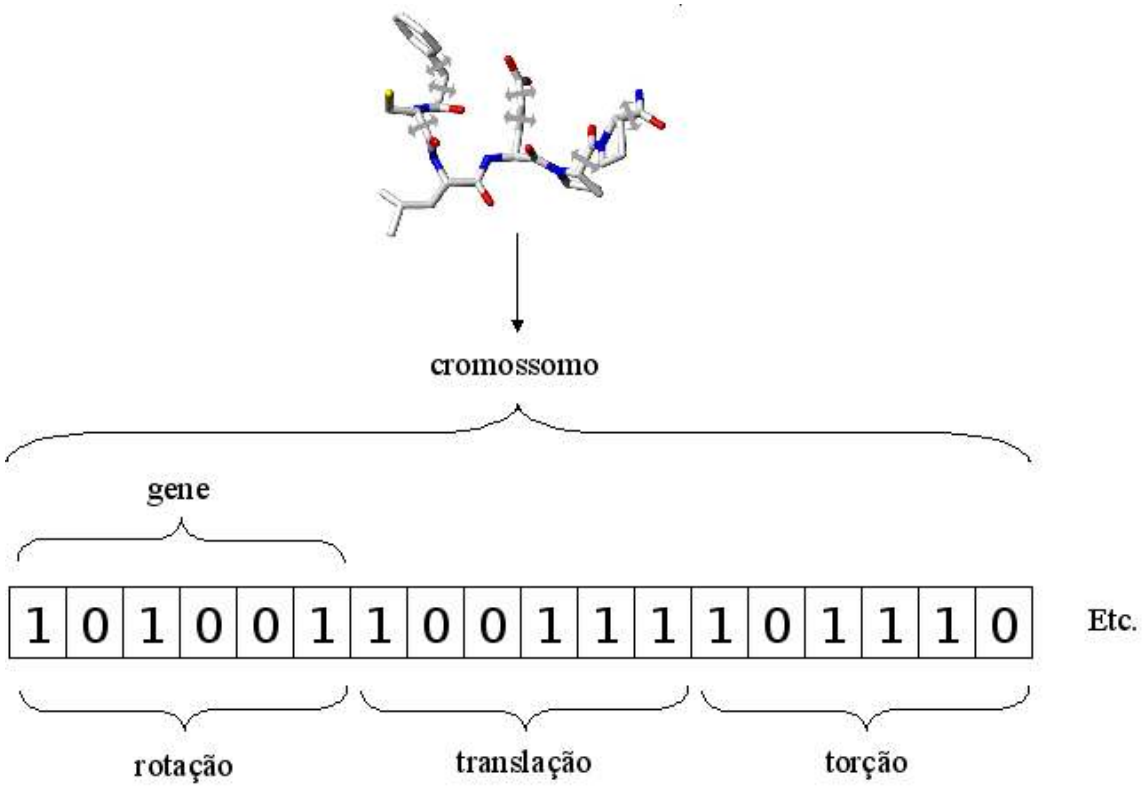


Figura 10. A busca conformacional do ligante pode ser feita usando o algoritmo genético Lamarckiano. Este algoritmo representa as diferentes variáveis de grau de liberdade do ligante como genes, e cada gene contém uma sequência de bits representativos do modo de ligação do ligante, e o conjunto destes genes forma o cromossomo de um indivíduo.

A separação do cálculo das grades de afinidade do processo de simulação do *docking* molecular possibilita a modularização do procedimento, permitindo a exploração de um vasto número de representações de interações moleculares de forma rápida.

A afinidade de ligação calculada pelo programa é igual à diferença entre as energias do ligante e da proteína em um estado isolado, e a energia do complexo formado. Esta avaliação é separada em dois passos: primeiramente é avaliada a variação da energia intramolecular nas moléculas isoladas e na conformação do complexo, em seguida é avaliada a variação da energia intermolecular resultante da formação do complexo [Huey et al., 2007]. O campo de força inclui seis termos de interação entre pares de átomos (*pair-wise*) (V) e uma estimativa da entropia conformacional perdida durante a ligação (ΔS_{conf}):

$$\Delta G = (V_{ligado}^{L-L} - V_{separado}^{L-L}) + (V_{ligado}^{P-P} - V_{separado}^{P-P}) + (V_{ligado}^{P-L} - V_{desligado}^{P-L}) + \Delta S_{Conf}$$

onde L se refere ao "ligante" e P se refere à "proteína". Os dois primeiros termos são as energias intramoleculares para os estados complexado e isolado do ligante, e os dois termos seguintes são as energias intramoleculares para os estados complexado e isolado da proteína. A variação da energia em função das interações intermoleculares entre os estados complexado e isolado está descrita no terceiro parênteses. Pressupõe-se que as duas moléculas estão suficientemente distantes uma da outra no estado isolado para que $V_{desligado}^{P-L}$ seja igual a zero.

Os termos atômicos entre pares de átomos incluem interações de dispersão/repulsão, ligação de hidrogênio, interação eletrostática, e desolvatação:

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hcomplexo} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2 / 2\sigma^2)}$$

onde W é a constante de ajuste (*weighting*). O primeiro termo é um potencial 6/12 para interações de dispersão/repulsão de van der Waals, onde os parâmetros A e B foram obtidos do campo de força Amber [Weiner et al., 1984]. O segundo termo é associado às ligações de hidrogênio direcionais baseado em um potencial 10/12 [Goodford, 1985], onde os parâmetros C e D são atribuídos para fornecer uma energia máxima de 5 kcal/mol em uma distância de 1,9 Å para O-H e N-H, e de 1

kcal/mol a uma distância de 2,5 Å para S-H. A direcionalidade da interação da ligação de hidrogênio *E(t)* é dependente do ângulo *t*. As interações eletrostáticas são avaliadas usando o potencial de Coulomb (terceiro termo). O termo final é um potencial de desolvatação baseado no volume (*V*) dos átomos ao redor de um dado átomo, ponderado por um parâmetro de solvatação e um termo exponencial baseado em distância [Stouten et al., 1993].

O termo referente à perda de entropia torsional durante a ligação (ΔS_{conf}) é diretamente proporcional ao número de ligações rotacionáveis na molécula (N_{tors}):

$$\Delta S_{conf} = W_{conf} N_{tors}$$

O número de ligações rotacionáveis inclui todos os graus de liberdade torsionais, incluindo a rotação de átomos de hidrogênio polares em grupos hidroxila.

4.6 NEQUIM Contact System

Com o objetivo de se comparar os diferentes modos de ligação de ligantes no sítio ativo das proteínas, foi desenvolvida uma ferramenta computacional que possibilita a análise detalhada das interações inter-atômicas entre ligantes e resíduos de aminoácidos através da criação de *fingerprints* de interação. Os resultados deste trabalho foram apresentados na IV Conferência Internacional da AB3C (X-Meeting 2008) [José et al., 2008].

O NEQUIM Contact System (NCS) usa uma representação binária 1D das interações moleculares presentes em um complexo tridimensional proteína/ligante. Primeiramente, é feita a identificação de todos os resíduos de aminoácidos envolvidos em algum tipo de interação com o ligante, e em seguida é feita a classificação destas interações. Foi implementado um total de seis *bits* para cada resíduo de aminoácido que está em contato com o ligante no sítio de ligação (Fig. 11). Os *bits* são ligados (valor=1) ou desligados (valor=0) se as seguintes interações estão presentes: 1) se algum contato está envolvido nesta posição; 2) se o contato ocorre pela cadeia lateral; 3) se o contato é uma ligação de hidrogênio; 4) se o contato é aromático; 5) se o contato é hidrofóbico; 6) se o contato é desestabilizante. Desta forma, cada resíduo é representado por um *bitstring* de 6 *bits* de comprimento.

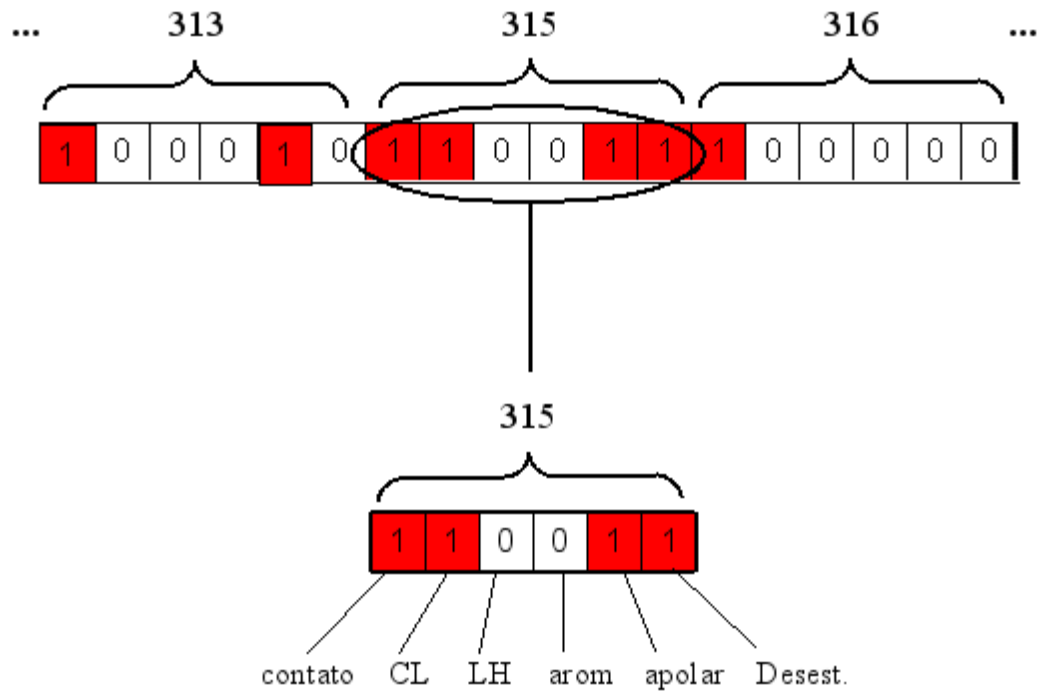


Figura 11. Mapa de contato gerado pelo programa NCS referente às interações entre um ligante e os resíduos de aminoácidos de uma proteína nas posições 313, 315 e 316 da sequência primária protéica. Abaixo, em destaque, os tipos de interações representadas pelo *bitstring*.

O *fingerprint* de interação completo para um complexo entre um ligante e uma proteína é finalmente construído pela concatenação sequencial dos *bitstrings* de cada resíduo de aminoácido no sítio de ligação da proteína, de acordo com a ordem ascendente de numeração dos resíduos.

A comparação entre diferentes modos de ligação de ligantes pode então ser feita através de cálculo de similaridade usando o coeficiente de Tanimoto [Tanimoto, 1957; Rogers, 1960]:

$$CT = N_{AB} / (N_A + N_B - N_{AB})$$

onde N_A e N_B representam os números de *bits* ligados nos *fingerprint* A e B, respectivamente, e N_{AB} representa o número de bits ligados tanto no *fingerprint* A quanto no *fingerprint* B.

4.7 Banco de Dados MySQL

Os bancos de dados são ferramentas de extrema importância na Bioinformática, pois permitem tanto o armazenamento quanto a busca e recuperação de informações biológicas. Dentre os tipos de sistemas de gerenciamento de banco de dados existentes, os dois mais utilizados são: sistemas de indexação de arquivos simples e relacionais (RDBMSs – *Relational Database Management Systems)* [Gibas e Jambeck, 2002].

Um banco de dados de arquivos simples não é realmente um banco de dados, é simplesmente uma coleção ordenada de arquivos semelhantes, geralmente em conformidade com um formato padrão de conteúdo. Os bancos de dados de arquivos simples se tornam úteis com a ordenação e a indexação. Um índice extrai um atributo específico de um arquivo e alinha o valor do atributo no índice com um nome de arquivo e uma localização [Celko, 1999; Gibas e Jambeck, 2002].

Por outro lado, um banco de dados relacional armazena dados em tabelas separadas em vez de colocar todos os dados em um único local. Os dados em uma tabela de banco de dados relacional são organizados em linhas, onde cada linha representa um registro no banco de dados. Uma linha pode conter várias informações separadas (campos), e cada campo pode conter uma informação distinta. Não pode consistir em um conjunto ou lista que possam ser divididos em partes menores. A função do RDBMS é fazer a conexão entre tabelas relacionadas, localizando rapidamente os elementos comuns que estabelecem esses relacionamentos. A rede de tabelas e relacionamentos que compõe um banco de dados é denominada esquema de banco de dados [Celko, 1999], que pode ser construído e visualizado utilizando programas específicos, como o DBDesigner (http://www.fabforce.net/dbdesigner4).

O MySQL é um DBMS relacional de código aberto que possibilita ao usuário criar, manter e gerenciar bancos de dados eletrônicos (http://www.mysql.com). As principais vantagens deste banco de dados são velocidade, robustez e facilidade de uso [DuBois, 2000]. No MySQL, o conceito da estrutura que mantém os blocos (ou registros) de informações é chamado de tabela. Estes registros, por sua vez, são constituídos de objetos menores que podem ser manipulados pelos usuários, conhecidos por tipos de dados (*datatypes*). Juntos, um ou mais *datatypes*, formam um registro (*record*). Uma hierarquia de banco de dados pode ser considerada como: Banco de dados > Tabela >Registro > Tipo de dados. Os tipos de dados possuem diversas formas e tamanhos, permitindo ao programador criar tabelas específicas de acordo com suas necessidades.

Neste trabalho, o banco de dados foi instalado em um servidor DELL Power Edge com o

sistema operacional Ubuntu-Linux, com a capacidade de 1 TB de HD, tendo grande capacidade para armazenamento de dados e rapidez nos processos de busca e aquisição das informações.

4.7.1 Programas, servidores e links no TargetSNPdb

Desde o advento da *World Wide Web*, diversos bancos de dados biológicos públicos se tornaram disponíveis para *download*. A Tabela 1 mostra alguns *sites* de onde foram obtidas informações biológicas contidas no banco de dados TargetSNPdb, além de informações dos programas utilizados para gerar novas análises. A tabela se divide em 4 partes, as quais apresentam informações sobre os tipos de dados, as fontes, a caracterização em S ou P (Servidor ou Programa) e o endereço eletrônico, respectivamente.

Tabela 1. Descrição dos dados contidos no banco de dados TargetSNPdb.

Tipo de Dado	Fonte	S / P	URL
Registros de nsSNPs	dbSNP	S	http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp
Dados protéicos	Swiss-Prot	S	http://ca.expasy.org/sprot/
Anotação de proteínas variantes	Swissvar	S	http://www.expasy.org/swissvar/
Dados protéicos	PDB	S	http://rcsb.org/pdb
Mapeamento das entradas do SwissProt às cadeias do PDB	PDBSWSdb	S	http://www.bioinf.org.uk/pdbsws/
Doenças genéticas humanas	OMIM	S	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
Associação de nsSNPs a doenças	Genetic Association Database	S	http://geneticassociationdb.nih.gov/
Dados de frequência populacional de nsSNPs	HapMap BioMart	S	http://hapmart.hapmap.org/BioMart/martview
Dados de enzimas metabolizadoras que possuem nsSNPs	JSNP	S	http://snp.ims.u-tokyo.ac.jp/
Classificação das vias metabólicas	PANTHERdb	S	http://www.pantherdb.org/pathway/
Mapeamento da posição de nsSNPs em estruturas protéicas	coliSNP	S	http://yayoi.kansai.jaea.go.jp/colisnp/
Artigos do PubMed com registros de nsSNPs	PubMed	S	http://www.ncbi.nlm.nih.gov/pubmed/
Genes contendo nsSNPs	NCBI Entrez Gene	S	http://www.ncbi.nlm.nih.gov/gene
	DrugBank	S	http://www.drugbank.ca/
	TTD	S	http://bidd.nus.edu.sg/group/cjttd/
Alvos terapêuticos e fármacos	KEGG	S	http://www.genome.jp/kegg/
Ferramenta de predição de impacto de mutações baseado em homologia de sequências	SIFT	P	http://sift.jcvi.org/
Ferramenta de predição de impacto de mutações baseado em homologia de estruturas	PolyPhen	P	http://genetics.bwh.harvard.edu/pph/
Docking molecular	AutoDock 4.0	P	http://autodock.scripps.edu
Modelagem de cadeias laterais de resíduos de aminoácidos	SCWRL4	P	http://dunbrack.fccc.edu/
Minimização de estrutura protéica	GROMACS	P	http://www.gromacs.org/

S: Servidor; P: Programa

5.1 Avaliação da precisão de vários métodos de modelagem de cadeias laterais de resíduos de aminoácidos

Com o objetivo de comparar a precisão de vários programas públicos de modelagem molecular de cadeias laterais de resíduos de aminoácidos, foi feita uma avaliação da porcentagem de ângulos diedros $\chi 1$, $\chi 2$ e $\chi 1+2$ das cadeias laterais dos resíduos de aminoácidos modelados preditos corretamente (Apêndice 8.1). Seguindo a convenção usual, para um dado resíduo de aminoácido, um ângulo diedro é definido como correto quando seu valor ocorre dentro do limite de 40° em comparação ao ângulo correspondente na estrutura cristalográfica da proteína modelada [Dunbrack et al., 1993; Jacobson et al., 2002].

As estruturas utilizadas neste estudo foram obtidas do banco de dados de estruturas protéicas PDB (http://www.pdb.org). Foram buscadas estruturas protéicas mutantes cristalizadas, e suas respectivas estruturas protéicas nativas cristalizadas, de forma que cada par de estruturas difere apenas por um resíduo de aminoácido. Desta forma, a precisão da modelagem de um resíduo de aminoácido na estrutura nativa pode ser comparada com o resíduo de aminoácido presente na estrutura mutante cristalizada. A lista de estruturas utilizadas neste estudo pode ser encontrada no Apêndice 8.2.

Primeiramente, foi feita uma análise da precisão da modelagem baseada nos diferentes tipos de resíduos de aminoácidos modelados (Fig. 12). Devido à variedade de espaço conformacional disponível dentre os diferentes tipos de resíduos [Feyfant et al., 2007], podemos observar que resíduos que possuem grande restrição de flexibilidade conformacional, como a prolina (Pro), ou o triptofano (Trp), apresentaram uma maior precisão de modelagem pelos diferentes métodos do que resíduos que possuem pequena restrição de flexibilidade conformacional, como a leucina (Leu), glutamina (Gln), ou a valina (Val). Podemos observar também que a precisão da modelagem de vários resíduos, como o triptofano (Trp), tirosina (Tyr), asparagina (Asp), fenilalanina (Phe), e a leucina (Leu), foi similar para os diferentes métodos utilizados, enquanto outros resíduos, como a glutamina (Gln), serina (Ser), isoleucina (Ile), cisteína (Cys), histidina (His), arginina (Arg), asparagina (Asp), e treonina (Thr) apresentaram alta variação de eficácia de predição dentre os diferentes métodos utilizados.

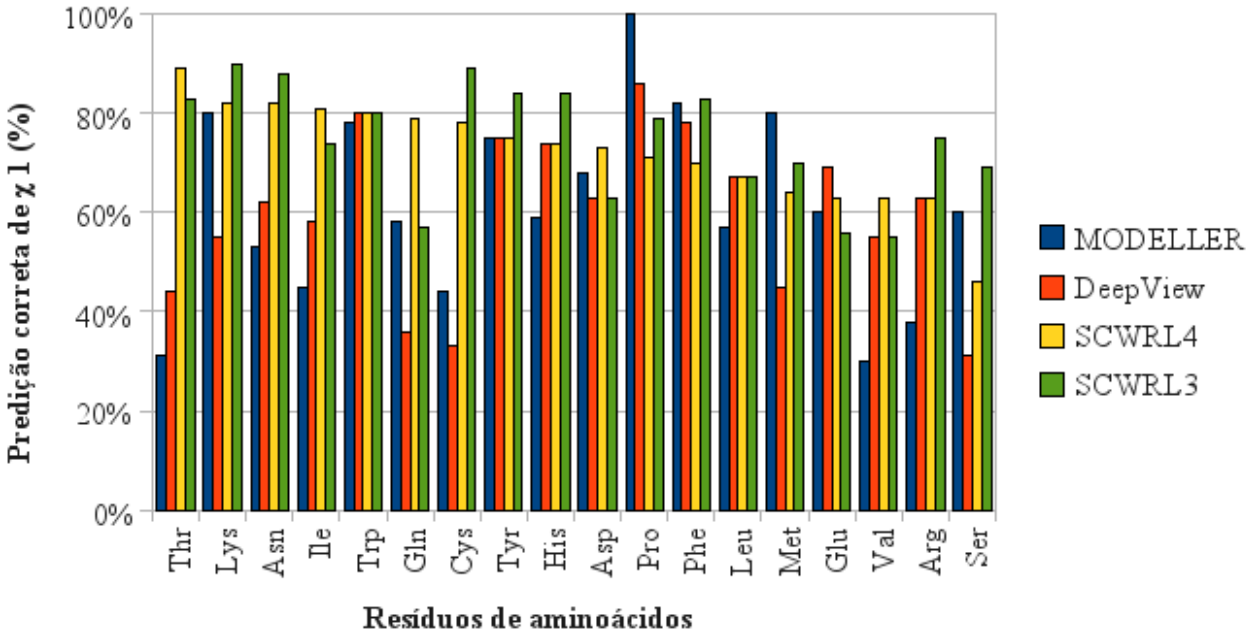


Figura 12. Precisão da modelagem de resíduos de aminoácidos referentes ao ângulo diedro χ1, em função do tipo de resíduo de aminoácido estudado, utilizando os programas MODELLER, DeepView, SCWRL3 e SCWRL4.

Outro fator que pode influenciar a precisão da modelagem é o grau de acessibilidade ao solvente dos resíduos de aminoácidos modelados, o qual reflete o grau de restrição para a busca de novas conformações para as cadeias laterais dos resíduos de aminoácidos da proteína. Foi feita a avaliação da precisão da modelagem aplicada a dois conjuntos diferentes de resíduos de aminoácidos: acessíveis ao solvente (AS), e inacessíveis ao solvente (IS). A acessibilidade ao solvente foi calculada usando-se o programa MODELLER [Sali et al., 1993]. A área fracionária da superfície foi obtida dividindo-se a área de contato de um dado resíduo de aminoácido pela área de contato padrão do resíduo correspondente no tripeptídeo Gly-X-Gly, onde X representa o dado resíduo de aminoácido. Resíduos que apresentaram uma fração da área de superfície acessível ao solvente menor ou igual a 30% em relação ao resíduo de aminoácido isolado foram incluídos no conjunto IS, e aqueles que apresentaram valores maiores que 30% foram incluídos no conjunto AS [Feyfant et al., 2007].

Considerando-se todos os resíduos modelados (AS e IS), os programas SCWRL3 e SCWRL4 apresentaram maiores níveis de predição de acerto em comparação aos programas DeepView e MODELLER. A predição dos ângulos diedros χ1, χ2 e χ1+2 preditos corretamente foi de 73%, 40% e

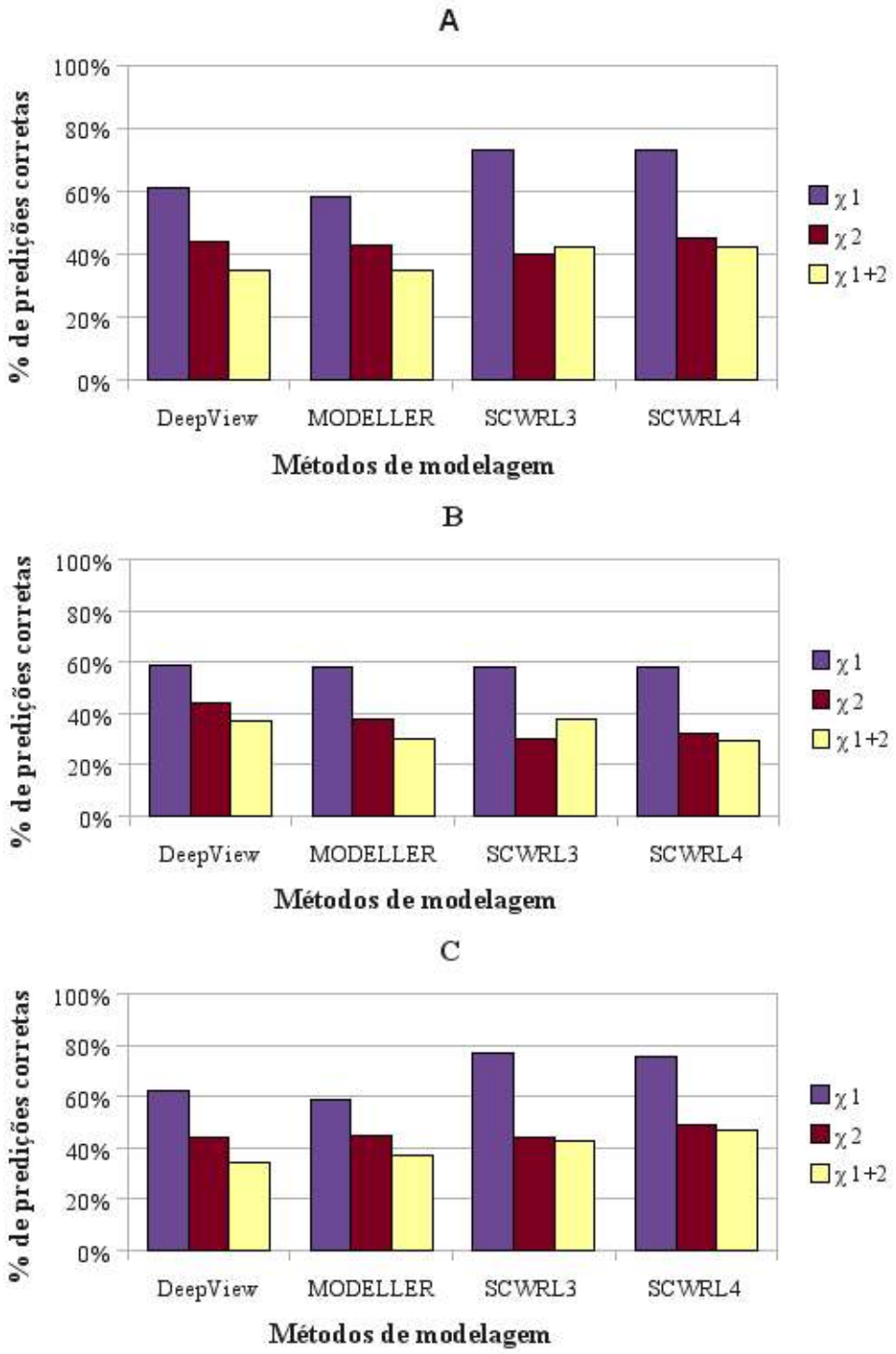


Figura 13. Precisão da modelagem de resíduos de aminoácidos referentes aos ângulos diedro χ1, χ2, e χ1+2 considerando todos os resíduos da proteína (A), apenas resíduos acessíveis ao solvente (B), e apenas resíduos inacessíveis ao solvente, utilizando os programas MODELLER, DeepView, SCWRL3 e SCWRL4.

42% para o SCWRL3, e 73%, 45% e 42% para o SCWRL4, respectivamente (Fig. 13A). O programa SCWRL4 apresentou um aumento de 5% na eficácia de predição do ângulo diedro $\chi 2$ com relação à versão anterior do programa.

Comparando-se os resultados obtidos de predição de ângulos diedros das cadeias laterais entre os conjuntos de resíduos de aminoácidos AS e IS, podemos ver que todos os programas apresentaram níveis de acerto mais altos para o conjunto IS do que para o conjunto AS (Fig. 13B e C). Como esperado, no caso do conjunto IS, a cadeia lateral dos resíduos de aminoácidos tem que se adaptar a um número limitado de conformações, devido a restrições estéricas em um ambiente inacessível ao solvente. Em contraste, no conjunto AS, as cadeias laterais dos resíduos de aminoácidos podem apresentar, em princípio, um número maior de conformações do que aquelas do conjunto IS, devido a um ambiente menos restritivo [Dunbrack et al., 1994; Feyfant et al., 2007].

Portanto, o espaço conformacional é menor no conjunto IS do que no conjunto AS, aumentando assim a probabilidade de uma predição correta, principalmente no caso de resíduos que apresentam menor liberdade conformacional, como mencionado anteriormente.

5.2 Avaliação da precisão do programa de *docking* molecular Autodock 4.0

Em estudos de *docking* molecular, a precisão da predição de afinidade de ligação é geralmente avaliada pelo método de *re-docking* de ligantes em complexos cristalizados que possuem dados experimentais de afinidade de ligação disponíveis. Assim, usando-se como receptores as mesmas estruturas dos complexos cristalizados, é possível comparar valores experimentais de afinidade de ligação com valores obtidos por programas de *docking* molecular.

Nossa análise baseou-se em um conjunto de 185 complexos cristalizados proteína/ligante (descrição dos dados utilizados no Apêndice 8.3) que apresentam dados experimentais de afinidade de ligação (pKi) descritos na base de dados PDBBind [Wang et al., 2005]. Foram escolhidos apenas complexos contendo ligações não-covalentes entre o ligante e a proteína, e cuja resolução cristalográfica era menor ou igual a 2,5 Å. Usando estas estruturas protéicas e seus respectivos ligantes cristalizados, fizemos o *re-docking* molecular utilizando o programa AutoDock 4.0 com os parâmetros descritos na Tabela 2.

Tabela 2. Parâmetros utilizados para o docking molecular utilizando o programa AutoDock 4.0.

Parâmetro	**Descrição**
tran0 random	# coordenadas iniciais do ligante
axisangle0 random	# orientação inicial do ligante
rmstol 2.0	# tolerância do cluster (Å)
ga_pop_size 150	# número de indivíduos na população
ga_num_evals 1000000	# número máximo de avaliações de energia
ga_num_generations 270000	# número máximo de gerações
ga_elitism 1	# número de indivíduos *top* que sobrevivem à próxima geração
ga_mutation_rate 0.02	# taxa de mutação gênica
ga_crossover_rate 0.8	# taxa de crossover
sw_max_its 300	# número de iterações da busca local Solis & Wets
ga_run 100	# número de cálculos GA-LS

Os resultados obtidos pelo *re-docking* molecular foram comparados com os dados experimentais, obtendo-se um coeficiente de correlação de Pearson igual a 0,47 (Fig. 14). Na maioria dos casos (80%), o valor de Energia Livre de Ligação (ΔG) calculado pelo Autodock 4.0 foi mais

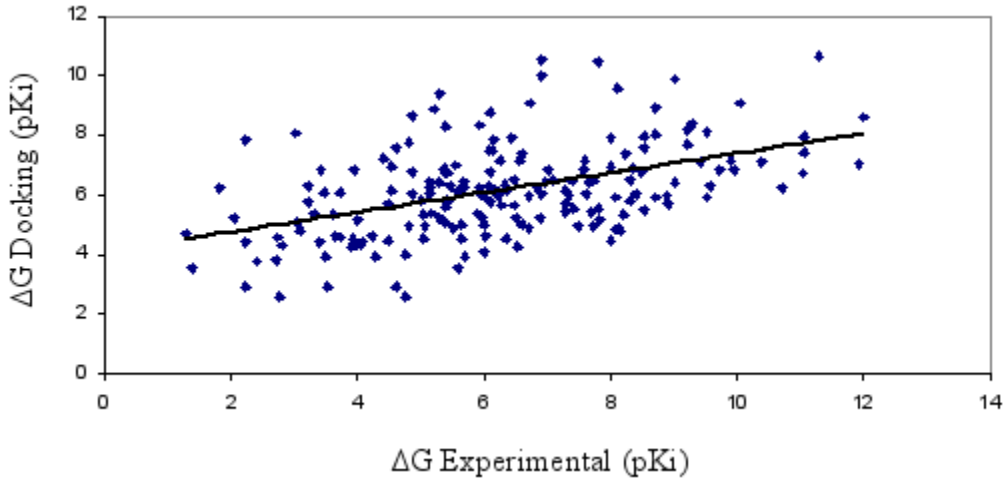


Figura 14. Correlação entre valores experimentais de Energia de Ligação (pKi) e valores de Energia Livre de Ligação (pKi) obtidos pelo *docking* molecular de 185 complexos ligante/proteína usando o programa AutoDock 4.0. O coeficiente de correlação obtido foi igual a 0,47

negativo do que o valor experimental. Tal discrepância se deve não apenas às limitações da função de *scoring* usada pelo programa, mas também a condições experimentais específicas que podem influenciar a afinidade de um ligante por um receptor, tais como o pH da solução, ou o estado tautomérico do ligante.

Além da predição de afinidade de ligação, outro resultado de grande relevância fornecido pelo *docking* molecular é o modo de ligação da estrutura do ligante no sítio ativo do receptor. Existem atualmente poucos estudos comparativos entre modos de ligação de ligantes em complexos cristalizados e modos de ligação preditos pelo método de *docking* molecular [Kolb e Irwin, 2009]. Assim, é de nosso interesse uma comparação conformacional dos resultados obtidos.

Na busca pelo modo de ligação correto de um ligante no sítio ativo de um receptor, estudos de *docking* molecular são frequentemente definidos como precisos quando o cálculo de RMSD resultante da sobreposição do modo de ligação do ligante cristalizado e o modo de ligação mais bem classificado pelo *docking* molecular apresenta um valor inferior a 2,0 Å [Goto et al., 2008]. Fizemos uma comparação entre os modos de ligação dos ligantes cristalizados com aqueles obtidos pelo *re-docking* molecular através da análise dos valores de RMSD resultantes da sobreposição destas estruturas, que é fornecido pelo arquivo de saída do AutoDock 4.0.

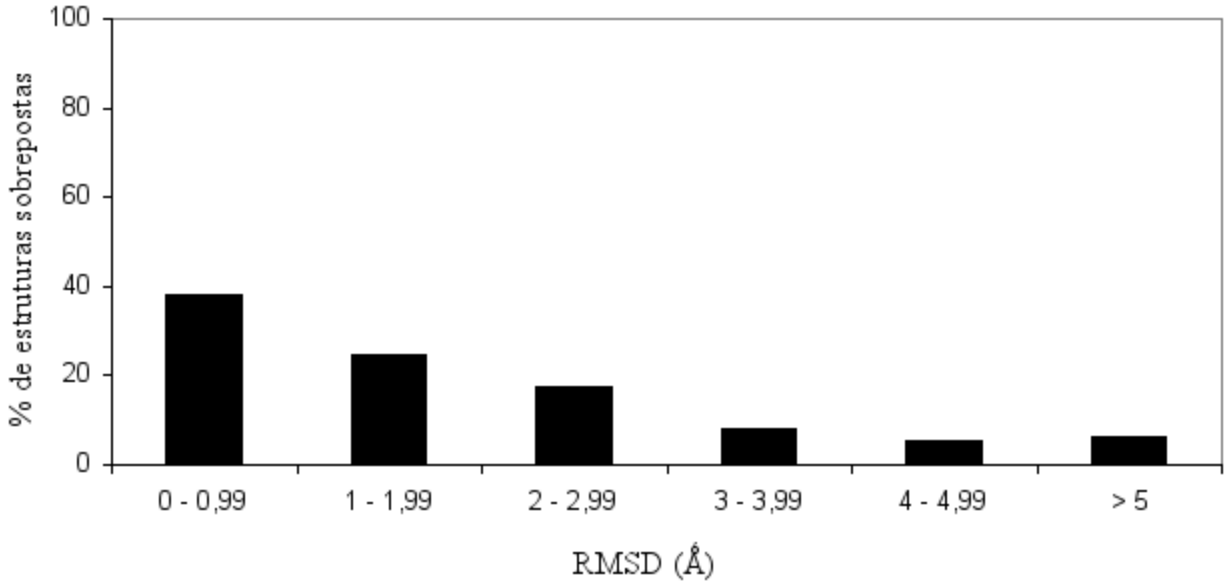


Figura 15. Distribuição dos valores de RMSD resultantes da sobreposição dos modos de ligação obtidos pelo *docking* molecular com suas respectivas estruturas cristalizadas.

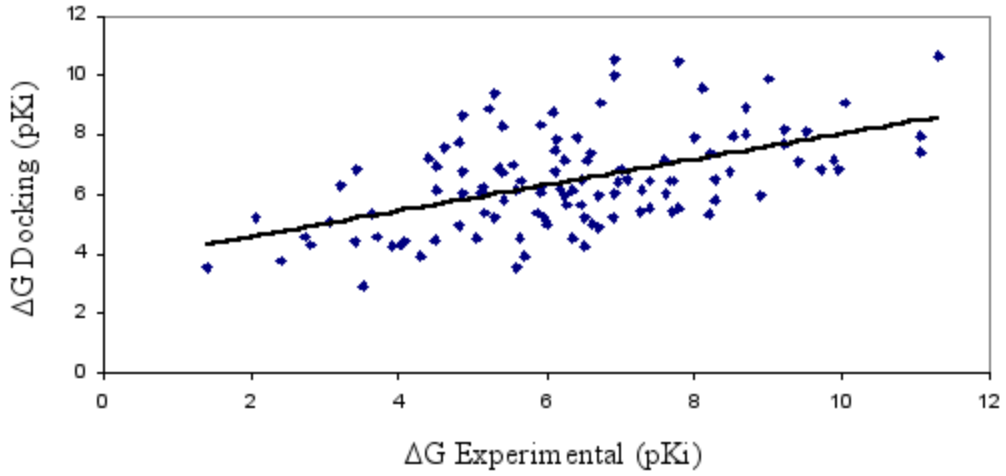


Figura 16. Correlação entre valores experimentais de Energia de Ligação (pKi) e valores de Energia Livre de Ligação (pKi) obtidos pelo *docking* molecular, considerando apenas resultados de *docking* molecular que apresentaram valores de RMSD de sobreposição abaixo de 2,0 Å em relação à estrutura cristalizada. O coeficiente de correlação obtido foi de 0,53.

Dentre 185 simulações de *re-docking* molecular feitos, um total de 118 (63 %) apresentaram valores de RMSD de sobreposição abaixo de 2,0 Å com relação às estruturas cristalizadas (Fig. 15), mostrando que na maioria dos casos a predição do modo de ligação do ligante foi satisfatória. Ao considerarmos apenas os resultados de afinidade de ligação destas simulações, a correlação torna-se maior (coeficiente de correlação = 0,53) (Fig. 16) do que aquela obtida anteriormente (coeficiente de correlação = 0,47), onde foram consideradas todas as simulações. Este resultado sugere que um aumento na precisão de predição do modo de ligação do ligante também contribui para um aumento na precisão da predição de afinidade de ligação do *docking* molecular.

Usando o programa NEQUIM Contact System (NCS), fizemos também a comparação entre os *fingerprints* de interação correspondentes aos modos de ligação obtidos pelo docking molecular e aqueles de suas respectivas estruturas cristalizadas (Fig. 17). Podemos ver que um total de 133 (72%) modos de ligação obtidos pelo *docking* molecular apresentaram valores de Coeficiente de Tanimoto maiores do que 0,6.

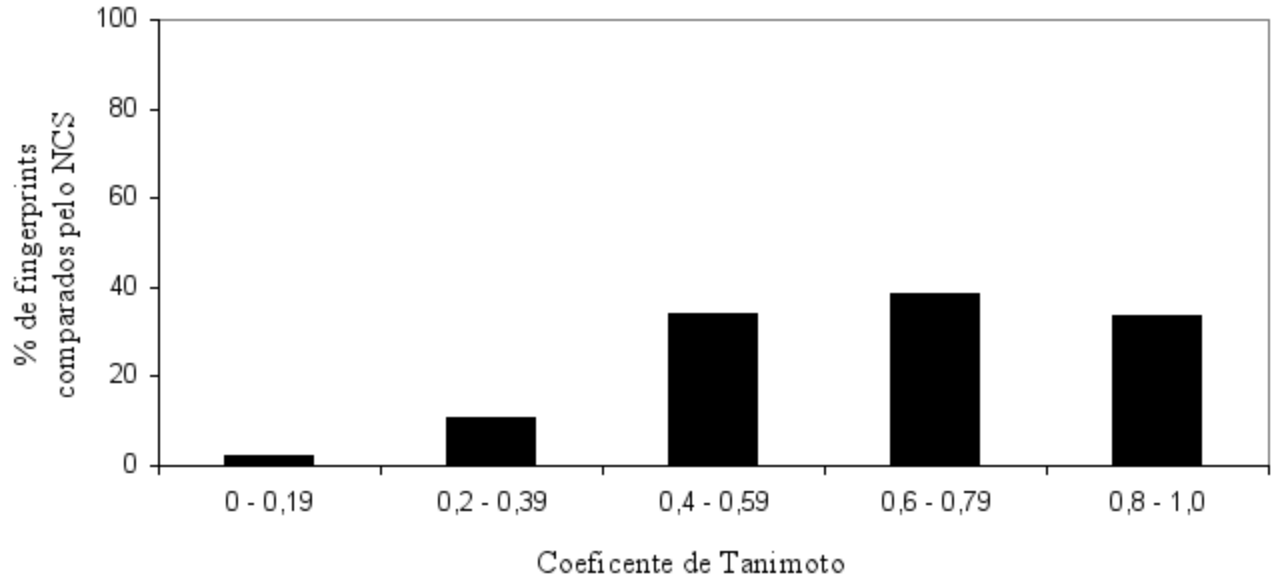


Figura 17. Distribuição dos valores de Coeficiente de Tanimoto resultantes de estudo de comparação de *fingerprints* dos modos de ligação obtidos pelo *docking* molecular e aqueles de suas respectivas estruturas cristalizadas, usando o programa NEQUIM Contact System (NCS) [José et al., 2008].

Em outro estudo, buscamos determinar o efeito do número de torções (ou graus de liberdade) das moléculas de ligante na precisão de cálculo do programa AutoDock 4.0. Na Figura 18, apresentamos os resultados de correlação experimental versus computacional do estudo de *re-docking* descrito acima, plotando separadamente diferentes grupos, definidos pelo número de torções dos ligantes estudados: 0-4, 5-9, 10-14, e >15 torções.

Observamos que a precisão do cálculo de afinidade de ligação pelo AutoDock 4.0 diminui à medida que o número de torções dos ligantes estudados aumenta para os grupos 0-4, 5-9 e 10-14. Já para o grupo de ligantes que apresenta mais de 15 torções, o valor da correlação é ligeiramente maior do que nos grupos 5-9 e 10-14.

Decidimos então testar a hipótese de que ligantes que apresentam números altos de torções requerem um número mais alto de avaliações de energia pelo algoritmo do AutoDock 4.0 para atingir resultados mais precisos, e até que ponto este aumento pode ser benéfico. Refizemos o estudo de *re-docking,* aumentando gradualmente o parâmetro referente ao número de avaliações de energia (ga_nums_evals) de acordo com o aumento no número de torções dos ligantes estudados (Tabela 3).

Tabela 3. Valores do parâmetro referente ao número de avaliações de energia (ga_nums_evals) utilizados para grupos de ligantes com diferentes graus de liberdade.

No. de graus de liberdade do ligante	No. de avaliações de Energia
0-4	2000000
5-9	4000000
10-14	6000000
>15	8000000

Como pode-se observar na Figura 19, neste estudo houve um aumento da correlação para os grupos de ligantes que possuem 0-4, 5-9, e 10-14 torções (0,64, 0,52, e 0,39), em comparação ao estudo anterior (0,55, 0,38, e 0,36, respectivamente) (Fig. 18) em que foi utilizado o mesmo número de avaliações de energia (1000000), independentemente do número de torções dos ligantes estudados. Quanto ao grupo de ligantes que apresenta mais de 15 torções, houve uma diminuição da correlação.

Considerando todos os resultados deste estudo em um único gráfico de correlação, observamos que houve um aumento no coeficiente de correlação de 0,47 no estudo anterior (Fig. 14) para 0,59 neste estudo (Fig. 20) confirmando a eficácia de adaptar o parâmetro referente ao número de avaliações de energia de acordo com o número de torções dos ligantes estudados.

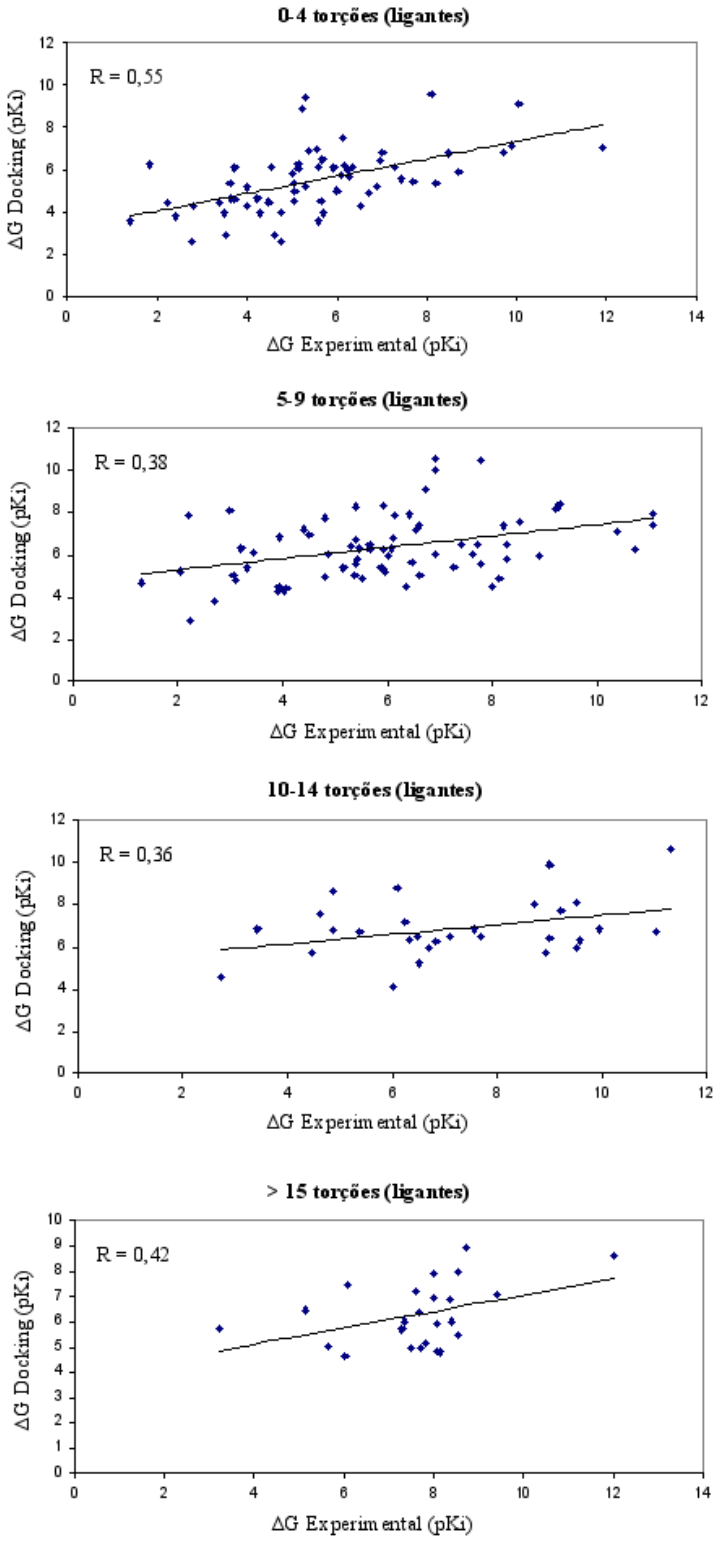


Figura 18. Correlação entre ΔG experimental e computacional do estudo de *re-docking*, plotando separadamente diferentes grupos, definidos pelo número de torções dos ligantes estudados: 0-4, 5-9, 10-14, e > 15 torções.

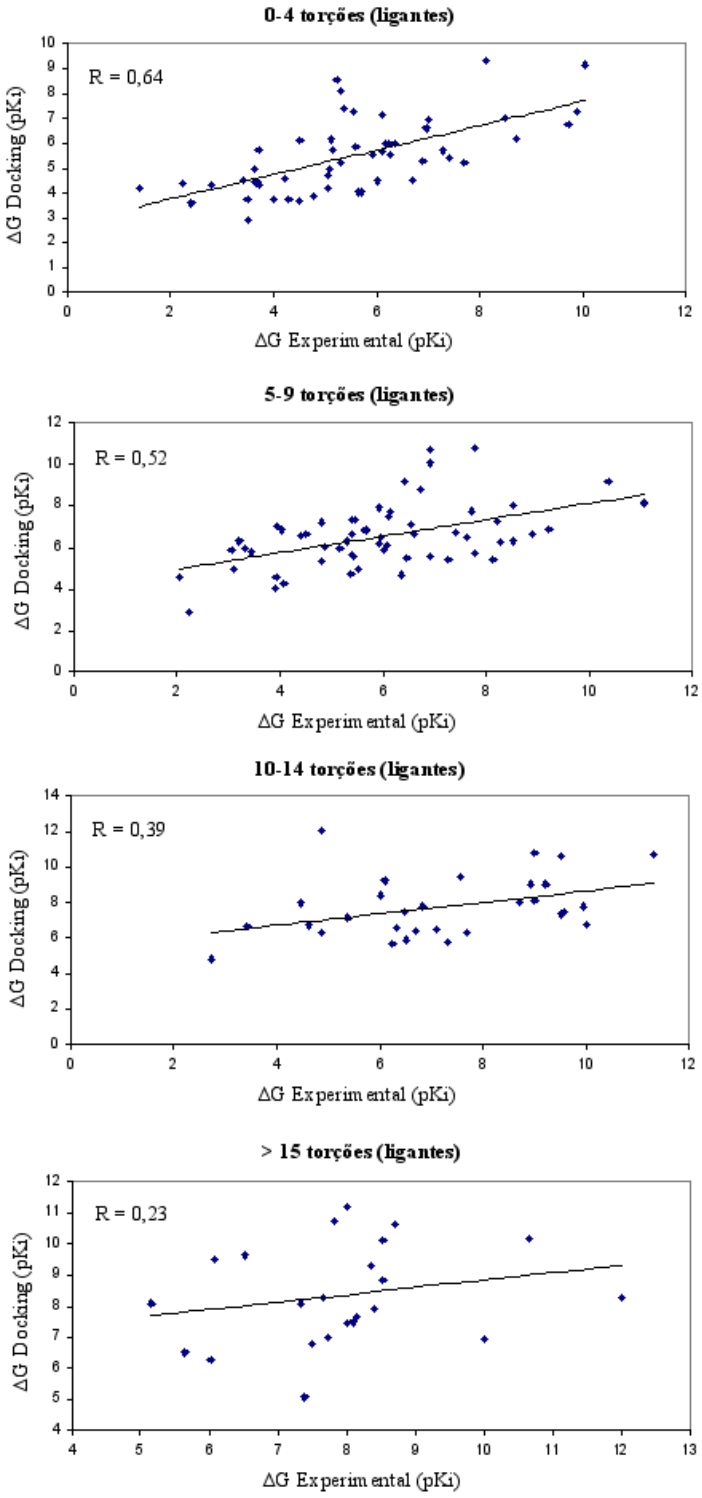


Figura 19. Correlação experimental versus computacional do estudo de *re-docking*, aumentando gradualmente o parâmetro referente ao número de avaliações de energia (ga_nums_evals) de acordo com o aumento no número de torções dos ligantes estudados.

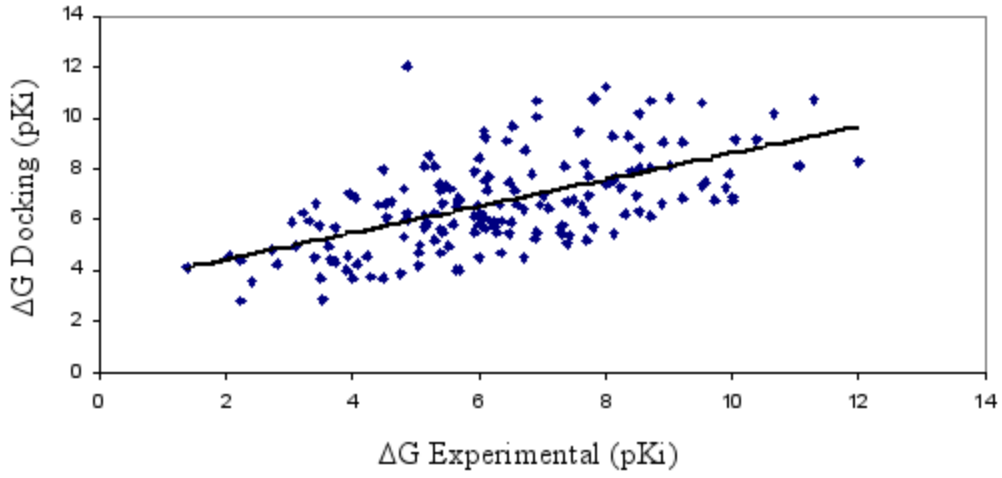


Figura 20. Correlação entre valores experimentais de Energia de Ligação (pKi) e valores de Energia Livre de Ligação (pKi) obtidos pelo *docking* molecular de 185 complexos ligante/proteína usando o programa AutoDock 4.0. Para ligantes que apresentaram 0-4, 5-9, 10-14, e >15 torções, foram feitas 2000000, 3000000, 4000000, e 5000000 avaliações de energia, respectivamente. O coeficiente de correlação obtido foi igual a 0,59.

Na tentativa de encontrar outros parâmetros que poderiam influenciar o resultado de *docking* molecular devido a diferenças em números de torções em ligantes, foi analisado também o impacto causado por mudanças nos parâmetros referentes ao número de indivíduos na população, número máximo de gerações, número de iterações da busca local Solis & Wets, e número de cálculos GA-LS. No entanto, as mudanças nestes parâmetros não influenciaram os resultados obtidos.

5.3 Controle da variação de resultados de afinidade em simulações de *docking* molecular repetidos

Em geral, repetições de simulações de *docking* molecular usando os mesmos parâmetros e as mesmas moléculas de entrada podem gerar certa variação no resultado de Energia Livre de Ligação (ΔG). O programa AutoDock 4.0 utiliza um gerador de números aleatórios para criar o modo de ligação inicial da molécula do ligante. Estes números aleatórios definem a localização, orientação, e valores de torção do ligante na grade de busca. Portanto, como estes valores aleatórios podem ser diferentes entre os cálculos de *docking* molecular em simulações repetidas, o processo de busca do algoritmo gera conformações aleatórias, podendo então gerar resultados diferentes.

Com o objetivo de contornar este problema analisamos o efeito de mudanças nos parâmetros referentes ao número de avaliações de energia e número máximo de gerações nos resultados de afinidade de ligação em simulações repetidas de *docking* molecular.

Elaboramos seis protocolos diferentes (Tabela 4), onde cada um apresentava parâmetros diferentes de avaliações de energia ou número máximo de gerações, e repetimos 50 vezes uma dada simulação de *docking* molecular (proteína 2hyy e ligante Imatinib) usando cada um dos protocolos.

Tabela 4. Protocolos utilizados em experimentos de docking molecular repetidos.

	Protocolo 1	**Protocolo 2**	**Protocolo 3**	**Protocolo 4**	**Protocolo 5**	**Protocolo 6**
No. de avaliações de Energia	1000000	2000000	1000000	3000000	4000000	5000000
No. máximo de gerações	270000	270000	540000	270000	270000	270000

Os resultados obtidos mostraram que, mantendo o número máximo de gerações constante (270000), o aumento no número de avaliações de energia reduziu o desvio padrão dos resultados de afinidade de ligação em simulações repetidas de *docking* molecular, atingindo o mínimo de variação quando foi usado o valor de 4000000 avaliações de energia (Tabela 5). No entanto, podemos observar que o aumento no número de avaliações de energia de 4000000 para 5000000 causou um aumento no desvio padrão. Com relação ao parâmetro referente ao número máximo de gerações, o aumento no

valor deste parâmetro não causou uma diminuição da variação dos resultados repetidos, e a diferença do valor da média obtido em relação à média obtida para o protocolo 1 não foi estatisticamente significante, como mostrado pelo valor P após o teste T. Com relação aos resultados obtidos utilizando os protocolos 2, 4, 5 e 6, podemos ver que o valor P obtido mostra que a diferença dos valores da média em relação ao protocolo 1 obtidos foram estatisticamente significantes (Valores P em negrito na Tabela 5).

Este estudo mostrou que o aumento no parâmetro de avaliações de energia, até certo ponto (no caso, 4000000 de avaliações de energia), aumenta a probabilidade do algoritmo de busca encontrar o mínimo de energia, possibilitando assim uma busca mais abrangente pelo modo de ligação correto.

Considerando o estudo anterior sobre a necessidade de padronizar o parâmetro de avaliações de energia de acordo com o número de torções dos ligantes, o ligante utilizado neste estudo (Imatinib) possui 7 graus de liberdade, e seria portanto recomendado 4000000 avaliações de energia para este ligante. O valor médio de Energia Livre de Ligação (-10,33 Kcal/mol) obtido quando utilizamos 4000000 avaliações de energia é bastante próximo do valor experimental obtido para o complexo cristalizado (-10,37 Kcal/mol) [Pricl et al., 2005]. Portanto, estes resultados sugerem que a otimização do parâmetro referente ao número de avaliações de energia de acordo com o número de torções dos ligantes estudados aumenta a probabilidade de se encontrar o mínimo de energia local, e também reduz a variação de resultados de afinidade em simulações de docking molecular repetidas, consequentemente resultando em cálculos mais precisos.

Tabela 5. Resultados obtidos de Energia Livre de Ligação (ΔG) para seis protocolos diferentes, variando-se o número de avaliações de energia e número máximo de gerações, e repetindo-se cada protocolo um número total de 50 vezes.

Experimento	Protocolo 1	Protocolo 2	Protocolo 3	Protocolo 4	Protocolo 5	Protocolo 6
1	-9,65	-10,23	-9,98	-10,58	-10,51	-10,31
2	-9,65	-10,33	-9,73	-10,22	-10,16	-10,32
3	-9,67	-10,08	-10,01	-10,45	-10,39	-10,2
4	-9,68	-10,04	-9,89	-10,04	-10,44	-10,17
5	-9,7	-10,16	-9,61	-10,44	-10,22	-10,2
6	-9,74	-10,4	-9,8	-10,21	-10,26	-10,33
7	-9,76	-10,45	-9,84	-10,17	-10,26	-10,09
8	-9,78	-10,06	-10,24	-10,43	-10,29	-10,52
9	-9,79	-9,89	-9,64	-10,47	-10,4	-10,38
10	-9,83	-10,09	-9,69	-10,31	-10,26	-10,45

13	-9,84	-10,18	-9,75	-10,37	-10,15	-10,56
14	-9,84	-10,44	-10,18	-10,35	-10,29	-10,53
15	-9,86	-10,08	-9,83	-10,19	-10,18	-10,63
16	-9,86	-10	-9,41	-10,25	-10,26	-10,23
17	-9,86	-10,21	-9,96	-10,44	-10,45	-10,59
18	-9,88	-10,4	-10,04	-10,32	-10,32	-10,53
19	-9,88	-10,07	-9,89	-10,14	-10,3	-10,46
20	-9,89	-9,98	-9,79	-10,13	-10,49	-10,32
21	-9,9	-10,04	-10	-10,34	-10,47	-10,19
22	-9,91	-10,02	-9,68	-10,08	-10,29	-10,48
23	-9,92	-10,12	-10,39	-10,63	-10,52	-10,31
24	-9,93	-10,36	-9,99	-10,2	-10,26	-10,52
25	-9,93	-10,28	-10,37	-10,16	-10,26	-10,35
26	-9,93	-10,08	-10,12	-10,29	-10,41	-10,48
27	-9,94	-10,21	-9,87	-10,2	-10,4	-10,14
28	-9,95	-10,14	-9,87	-10,2	-10,22	-10,34
29	-9,95	-10,15	-10,15	-10,42	-10,3	-10,61
30	-9,96	-10,34	-9,96	-10,27	-10,54	-10,43
31	-9,96	-10,09	-9,62	-9,98	-10,37	-10,39
32	-9,97	-10,32	-9,82	-10,41	-10,35	-10,39
33	-9,99	-10,04	-10,12	-10,34	-10,12	-10,57
34	-10	-10,16	-9,9	-10,21	-10,62	-10,28
35	-10,02	-10,18	-9,69	-10	-10,47	-10,23
36	-10,02	-9,91	-10,04	-10,12	-10,53	-10,35
37	-10,04	-10,07	-9,98	-10,19	-10,49	-10,3
38	-10,08	-10,38	-9,94	-10,22	-10,22	-10,4
39	-10,08	-10,34	-10,02	-10,3	-10,21	-10,54
40	-10,1	-10,1	-9,8	-10,33	-10,16	-10,27
41	-10,1	-10,02	-9,89	-10,29	-10,35	-10,64
42	-10,12	-10,53	-10,09	-10,35	-10,28	-10,27
43	-10,13	-10,07	-10,06	-10,15	-10,39	-10,54
44	-10,13	-10,13	-9,72	-10,39	-10,03	-10,27
45	-10,15	-10,24	-10,01	-10,29	-10,19	-10,48
46	-10,18	-10,03	-9,88	-10,26	-10,27	-10,25
47	-10,24	-10,17	-10,03	-10,07	-10,39	-10,42
48	-10,26	-10,39	-9,98	-10,2	-10,27	-10,14
49	-10,38	-10,35	-9,92	-10,35	-10,28	-10,4
50	-10,39	-10,35	-9,91	-10,24	-10,29	-10,39
Média	-9,95	-10,18	-9,92	-10,27	-10,33	-10,38
Desvio Padrão	0,17	0,16	0,19	0,14	0,13	0,14
Dif. da Média em relação ao Protocolo 1	-	0,23	-0,03	0,32	0,38	0,43
Valor P (em comparação ao Protocolo 1)	-	0,0001	0,4751	0,0001	0,0001	0,0001

5.4 Avaliação da capacidade do programa AutoDock 4.0 de detectar mutações pontuais que alteram a afinidade de ligação

Nesta etapa, apresentamos um estudo de caso que foi feito a fim de verificar se o programa AutoDock 4.0 pode ser utilizado para detectar substituições de resíduos de aminoácidos que causam impacto na afinidade de ligação entre ligantes e alvos terapêuticos.

Considerando que a função de *scoring* utilizada para predição de valor de Energia Livre de Ligação (ΔG) pelo AutoDock 4.0 apresenta um erro padrão de ~2,2 Kcal/mol, segundo estudos de *re-docking* realizados pelos criadores do programa, utilizando uma grande variedade de complexos cristalizados [Morris et al., 1998; Huey et al., 2007], apenas substituições de resíduos de aminoácidos que resultam em um aumento no valor de ΔG maior do que 2,2 Kcal/mol em relação à estrutura protéica nativa podem ser definidas como potencialmente capazes de afetar diretamente a afinidade de ligação de um ligante.

Os resultados deste estudo foram apresentados em pôster na conferência internacional *Intelligent Systems for Molecular Biology*, realizada em Fortaleza no ano de 2006.

5.4.1 Estudo de Caso

Uma abordagem computacional para o estudo do efeito de mutações pontuais no domínio ABL da tirosina quinase receptora do medicamento Imatinib

1. Introdução

A leucemia mielóide crônica (LMC) resulta de um defeito genético em células tronco hematopoiéticas, caracterizado por uma translocação recíproca entre os cromossomos 9 e 22, formando o cromossomo Filadélfia (cromossomo Ph) [Nowell et al., 1960; Rowley, 1973], detectado em mais de 90% dos pacientes com esta doença [Shepherd et al., 1995]. Esta translocação funde uma região do gene *BCR* com porções do gene *ABL,* codificando assim uma proteína quimérica (BCR-ABL) com atividade de tirosina quinase (Fig. 1) [de Klein et al., 1982].

As tirosinas quinase são enzimas que têm como função a transferência de um grupo fosfato de uma molécula de ATP para um resíduo de tirosina em um substrato. Normalmente a estrutura da tirosina quinase ABL ocorre em um estado inativo por um mecanismo de auto inibição [Sicheri et al., 1997; Pluk et al., 2002], onde o domínio SH3 inibe o domínio catalítico (CAT) ao se ligar à região de ligação SH2-CAT. O *cap* terminal-N, composto por um grupo miristato e os domínios 1a e 1b, trava o domínio SH3 nesta configuração ao se ligar a este domínio e ao CAT (Fig. 2A) [Sawyers, 2002a; Nagar et al., 2003]. Como resultado, o *loop* de ativação (loop-A) da enzima é mantido na posição fechada, impedindo a entrada de uma molécula de ATP e de substrato. No entanto, a associação de outras proteínas com estes domínios pode desencadear a abertura do loop-A e a consequente ativação da enzima, permitindo assim a entrada do ATP na região do sítio de ligação da enzima [Nagar et al. 2003].

A enzima quimérica BCR-ABL não possui o *cap* terminal-N, que tem grande importância na regulação da atividade enzimática [Sawyers, 2002a]. No entanto, os domínios SH2, SH3 e o CAT estão presentes, o que possibilita o funcionamento da tirosina quinase (Fig. 2B). Portanto, através da constante fosforilação dos substratos, a BCR-ABL ativa, de forma desregulada, a via de transdução de sinais que irá eventualmente induzir o processo de proliferação celular, sobrevivência celular e diferenciação, levando ao desenvolvimento da LMC.

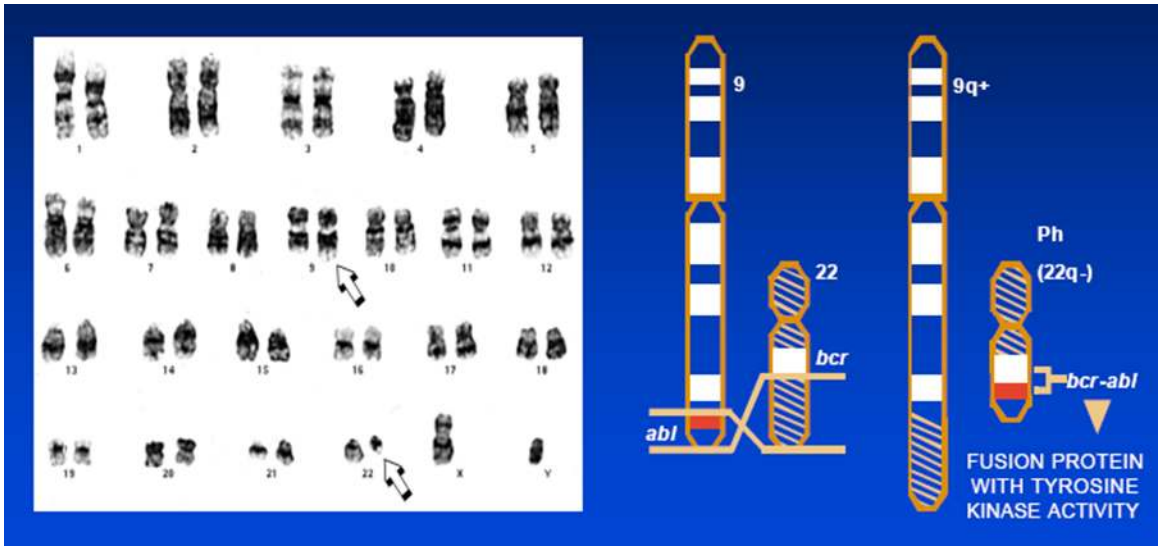


Figura 1. Esquerda: Cariótipo de uma célula tronco hematopoiética de um paciente afetado pela leucemia mielóide crônica. Direita: Translocação recíproca entre os cromossomos 9 e 22, formando o cromossomo Filadélfia (cromossomo Ph), que codifica a proteína quimérica BCR-ABL.

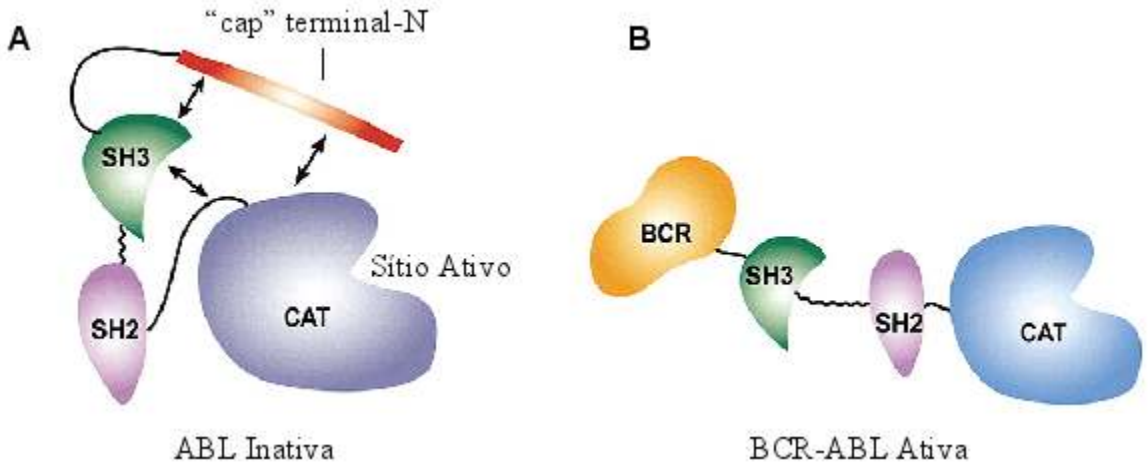


Figura 2. A. Estrutura da enzima ABL na conformação inativa regulada, com o domínio SH3 inibindo o domínio catalítico (CAT) ao se ligar à região de ligação SH2-CAT. O *cap* terminal-N trava o domínio SH3 nesta configuração ao se ligar a este domínio e ao CAT. B. A oncoproteína BCR-ABL não possui o *cap* terminal-N e, portanto, o CAT é ativado de forma desregulada.

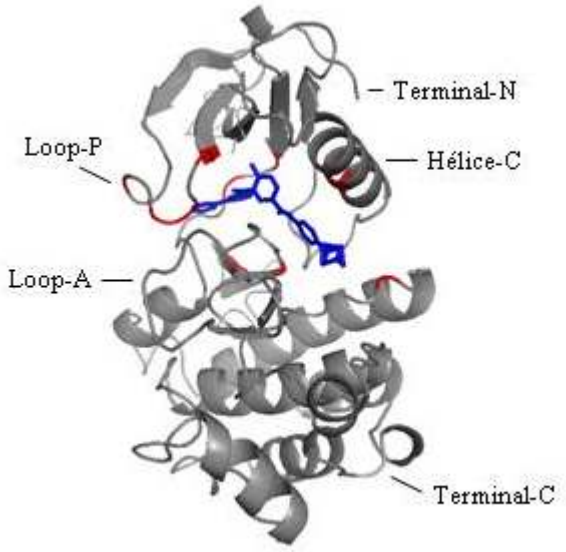
Com o objetivo de tratar pacientes com LMC, o inibidor de tirosina quinase Imatinib (STI571; Gleevec) foi desenvolvido para inibir a atividade da enzima BCR-ABL através da ligação específica e estabilização da forma inativa da enzima mutante, prevenindo a ativação do loop-A [Druker et al., 2001a; Druker et al., 2001b; Schindler et al., 2000]. Estima-se que 90% dos pacientes com LMC em fase inicial da doença apresentam uma resposta hematológica completa após o tratamento com o Imatinib [Druker et al., 2001b; Kantarjian et al., 2002].

No entanto, a maioria dos pacientes tratados nas fases avançadas da LMC apresenta falhas de resposta ou recaídas após uma resposta inicial ao tratamento [Druker et al., 2001a; Sawyers, 2002b; Shah et al., 2002]. Mutações no domínio quinase da enzima são os mecanismos mais associados à resistência, ocorrendo a diminuição da sensibilidade ao Imatinib nestes pacientes [Von Bubnoff et al., 2003].

Ao nível molecular, mutações gênicas que causam substituições de resíduos de aminoácidos na enzima BCR-ABL podem reduzir a afinidade do Imatinib por um mecanismo direto ou indireto. No caso de um mecanismo direto, as mutações podem reduzir a afinidade do Imatinib através de mudanças nas cadeias laterais de resíduos de aminoácidos que contribuem com interações favoráveis à ligação do Imatinib, ou como resultado de mudanças topográficas que afetam a conformação de ligação do Imatinib [Weisberg et al., 2007]. Mutações associadas à resistência ao Imatinib pelo mecanismo indireto reduzem a afinidade do Imatinib através da desestabilização da conformação inativa do loop A, ou através da estabilização da conformação ativa da enzima, resultando em um aumento da atividade da enzima [Roumiantsev et al., 2002; Cowan-Jacob et al., 2007].

Neste estudo, usamos o programa de *docking* molecular AutoDock 4.0 para investigar individualmente o impacto direto causado por 12 mutações diferentes (Gly250Glu, Gln252His, Tyr253Phe, Glu255Lys, Val256Glu, Glu286Leu, Met290Ala, Thr315Ile, Phe317Leu, Phe359Val, Leu370Gly e Val379Ile) na interação com o Imatinib. Todas as mutações estudadas estão localizadas no sítio ativo do domínio ABL da enzima BCR-ABL, e foram descritas na literatura como associadas a diferentes graus de resistência ao Imatinib, apesar de que o mecanismo de resistência de várias mutações ainda não foi confirmado (Fig. 3) [Shah et al., 2002; Corbin et al., 2002; Roche-Lestienne, 2002; Roumiantsev et al., 2002; Branford et al., 2003].

A



B

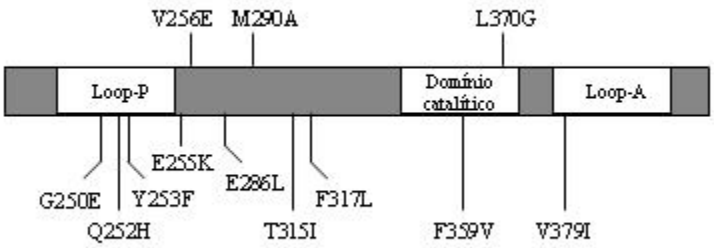


Figura 3. **A**. Representação esquemática do complexo formado pelo Imatinib (azul) e o domínio quinase da enzima BCR-ABL (cinza), mostrando as regiões do loop-A, loop-P, domínio catalítico, terminais N e C. As posições das mutações pontuais estudadas (Gly250Glu, Gln252His, Tyr253Phe, Glu255Lys, Val256Glu, Glu286Leu, Met290Ala, Thr315Ile, Phe317Leu, Phe359Val, Leu370Gly e Val379Ile) estão ilustradas em vermelho. **B**. Posições relativas das mutações pontuais estudadas ao longo da cadeia primária da enzima.

2. Materiais e Métodos

2.1 Modelagem Molecular e Minimização

Usando uma estrutura protéica (cristalizada por raio X) nativa do domínio ABL da enzima BCR-ABL (PDB id 2hyy), foi feita a modelagem de cada uma das mutações de interesse (Gly250Glu, Gln252His, Tyr253Phe, Glu255Lys, Val256Glu, Glu286Leu, Met290Ala, Thr315Ile, Phe317Leu, Phe359Val, Leu370Gly e Val379Ile) separadamente usando o programa SCWRL4, gerando 12 estruturas mutantes modeladas. Em seguida, estas estruturas, e também a estrutura nativa, foram minimizadas usando o programa GROMACS (1000 passos do método de *Steepest Descent* e 500 passos do método de Gradiente Conjugado).

2.2 Preparação da estrutura do inibidor

A molécula tridimensional do medicamento Imatinib utilizada para o *docking* molecular foi obtida do complexo cristalizado do domínio ABL/Imatinib (PDB id 2hyy). Usando o software AutoDockTools (ADT), cargas atômicas parciais foram adicionadas a esta molécula usando o método de Gasteiger Marsili [Gasteiger e Marsili, 1980; Morris et al., 1998].

2.3 *Docking* molecular

Os cálculos de *docking* molecular foram realizados utilizando o software público AutoDock 4.0. Antes do processo de *docking*, mapas de grade representando as energias de interação entre os vários tipos de átomos do inibidor e os átomos de resíduos de aminoácidos no sítio ativo da enzima foram calculados com o pacote AutoGrid do AutoDock 4.0. O centro da grade foi definido como o centro do sítio ativo da enzima. Foi usada uma grade de 70x60x60 ao longo dos eixos X, Y e Z, separada por pontos espaçados por 0,375 Å.

Usando o ADT, átomos de hidrogênio polares foram adicionados geometricamente à estrutura do receptor protéico, e cargas atômicas parciais foram adicionadas utilizando o método de Gasteiger Marsili [Gasteiger e Marsili, 1980; Morris et al., 1998]. O ADT também foi utilizado para designar o número de torções, e para adicionar átomos polares de hidrogênio na molécula do Imatinib.

O algoritmo genético Lamarckiano foi utilizado para busca global nas simulações de docking, e o algoritmo Solis Wets para a otimização local subsequente. A população foi composta por 150 indivíduos, o número máximo de avaliações de energia definido como 4000000, o número máximo de

gerações como 270000, e o número de corridas como 100. Uma taxa de mutação máxima de 0,02, um elitismo de 1, uma taxa de *crossover* de 0,8 e taxa de busca local de 0,06 foram utilizados. Os valores padrão foram usados para todos os parâmetros restantes.

3. Resultados e Discussão

O resultado obtido pelo *docking* molecular do Imatinib com a estrutura nativa do receptor mostra que o modo de ligação obtido por esta simulação se assemelha àquele obtido pela cristalização do complexo da estrutura nativa contendo o Imatinib em seu sítio ativo. A sobreposição dos dois modos de ligação revela um valor de RMSD (*root mean square deviation*) igual a 1,5Å (Fig. 4).

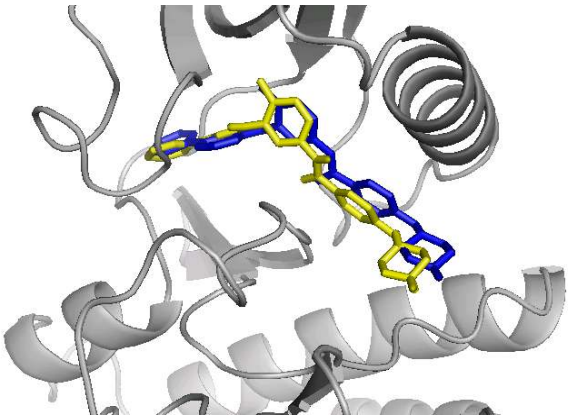


Figura 4. Sobreposição dos modos de ligação do Imatinib obtidos por cristalização (azul) e docking (amarelo) (RMSD = 1,5 Å).

Dentre as 12 estruturas mutantes estudadas, apenas a estrutura que contém a mutação Thr315Ile apresentou uma diferença significativa de afinidade de ligação com o Imatinib (+3,00 Kcal/mol) em relação ao resultado obtido utilizando-se a estrutura nativa (Tabela 1). A mutação de uma treonina para uma isoleucina na posição 315 causa uma alteração no número de ligações de hidrogênio presentes no complexo cristalizado (Fig. 5), resultando na diminuição da afinidade de ligação deste fármaco, conforme relatos experimentais já evidenciaram [Cowan-Jacob et al., 2007], sendo assim uma das mutações mais comuns em pacientes resistentes ao Imatinib, [Shah et al., 2002].

Com relação às outras mutações, apesar de estarem na região de ligação com o Imatinib, a diferença de afinidade de ligação em relação ao resultado obtido com a estrutura nativa para todas elas

está dentro do valor do erro padrão da função de *scoring* do AutoDock 4.0 (~2,2 Kcal/mol) [Morris et al., 1998; Huey et al., 2007] e, portanto, esta abordagem não foi capaz de notar perturbações significativas para estas mutações.

Tabela 1. Resultados do *docking* molecular da interação entre o Imatinib e 13 estruturas diferentes do domínio ABL da tirosina quinase. Em negrito, a maior diferença de energia em relação à estrutura nativa, referente à mutação Thr315Ile.

Estruturas	ΔG Docking (Kcal/ mol)	Diferença de Energia em relação à estrutura nativa (Kcal/mol)
Nativa	-10,33 (-10,37)*	-
Gly250Glu	-10,13	0,21
Gln252His	-10,02	0,32
Tyr253Phe	-11,27	-0,93
Glu255Lys	-10,39	-0,05
Val256Glu	-9,6	0,74
Glu286Leu	-9,14	1,2
Met290Ala	-10,32	0,02
Thr315Ile	-7,34 (-7,23)*	**3,00**
Phe317Leu	-10,15	0,19
Phe359Val	-9,72	0,62
Leu370Gly	-11,27	-0,93
Val379Ile	-9,88	0,46

* Valores experimentais de ΔG obtidos de artigo de Pricl et al. aparecem em parênteses [Pricl et al., 2005]

O pressuposto básico desta abordagem é que as mutações não causam grandes mudanças conformacionais na estrutura da proteína. Através do método de minimização de energia das estruturas mutantes modeladas, é possível determinar o mínimo local de energia da estrutura modelada. No entanto, este método não é adequado para se determinar mutações que poderiam causar uma desestabilização da conformação inativa do loop-A, podendo consequentemente afetar indiretamente a interação com o Imatinib.

O método mais adequado para este tipo de estudo seria a dinâmica molecular, através do qual é possível gerar estruturas que seriam então usadas para o *docking* molecular, possibilitando assim uma avaliação mais precisa do impacto causado por mutações que possivelmente causam grandes mudanças conformacionais na estrutura protéica. De fato, tal efeito foi estudado através de simulações de dinâmica molecular com diversas mutações no sítio ativo do domínio ABL da enzima BCR-ABL, onde

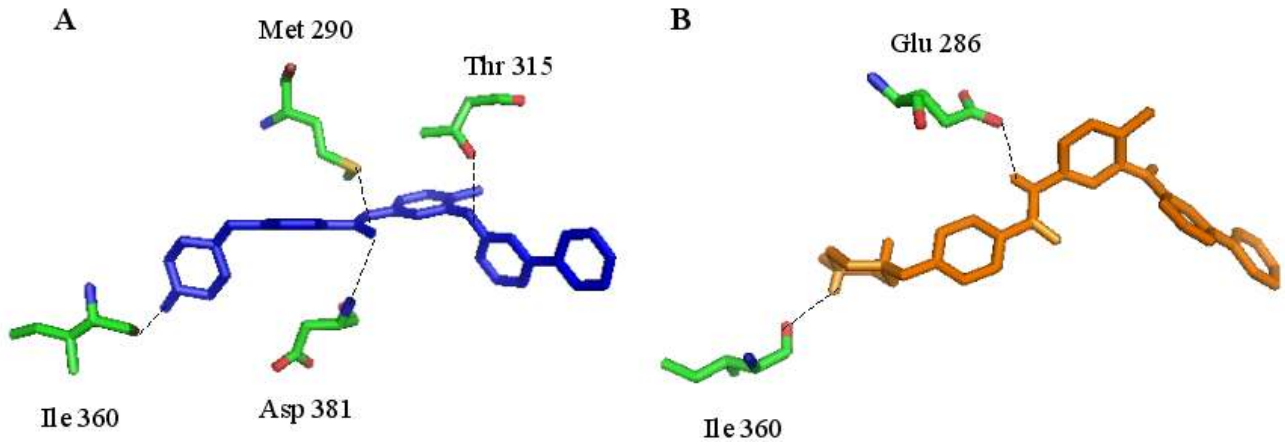


Figura 5. Comparação dos resíduos de aminoácido que fazem ligações de hidrogênio com o modo de ligação da molécula de Imatinib na estrutura cristalizada nativa (A) (Met290, Thr315, Ile360 e Asp381) e com o modo de ligação resultante do *docking* molecular com a estrutura que contém a mutação Thr315Ile (B) (Glu286 e Ile360), que resulta na perda de duas ligações de hidrogênio.

foi sugerido que várias destas mutações causam um impacto indireto na interação com o Imatinib [Lee et al., 2008].

Dito isto, os resultados do presente trabalho sugerem que, através da modelagem molecular, seguida da minimização da estrutura mutante e do *docking* molecular, é possível detectar mutações que afetam diretamente a afinidade entre um receptor em um ligante (Fig. 6). Espera-se que o aprimoramento e refinamento das funções de scoring dos programas de *docking* molecular poderá, futuramente, possibilitar também a detecção de mutações que causam impactos sutis na afinidade de ligação.

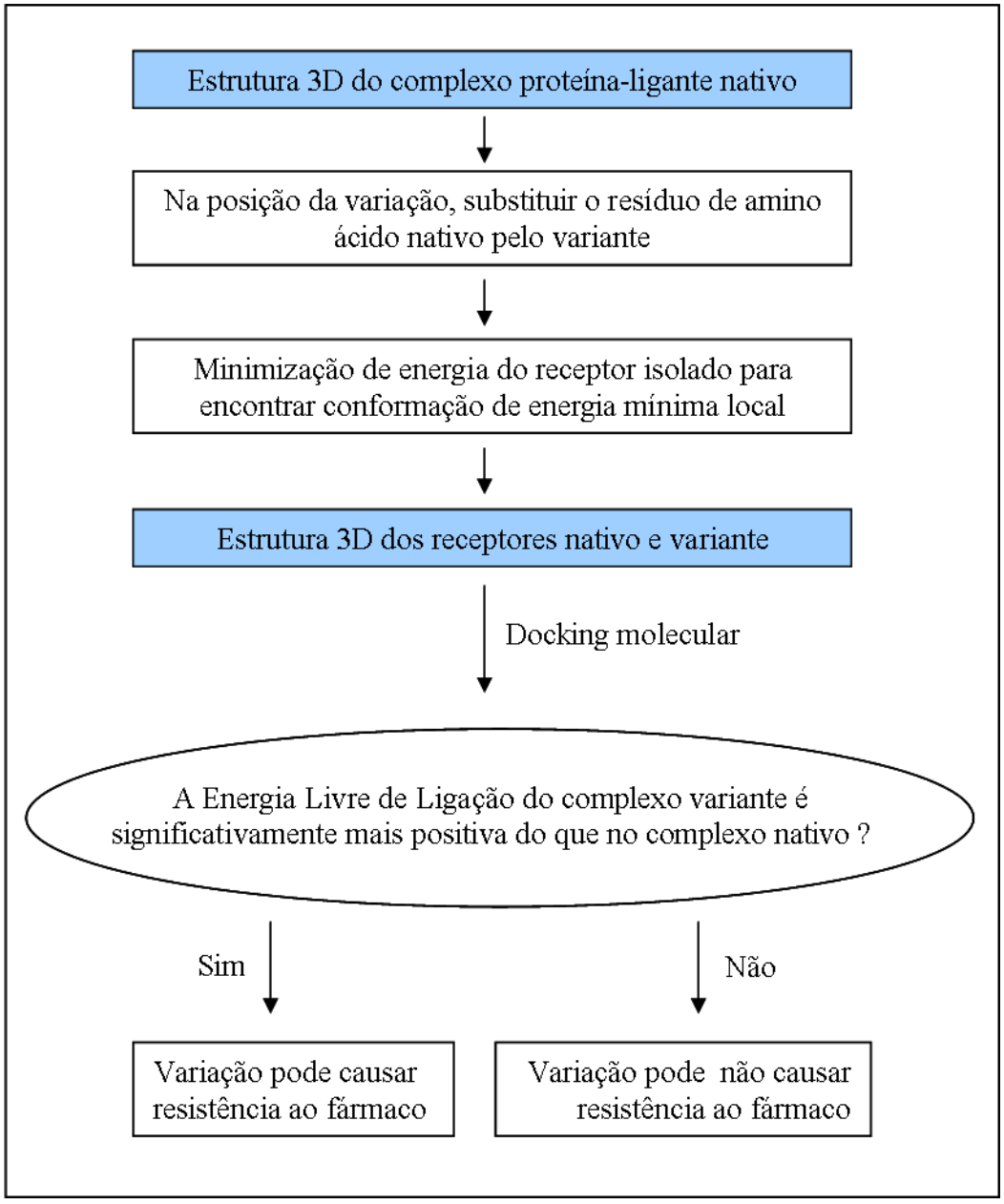


Figura 6. Fluxograma mostrando protocolo de uma abordagem computacional utilizada neste trabalho para determinar o impacto causado por substituições de resíduos de aminoácidos em complexos proteína/ligante.

Referências

. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem. 1998;19(14):1639-1662.

. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. J Comput Chem. 2007;28(6):1145-1152.

. Nowell P, Hungerford D. A minute chromosome in human chronic granulocytic leukemia. Science. 1960;132:1497.

. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature. 1973;243(5405):290-293.

. Shepherd P, Suffolk R, Halsey J, Allan N. Analysis of molecular breakpoint and m-RNA transcripts in a prospective randomized trial of interferon in chronic myeloid leukaemia: no correlation with clinical features, cytogenetic response, duration of chronic phase, or survival. Br J Haematol. 1995;89(3):546-554.

. De Klein A, Hagemeijer A, Bartram CR, Houwen R, Hoefsloot L, Carbonell F, Chan L, Barnett M, Greaves M, Kleihauer E. bcr rearrangement and translocation of the c-abl oncogene in Philadelphia positive acute lymphoblastic leukemia. Blood. 1986;68(6):1369-1375.

. Sicheri F, Kuriyan J. Structures of Src-family tyrosine kinases. Curr Opin Struct Biol. 1997;7(6):777-785.

. Pluk H, Dorey K, Superti-Furga G. Autoinhibition of c-Abl. Cell. 2002;108(2):247-259.

. Sawyers CL. Disabling Abl-perspectives on Abl kinase regulation and cancer therapeutics. Cancer Cell. 2002a;1(1):13-15.

. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J. Structural basis for the autoinhibition of c-Abl tyrosine kinase. Cell. 2003;112(6):859-871.

. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N Engl J Med.

2001a;344(14):1038-1042.

. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001b;344(14):1031-1037.

. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science. 2000;289(5486):1938-1942.

. Kantarjian H, Sawyers C, Hochhaus A, Guilhot F, Schiffer C, Gambacorti-Passerini C, Niederwieser D, Resta D, Capdeville R, Zoellner U, Talpaz M, Druker B, Goldman J, O'Brien SG, Russell N, Fischer T, Ottmann O, Cony-Makhoul P, Facon T, Stone R, Miller C, Tallman M, Brown R, Schuster M, Loughran T, Gratwohl A, Mandelli F, Saglio G, Lazzarino M, Russo D, Baccarani M, Morra E. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. N Engl J Med. 2002;346(9):645-652.

. Sawyers CL, Hochhaus A, Feldman E, Goldman JM, Miller CB, Ottmann OG, Schiffer CA, Talpaz M, Guilhot F, Deininger MW, Fischer T, O'Brien SG, Stone RM, Gambacorti-Passerini CB, Russell NH, Reiffers JJ, Shea TC, Chapuis B, Coutre S, Tura S, Morra E, Larson RA, Saven A, Peschel C, Gratwohl A, Mandelli F, Ben-Am M, Gathmann I, Capdeville R, Paquette RL, Druker BJ. Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. Blood. 2002;99(10):3530-3539.

. Shah NP, Nicoll JM, Nagar B, Gorre ME, Paquette RL, Kuriyan J, Sawyers CL. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. Cancer Cell. 2002;2(2):117-125.

. von Bubnoff N, Peschel C, Duyster J. Resistance of Philadelphia-chromosome positive leukemia towards the kinase inhibitor imatinib (STI571, Glivec): a targeted oncoprotein strikes back. Leukemia. 2003;17(5):829-838.

. Weisberg E, Manley PW, Cowan-Jacob SW, Hochhaus A, Griffin JD. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. Nat Rev Cancer. 2007;7(5):345-356.

. Roumiantsev S, Shah NP, Gorre ME, Nicoll J, Brasher BB, Sawyers CL, Van Etten RA. Clinical resistance to the kinase inhibitor STI-571 in chronic myeloid leukemia by mutation of Tyr-253 in the

Abl kinase domain P-loop. Proc Natl A0cad Sci U S A. 2002;99(16):10700-10705.

. Cowan-Jacob SW, Fendrich G, Floersheimer A, Furet P, Liebetanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D, Manley PW. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. Acta Crystallogr D Biol Crystallogr. 2007;63(1):80-93.

. Corbin AS, Buchdunger E, Pascal F, Druker BJ. Analysis of the structural basis of specificity of inhibition of the Abl kinase by STI571. J Biol Chem. 2002;277(35):32214-32219.

. Roche-Lestienne C, Soenen-Cornu V, Grardel-Duflos N, Laï JL, Philippe N, Facon T, Fenaux P, Preudhomme C. Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. Blood. 2002;100(3):1014-1018.

. Branford S, Rudzki Z, Walsh S, Parkinson I, Grigg A, Szer J, Taylor K, Herrmann R, Seymour JF, Arthur C, Joske D, Lynch K, Hughes T. Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. Blood. 2003;102(1):276-283.

. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. Tetrahedron. 1980;36(22):3219-3228.

. Pricl S, Fermeglia M, Ferrone M, Tamborini E. T315I-mutated Bcr-Abl in chronic myeloid leukemia and imatinib: insights from a computational study. Mol Cancer Ther. 2005;4(8):1167-1174.

. Lee TS, Potts SJ, Kantarjian H, Cortes J, Giles F, Albitar M. Molecular basis explanation for imatinib resistance of BCR-ABL due to T315I and P-loop mutations from molecular dynamics simulations. Cancer. 2008;112(8):1744-1753.

5.5 Análise funcional e estrutural do impacto causado por SNPs no gene *IGF1R* utilizando métodos de Bioinformática e Quimioinformática

Nesta etapa, foi realizado um estudo com o objetivo de avaliar o impacto causado por SNPs no gene que codifica o receptor do fator de crescimento insulina-símile tipo 1 (*IGF1R*). A proteína receptora codificada por este gene é uma importante mediadora da proliferação e sobrevivência celular em humanos, e está implicada no desenvolvimento de várias doenças em pacientes que apresentam disfunções no gene *IGF1R*, dentre elas o câncer de mama e de próstata. Vários SNPs presentes neste gene têm sido associados a doenças em humanos. No entanto, devido ao grande número de SNPs neste gene, é necessário diferenciar SNPs funcionais daqueles não-funcionais, podendo estes ser usados como marcadores diagnósticos e prognósticos do câncer.

Neste estudo, foram analisados todos os SNPs conhecidos no gene *IGF1R*, e o impacto funcional e estrutural destes foi investigado através da utilização de várias ferramentas computacionais, dentre elas os algoritmos SIFT e PolyPhen. Vários SNPs analisados, dentre eles seis nsSNPs identificados como deletérios tanto pelo SIFT quanto pelo PolyPhen, podem ter um efeito deletério nas células afetadas. Através da modelagem molecular de um destes nsSNPs (rs61740868) na estrutura da proteína IGF1R, seguida da minimização de energia, foi também mostrado que este nsSNP causa uma alteração desfavorável da energia conformacional da proteína, decorrente da substituição de um resíduo de arginina para uma cisteína na superfície da proteína. Este trabalho foi publicado na revista *Journal of Biomedicine and Biotechnology.*

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 715139, 8 pages
doi:10.1155/2010/715139

Research Article

A Comprehensive *In Silico* Analysis of the Functional and Structural Impact of SNPs in the *IGF1R* Gene

S. A. de Alencar[1,2] and Julio C. D. Lopes[2]

[1] *Departamento de Bioquímica e Imunologia, Bioinformática, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil*
[2] *Chemoinformatics Group, NEQUIM, Departamento de Química, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil*

Correspondence should be addressed to S. A. de Alencar, sergiodealencar@gmail.com and Julio C. D. Lopes, jlopes.ufmg@gmail.com

Received 1 February 2010; Accepted 28 April 2010

Academic Editor: Ravindra N. Chibbar

Insulin-like growth factor 1 receptor (IGF1R) acts as a critical mediator of cell proliferation and survival. Many single nucleotide polymorphisms (SNPs) found in the *IGF1R* gene have been associated with various diseases, including both breast and prostate cancer. The genetics of these diseases could be better understood by knowing the functions of these SNPs. In this study, we performed a comprehensive analysis of the functional and structural impact of all known SNPs in this gene using publicly available computational prediction tools. Out of a total of 2412 SNPs in *IGF1R* retrieved from dbSNP, we found 32 nsSNPs, 58 sSNPs, 83 mRNA 3′ UTR SNPs, and 2225 intronic SNPs. Among the nsSNPs, a total of six missense nsSNPs were found to be damaging by both a sequence homology-based tool (SIFT) and a structural homology-based method (PolyPhen), and one nonsense nsSNP was found. Further, we modeled mutant proteins and compared the total energy values with the native IGF1R protein, and showed that a mutation from arginine to cysteine at position 1216 (rs61740868) on the surface of the protein caused the greatest impact on stability. Also, the FASTSNP tool suggested that 31 sSNPs and 3 intronic SNPs might affect splicing regulation. Based on our investigation, we report potential candidate SNPs for future studies on *IGF1R* mutations.

1. Introduction

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome is altered. SNPs make up about 90% of all human genetic variation, occurring every 100–300 bases along the 3-billion-base human genome, although their density vary between regions [1]. SNPs are found in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function; however, others could predispose people to disease or influence their response to a drug. Nonsynonymous SNPs (nsSNPs) that lead to an amino acid residue substitution in the protein product are of particular interest because they are responsible for nearly half of the known genetic variations related to human inherited disease [2]. Coding synonymous SNPs (sSNPs) and SNPs occurring outside gene promoter or coding regions may nevertheless still have consequences for gene expression, splicing, or transcription-factor binding [3, 4].

The identification of SNPs responsible for specific phenotypes appears to be a problem that is very difficult to solve, requiring multiple testing of hundreds or thousands of SNPs in candidate genes [5]. However, the question of how to choose the set of SNPs to be screened is critical to the success of association studies. A possible way to overcome this problem would be to prioritize SNPs according to their functional significance [6, 7] by using Bioinformatics prediction tools, which may help discriminate neutral SNPs from SNPs of likely functional importance and could also be useful to reveal the structural basis of disease mutations. Without any careful preselection of SNPs to be screened, a huge number of individuals might be required to detect association at a reasonable level of statistical significance [5].

Although wetlab-based approaches used to identify disease-associated SNPs from a large number of neutral SNPs remain crucial evidence for the functional role of SNPs [8], numerous disease associations published could not be confirmed by subsequent independent studies [6, 9]. Hence, independent evidence of functionality of SNPs obtained by using prediction tools could also serve as additional argument to discriminate true associations from false positives [5], as shown recently by the functional SNP analysis of the *BRCA1*, *ABL1*, *ERBB2*, *CFTR*, and *EGFR* genes [10–14].

Insulin-like growth factor 1 receptor (IGF1R) is a growth factor receptor tyrosine kinase that acts as a critical mediator of cell proliferation and survival. This receptor is implicated in several cancers, including both breast and prostate cancer [15, 16]. Evidence suggests that IGF1R signaling is required for survival and growth when prostate cancer cells progress to androgen independence [17], as increased levels of the receptor are expressed in the majority of primary and metastatic prostate cancer patient tumors [18]. There have also been studies showing associations of *IGF1R* polymorphisms in dementia and ischemic stroke [19, 20].

Although there are presently several articles describing the association of SNPs in the *IGF1R* gene with different types of diseases, computational analysis has not yet been undertaken on the functional consequences of SNPs in this gene. We applied different publicly available computational algorithms, namely, Sorting Intolerant From Tolerant (SIFT) [21], Polymorphism Phenotyping (PolyPhen) [22], and Function Analysis and selection tool for single nucleotide polymorphisms (FASTSNP) to identify likely deleterious SNPs which could affect protein function [23].

The SIFT algorithm predicts whether an amino acid substitution affects protein function based on sequence homology among related genes and domains over evolutionary time, and the physical-chemical properties of the amino acid residues [24–26]. Sequence conservation and the nature of the amino acid residues involved are also incorporated by PolyPhen, but it also values the location of the substitution within known structures and structural features of the protein available in the annotated database SwissProt [5, 27]. By accessing a variety of heterogeneous biological databases and analytical tools, FASTSNP is able to identify SNPs most likely to have functional effects, such as changes to the transcriptional level and pre-mRNA splicing [23].

SIFT and PolyPhen were approximately 80% successful in benchmarking studies employing amino acid substitutions assumed to have a major negative impact on the residual activity of the variant protein as the test set [22, 25, 27–29] and it has been estimated that the "false negative" and "false positive" error rates of SIFT is 31% and 20%, and 31% and 9% for PolyPhen [26]. FASTSNP was used to analyze 1569 SNPs from the SNP500 cancer database, and results showed that SNPs with a high predicted risk exhibited low allele frequencies for the minor alleles, which is consistent with the finding that a strong selective pressure exist for functional polymorphisms [23, 30].

As the majority of disease mutations affect protein stability [31, 32], we also proposed modeled protein structures for the mutant proteins and compared them with the native protein in order to evaluate stability changes.

2. Materials and Methods

2.1. Evaluation of the Functional Impact of Coding nsSNPs Using a Sequence Homology Tool (SIFT). SIFT takes a query sequence and uses multiple alignment information to predict tolerated and deleterious substitutions for every position of the query sequence (http://sift.jcvi.org) [21]. It is a multistep procedure that, given a protein sequence, (1) searches for similar sequences, (2) chooses closely related sequences that may share similar function, (3) obtains the multiple alignment of these chosen sequences, and (4) calculates normalized probabilities for all possible substitutions at each position from the alignment. Substitutions at each position with normalized probabilities less than a tolerance index of 0.05 are predicted to be intolerant or deleterious; those greater than or equal to 0.05 are predicted to be tolerated [24, 26].

The analysis was performed by allowing the algorithm to search for homologous sequences using the default settings (UniProt-TrEMBL 39.6 database, median conservation of sequences of 3.00, and allowance to remove sequences more than 90% identical to query sequence). The IGF1R FASTA amino acid sequence of the NCBI Protein accession id NP_000866.1 was used as the query sequence, and a total of 24 *IGF1R* nsSNPs filtered from the dbSNP database were analyzed.

2.2. Evaluation of the Functional Impact of Coding nsSNPs Using a Structural Homology-Based Method (PolyPhen). PolyPhen prediction is based on straightforward empirical rules which are applied to the sequence, phylogenetic and structural information characterizing the substitution [5]. The online input form available at http://coot.embl.de/PolyPhen was filled with the IGF1R amino acid sequence in FASTA format (NCBI Protein accession id NP_000866.1), and the position and substitution of each of the 24 nsSNPs analyzed by SIFT were also submitted for PolyPhen analysis. PolyPhen then searched for 3D protein structures, multiple alignments of homologous sequences and amino acid contact information in several protein structure databases, calculated position-specific independent counts (PSIC) scores for each of the two amino acid residues entered (the original residue and the nsSNP), and then computed the PSIC scores difference of the two residues. The higher a PSIC score difference, the higher functional impact a particular amino acid substitution is likely to have. A PSIC score difference of 1.5 and above is considered to be damaging. The query options were left with default values.

2.3. Functional Significance of SNPs in Regulatory Regions. The online tool FASTSNP [23] was used to determine the impact of the sSNPs, 3′ UTR regions SNPs and intronic

Table 1: List of nsSNPs that were analysed by SIFT and PolyPhen.

dbSNP ID	Alleles	AA change	Tolerance index	PSIC	Heterozygosity	Validation
rs70958401	C/T	Arg/Trp	0.18	**1.892**	0.039	
rs70958396	G/A	Ala/Thr	0.41	0.011	0.039	
rs61740877	G/A	Val/Ile	0.77	0.019	n/a	
rs61740868	C/T	Arg/Cys	**0.00**	**2.609**	n/a	1
rs61731172	G/A	Arg/Gln	0.74	0.137	n/a	
rs56248469	G/A	Arg/His	0.57	0.613	n/a	
rs45611935	A/G	Asn/Ser	0.77	0.387	n/a	
rs45597432	T/C	Ile/Thr	0.96	0.079	n/a	
rs45578132	T/C	Val/Ala	**0.00**	**2.027**	n/a	
rs45553041	G/A	Arg/His	**0.00**	**2.196**	0.012	
rs45526336	G/A	Glu/Lys	**0.00**	1.470	n/a	
rs45524940	A/G	Thr/Ala	**0.01**	**2.296**	n/a	
rs45522834	C/T	Thr/Ile	0.29	1.220	n/a	
rs45512296	G/A	Arg/His	**0.01**	**2.128**	n/a	
rs45504297	T/C	Leu/Pro	**0.00**	**2.372**	n/a	
rs45493995	G/T	Ser/Ile	0.30	0.400	n/a	
rs45475702	G/A	Val/Ile	0.55	0.296	n/a	
rs45451896	G/T	Arg/Leu	0.25	0.305	n/a	
rs45445894	G/A	Val/Met	**0.03**	0.947	0.011	
rs35224135	G/A	Ala/Thr	0.31	1.026	0.005	2
rs34516635	G/A	Arg/His	1.00	1.339	0.005	1; 2
rs34102392	G/A	Ala/Thr	0.17	1.097	n/a	
rs34061581	A/G	His/Arg	0.25	1.346	0.005	2
rs33958176	G/A	Arg/Gln	0.59	**1.503**	n/a	1; 2

Prediction scores found to be functionally significant by SIFT and PolyPhen are shown in bold.Validation Status Description: (1) validated by multiple, independent submissions to the refSNP cluster; (2) validated by frequency or genotype data: minor alleles observed in at least two chromosomes.

SNPs on the regulation of the *IGF1R* gene. The FASTSNP server (http://FASTSNP.ibms.sinica.edu.tw) follows the decision tree principle with external Web service access to TFSearch, which predicts whether a non-coding SNP alters the transcription factor binding site of a gene. The score is given on the basis of levels of risk with a ranking of 0, 1, 2, 3, 4, or 5. This signifies the levels of no, very low, low, medium, high, and very high effect, respectively.

2.4. Modeling of nsSNPs on Protein Structures and Calculation of their RMSD Difference. Structural analysis was performed in order to evaluate and compare the stability of native and mutant structures. Information about mapping the nsSNPs in the protein structure was obtained from dbSNP [33]. The highest resolution (2.00 Å) native structure of the IGF1R protein available in the Protein Data Bank (PDB) [34] has an id of 2oj9 [35]. The positions of the studied nsSNPs mutations on PDBid 2oj9 were confirmed by pairwise alignment between the FASTA amino acid sequence of the IGF1R protein obtained from the NCBI (NP_000866.1) and the 2oj9 FASTA amino acid sequence, using the Sequence Manipulation Suite [36]. The amino acid residue substitutions were performed using the Swiss-Pdb Viewer [37], followed by energy minimization of the modeled 3D structures using the GROMACS software version 4.0 [38]. The algorithms used for energy minimization were the steepest descent (1000 steps), followed by conjugate gradient (1500 steps) alternating with the steepest descent every 100 steps. The comparison between the resulting native and modeled structures was made by the calculation of the potential energy and RMSD values.

3. Results and Discussion

3.1. SNP Dataset. Polymorphism data of the *IGF1R* gene investigated in this paper was retrieved from the dbSNP database [33]. It contained a total of 2412 SNPs, out of which 32 (1.3%) were nsSNPs, 58 (2.4%) were sSNPs, 83 (3.4%) occurred in the mRNA 3′ UTR, and 2225 (92.2%) occurred in intronic regions. SNPs in the 5′ UTR region were not found. It can be seen from the distribution in Figure 1 that the vast majority of SNPs occur in the intronic region, and that there are more 3′ UTR region SNPs than nsSNPs or sSNPs. We selected missense nsSNPs, sSNPs, 3′ UTR SNPs, and intronic SNPs for our investigation.

3.2. Deleterious nsSNPs by SIFT Program. Protein sequence with mutational position and amino acid residue variants associated to 24 missense nsSNPs were submitted as input to the SIFT server, and the results are shown in Table 1, along with the corresponding heterozygosity and validation status description for each SNP, when available from

Table 2: List of SNPs predicted to be functionally significant by FASTSNP.

dbSNP ID	Nucleotide change	Region	Level of risk	Possible functional effect	Heterozygosity	Validation
rs45437300	A/T	coding	Very High-Very High (5-5)	Nonsense	n/a	
rs55895813	A/G	intronic	Medium-High (3-4)	Splicing site	n/a	
rs36108138	A/C	intronic	Medium-High (3-4)	Splicing site	n/a	
rs45495500	C/T	intronic	Medium-High (3-4)	Splicing site	n/a	
rs34226328	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.006	2
rs35041862	C/G	coding	Low-Medium (2-3)	Splicing regulation	0.017	2
rs55770488	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs35385418	A/G	coding	Low-Medium (2-3)	Splicing regulation	0.022	1; 2
rs45504194	A/G	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs45582234	G/T	coding	Low-Medium (2-3)	Splicing regulation	0.012	
rs17847210	G/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	1
rs56013396	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs35171849	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.011	1; 2
rs35812156	A/C	coding	Low-Medium (2-3)	Splicing regulation	n/a	1
rs55954954	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs45506098	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.013	1; 4; 5
rs45598332	G/T	coding	Low-Medium (2-3)	Splicing regulation	0.013	
rs45615734	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs45486504	C/G	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs3743262	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.255	1; 4; 5
rs45627636	A/G	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs45443393	A/G	coding	Low-Medium (2-3)	Splicing regulation	0.011	
rs45459793	A/G	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs56400113	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs35449468	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.006	1
rs17847208	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.005	1; 2
rs2229765	A/G	coding	Low-Medium (2-3)	Splicing regulation	0.458	1; 2; 3; 4
rs28664854	A/G	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs35362396	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.005	

Table 2: Continued.

dbSNP ID	Nucleotide change	Region	Level of risk	Possible functional effect	Heterozygosity	Validation
rs45598038	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	1
rs34364279	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.006	
rs45468291	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs56020698	C/T	coding	Low-Medium (2-3)	Splicing regulation	n/a	
rs17847203	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.170	1; 2
rs45453791	C/T	coding	Low-Medium (2-3)	Splicing regulation	0.039	1

Validation Status Description: (1) Validated by multiple, independent submissions to the refSNP cluster; (2) Validated by frequency or genotype data: minor alleles observed in at least two chromosomes; (3) All alleles have been observed in at least two chromosomes apiece; (4) Genotyped by HapMap project; (5) SNP has been sequenced in 1000 Genome project.

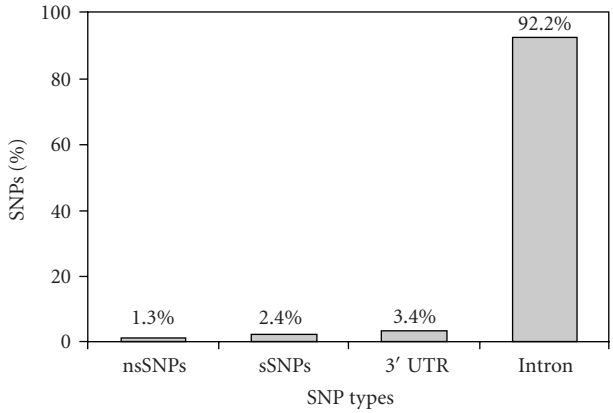


Figure 1: Distribution of *IGF1R* non-synonymous SNPs (nsSNPs), synonymous SNPs (sSNPs), 3′ UTR SNPs, and intronic SNPs.

dbSNP. According to the classification proposed by Ng and Henikoff [24] and Xi et al. [28], the lower the tolerance index, the higher the functional impact a particular amino acid residue substitution is likely to have and vice versa. Among the 24 nsSNPs analyzed, 8 nsSNPs were identified to be deleterious with a tolerance index score ≤0.05. Five nsSNPs (rs61740868, rs45578132, rs45553041, rs45526336, and rs45504297) showed a highly deleterious tolerance index score of 0.00. The remaining deleterious nsSNPs showed tolerance index scores of 0.01 (rs45524940 and rs45512296) and 0.03 (rs45445894). Four deleterious nsSNPs showed a nucleotide change from G/A, four a change from C/T, two a change from T/C, and one a change from A/G.

3.3. Damaged nsSNPs by PolyPhen Server. All the 24 protein sequences of missense nsSNPs submitted to SIFT were also submitted to the PolyPhen server. A PSIC score difference of 1.5 and above is considered to be damaging. Eight nsSNPs (rs70958401, rs61740868, rs45578132, rs45504297, rs45553041, rs45512296, rs45524940, and rs33958176) were considered to be damaging and exhibited a range of PSIC score difference between 1.503 and 2.609 (Table 1). Out of these damaging nsSNPs, two changed from positively charged amino acid in the native protein to hydrophobic amino acid in the mutant type, two from aliphatic nonpolar amino acid to non-polar amino acid, two from positively charged amino acid to aromatic positively charged amino acid, one from polar amino acid to non-polar amino acid, and one from positively charged to polar amino acid, respectively. It can be seen from Table 1 that there was significant correlation between the results obtained from the evolutionary-based approach SIFT and the structural-based approach PolyPhen for six nsSNPs predicted to be damaging by PolyPhen, suggesting that these nsSNPs may disrupt both the protein function and structure. The most damaging nsSNP (rs61740868) showed a PSIC score of 2.609, due to a mutation from arginine to cysteine.

3.4. SNPs in Regulatory Regions. According to FASTSNP, out of 58 sSNPs in the *IGF1R* gene, 31 sSNPs were predicted to be damaging with a risk ranking of 2-3, and a possible functional effect on splicing regulation (Table 2). Among these, the A/G polymorphism (rs2229765) has been shown experimentally to affect the susceptibility to ischemic stroke in Chinese population [19] to be associated with higher plasma concentrations of circulating IGF1R and premature pubarche [39, 40] and adult height variation in the human population [41]. Out of 2225 SNPs which occur in the intronic region of the *IGF1R* gene, 3 SNPs (rs55895813, rs36108138 and rs45495500) were predicted to affect the splicing site (3-4 risk) (Table 2).

It can be seen from Table 2 that a coding nonsense SNP (rs45437300) due to a nucleotide change from A to T was detected and showed a very high (5-5) level of risk, as it can truncate and even inactivate the IGF1R protein, causing disease as a result.

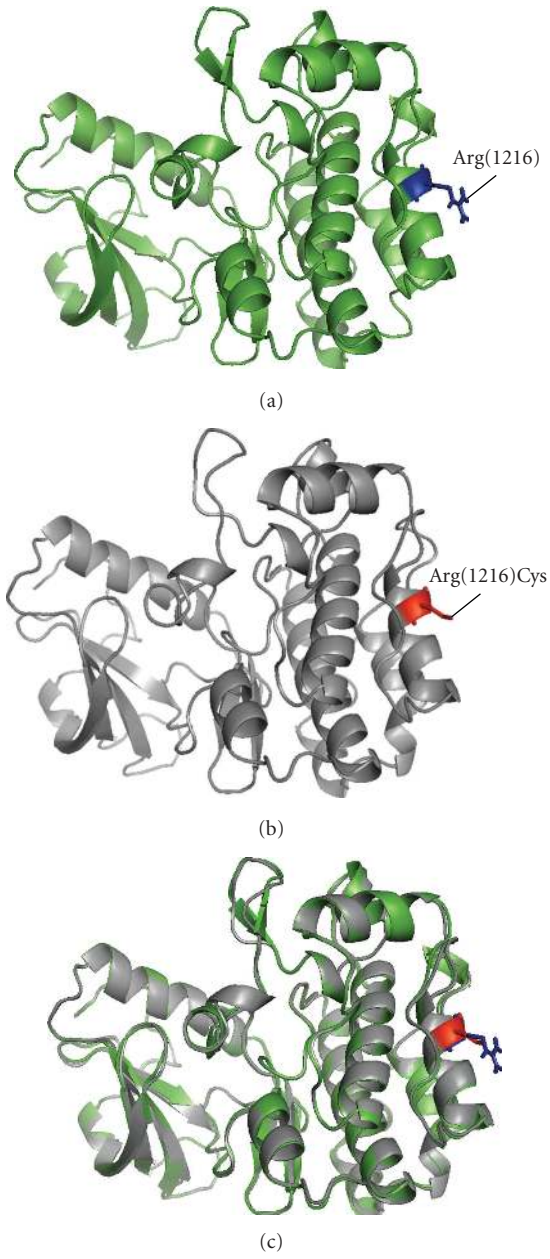


Figure 2: (a) Native structure (2jo9) showing arginine at position 1216. (b) Mutant modeled structure (2jo9 R1216C) showing cysteine residue at position 1216. (c) Superimposed structure of native structure (2jo9) (green) with mutant modeled structure (2jo9 R1216C) (gray).

3.5. Structural Analysis of Mutant Structures. Out of eight nsSNPs predicted to be deleterious by SIFT or PolyPhen, four (rs61740868, rs45526336, rs45512296, and rs45504297) were mapped to the PDB ID 2oj9 native structure. The amino acid residue substitutions were performed by Swiss-Pdb Viewer independently to get four mutant modeled structures (2oj9 R1216C, 2oj9 E1253K, 2oj9 R1216H, and 2oj9 L1211P, respectively). Then, energy minimizations were performed by GROMACS for the native structure (2oj9) and the mutant modeled structures.

Table 3: RMSD and total energy of native structure (2jo9) and mutant modeled structures.

dbSNP ID	Amino Acid change	RMSD between native and mutant structures	Total energy after minimization (KJ/mol)
rs61740868	Arg1216Cys	0,48	−13343.28
rs45526336	Glu1253Lys	0,38	−13887.05
rs45512296	Arg1216His	0,46	−13483.34
rs45504297	Leu1211Pro	0,22	−13782.33

Total energy of native structure (2jo9) after energy minimization: −13841.67.

The total energy for the native structure (2oj9) and the four mutant modeled structures 2oj9 R1216C, 2oj9 E1253K, 2oj9 R1216H, and 2oj9 L1211P was −13841.67, −13343.28, −13887.05, −13483.34, and −13782.33 KJ/mol, respectively (Table 3). Three out of four mutant modeled structures (2oj9 R1216C, 2oj9 R1216H, and 2oj9 L1211P) showed an increase in energy (less favorable change) in comparison with the native structure. This result correlates with the structural homology method (PolyPhen) results, which predicted all these three mutants to be deleterious (PSIC scores 2.609, 2.128, and 2.372, resp.) (Table 1). The mutant model 2oj9 R1216C showed the greatest increase in energy, which may be explained by the energetically unfavorable substitution of a positively charged arginine amino acid residue to a non-polar cysteine amino acid residue at the surface of the protein structure (Figure 2).

It can be seen from Table 3 that the RMSD values between the native structure (2oj9) and the mutant modeled structures are all similar, ranging from 0.22 Å to 0.48 Å. Because these values are low, we can suggest that these mutations do not cause a significant change in the mutant structures with respect to the native protein structure.

4. Conclusion

In this paper, we investigated the functional and structural impact of SNPs in the *IGF1R* gene using computational prediction tools. Out of a total of 2412 SNPs in the *IGF1R* gene, 32 SNPs were found to be non-synonymous, 58 were synonymous, 83 occurred in the mRNA 3′ UTR, and 2225 were found in intronic regions. Out of 24 missense nsSNPs, eight were found to be deleterious by SIFT, and eight were found to be damaging by the PolyPhen tool. A total of six nsSNPs were found to be damaging by both SIFT and PolyPhen tools. The structural analysis results showed that the amino acid residue substitutions which had the greatest impact on the stability of the IGF1R protein were mutations 2oj9 R1216C (rs61740868) and R1216H (rs45512296). Among the nsSNPs studied, a nonsense SNP (rs45437300) was found. Out of 58 sSNPs, 31 were predicted to affect splicing regulation by FASTSNP, including an sSNP (rs2229765) associated with several diseases. In the intronic region, 3 SNPs (rs55895813, rs36108138, and rs45495500) were predicted to affect splicing regulation. Based on our

results, we conclude that these SNPs should be considered important candidates in causing diseases related to *IGF1R* malfunction.

Acknowledgments

This work was supported by an FAPEMIG fellowship (S.A. de Alencar) and a CNPQ grant (J.C.D. Lopes).

References

[1] J.-E. Lee, J. H. Choi, J. H. Lee, and M. G. Lee, "Gene SNPs and mutations in clinical genetic testing: haplotype-based testing and analysis," *Mutation Research*, vol. 573, no. 1-2, pp. 195–204, 2005.

[2] M. Krawczak, E. V. Ball, I. Fenton et al., "Human gene mutation database—a biomedical information and research resource," *Human Mutation*, vol. 15, no. 1, pp. 45–51, 2000.

[3] L. Prokunina and M. E. Alarcón-Riquelme, "Regulatory SNPs in complex diseases: their identification and functional validation," *Expert Reviews in Molecular Medicine*, vol. 6, no. 10, 2004.

[4] P. D. Stenson, M. Mort, E. V. Ball, et al., "The human gene mutation database: 2008 update," *Genome Medicine*, vol. 1, no. 1, p. 13, 2009.

[5] V. Ramensky, P. Bork, and S. Sunyaev, "Human non-synonymous SNPs: server and survey," *Nucleic Acids Research*, vol. 30, no. 17, pp. 3894–3900, 2002.

[6] T. Emahazion, L. Feuk, M. Jobs et al., "SNP association studies in Alzheimer's disease highlight problems for complex disease analysis," *Trends in Genetics*, vol. 17, no. 7, pp. 407–413, 2001.

[7] N. J. Schork, D. Fallin, and J. S. Lanchbury, "Single nucleotide polymorphisms and the future of genetic epidemiology," *Clinical Genetics*, vol. 58, no. 4, pp. 250–264, 2000.

[8] C. G. P. Doss, C. Sudandiradoss, R. Rajasekaran et al., "Applications of computational algorithm tools to identify functional SNPs," *Functional and Integrative Genomics*, vol. 8, no. 4, pp. 309–316, 2008.

[9] N. J. Risch, "Searching for genetic determinants in the new millennium," *Nature*, vol. 405, no. 6788, pp. 847–856, 2000.

[10] R. Rajasekaran, C. Sudandiradoss, C. G. P. Doss, and R. Sethumadhavan, "Identification and in silico analysis of functional SNPs of the BRCA1 gene," *Genomics*, vol. 90, no. 4, pp. 447–452, 2007.

[11] C. G. P. Doss, C. Sudandiradoss, R. Rajasekaran, R. Purohit, K. Ramanathan, and R. Sethumadhavan, "Identification and structural comparison of deleterious mutations in nsSNPs of ABL1 gene in chronic myeloid leukemia: a bio-informatics study," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 607–612, 2008.

[12] R. Rajasekaran, C. G. P. Doss, C. Sudandiradoss, K. Ramanathan, R. Purohit, and R. Sethumadhavan, "Effect of deleterious nsSNP on the HER2 receptor based on stability and binding affinity with herceptin: a computational approach," *Comptes Rendus Biologies*, vol. 331, no. 6, pp. 409–417, 2008.

[13] C. G. P. Doss, R. Rajasekaran, C. Sudandiradoss, K. Ramanathan, R. Purohit, and R. Sethumadhavan, "A novel computational and structural analysis of nsSNPs in CFTR gene," *Genomic Medicine*, vol. 2, no. 1-2, pp. 23–32, 2008.

[14] R. Rajasekaran and R. Sethumadhavan, "In Silico identification of significant detrimental missense mutations of EGFR and their effect with 4-anilinoquinazoline-based drugs," *Applied Biochemistry and Biotechnology*, vol. 160, no. 6, pp. 1723–1733, 2010.

[15] G. S. Warshamana-Greene, J. Litz, E. Buchdunger, C. García-Echeverría, F. Hofmann, and G. W. Krystal, "The insulin-like growth factor-I receptor kinase inhibitor, NVP-ADW742, sensitizes small cell lung cancer cell lines to the effects of chemotherapy," *Clinical Cancer Research*, vol. 11, no. 4, pp. 1563–1571, 2005.

[16] H. E. Jones, L. Goddard, J. M. W. Gee et al., "Insulin-like growth factor-I receptor signalling and acquired resistance to gefitinib (ZD1839; Iresa) in human breast and prostate cancer cells," *Endocrine-Related Cancer*, vol. 11, no. 4, pp. 793–814, 2004.

[17] S. L. Krueckl, R. A. Sikes, N. M. Edlund et al., "Increased insulin-like growth factor I receptor expression and signaling are components of androgen-independent progression in a lineage-derived prostate cancer progression model," *Cancer Research*, vol. 64, no. 23, pp. 8620–8629, 2004.

[18] G. O. Hellawell, G. D. H. Turner, D. R. Davies, R. Poulsom, S. F. Brewster, and V. M. Macaulay, "Expression of the type 1 insulin-like growth factor receptor is up-regulated in primary prostate cancer and commonly persists in metastatic disease," *Cancer Research*, vol. 62, no. 10, pp. 2942–2950, 2002.

[19] J. Cheng, J. Liu, X. Li, et al., "Insulin-like growth factor-1 receptor polymorphism and ischemic stroke: a case-control study in Chinese population," *Acta Neurologica Scandinavica*, vol. 118, no. 5, pp. 333–338, 2008.

[20] J. Garcia, A. Ahmadi, A. Wonnacott, et al., "Association of insulin-like growth factor-1 receptor polymorphism in dementia," *Dementia and Geriatric Cognitive Disorders*, vol. 22, no. 5-6, pp. 439–444, 2006.

[21] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[22] S. Sunyaev, V. Ramensky, and P. Bork, "Towards a structural basis of human non-synonymous single nucleotide polymorphisms," *Trends in Genetics*, vol. 16, no. 5, pp. 198–200, 2000.

[23] H.-Y. Yuan, J.-J. Chiou, W.-H. Tseng et al., "FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization," *Nucleic Acids Research*, vol. 34, pp. W635–W641, 2006.

[24] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Research*, vol. 11, no. 5, pp. 863–874, 2001.

[25] P. C. Ng and S. Henikoff, "Accounting for human polymorphisms predicted to affect protein function," *Genome Research*, vol. 12, no. 3, pp. 436–446, 2002.

[26] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," *Annual Review of Genomics and Human Genetics*, vol. 7, pp. 61–80, 2006.

[27] S. Sunyaev, W. Lathe III, and P. Bork, "Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms," *Current Opinion in Structural Biology*, vol. 11, no. 1, pp. 125–130, 2001.

[28] T. Xi, I. M. Jones, and H. W. Mohrenweiser, "Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function," *Genomics*, vol. 83, no. 6, pp. 970–979, 2004.

[29] D. Chasman and R. M. Adams, "Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation," *Journal of Molecular Biology*, vol. 307, no. 2, pp. 683–706, 2001.

[30] M. Cargill, D. Altshuler, J. Ireland et al., "Characterization of single-nucleotide polymorphisms in coding regions of human genes," *Nature Genetics*, vol. 22, no. 3, pp. 231–238, 1999.
[31] Z. Wang and J. Moult, "SNPs, protein structure, and disease," *Human Mutation*, vol. 17, no. 4, pp. 263–270, 2001.
[32] N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik, "How protein stability and new functions trade off," *PLoS Computational Biology*, vol. 4, no. 2, Article ID e1000002, 2008.
[33] S. T. Sherry, M.-H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.
[34] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.
[35] U. Velaparthi, M. Wittman, P. Liu et al., "Discovery and initial SAR of 3-(1H-benzo[d]imidazol-2-yl)pyridin-2(1H)-ones as inhibitors of insulin-like growth factor 1-receptor (IGF-1R)," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 8, pp. 2317–2321, 2007.
[36] P. Stothard, "The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences," *BioTechniques*, vol. 28, no. 6, pp. 1102–1104, 2000.
[37] N. Guex, A. Diemand, and M. C. Peitsch, "Protein modelling for all," *Trends in Biochemical Sciences*, vol. 24, no. 9, pp. 364–366, 1999.
[38] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.
[39] M. Bonafè, M. Barbieri, F. Marchegiani et al., "Polymorphic variants of insulin-like growth factor I (IGF-I) receptor and phosphoinositide 3-kinase genes affect IGF-I plasma levels and human longevity: cues for an evolutionarily conserved mechanism of life span control," *Journal of Clinical Endocrinology and Metabolism*, vol. 88, no. 7, pp. 3299–3304, 2003.
[40] M. B. Roldan, C. White, and S. F. Witchel, "Association of the GAA1013 → GAG polymorphism of the insulin-like growth factor-1 receptor (IGF1R) gene with premature pubarche," *Fertility and Sterility*, vol. 88, no. 2, pp. 410–417, 2007.
[41] V. M. Chia, L. C. Sakoda, B. I. Graubard et al., "Risk of testicular germ cell tumors and polymorphisms in the insulin-like growth factor genes," *Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 3, pp. 721–726, 2008.

5.6 TargetSNPdb

Depois de avaliarmos a utilidade de várias ferramentas computacionais para o estudo do impacto de substituições de resíduos de aminoácidos na função protéica, foi construído um banco de dados, o TargetSNPdb, que contém resultados das análises feitas, juntamente com informações já existentes obtidas de outras fontes, tais como de doenças, vias metabólicas, alvos terapêuticos, fármacos, enzimas metabolizadoras de fármacos, e anotações de sequências protéicas, possibilitando a integração de diversas informações relevantes ao estudo do impacto de nsSNPs na função protéica.

Este trabalho, intitulado "TargetSNPdb: a database of preliminary analysis data of nsSNPs on drug target and disease associated genes" será submetido à revista *Nucleic Acids Research.* Uma descrição mais detalhada sobre este banco de dados será apresentada a seguir.

TargetSNPdb: a database of preliminary analysis data of the impact of nsSNPs on drug target and disease associated genes

S.A. de Alencar[1,2*], E.C. Santos[1,2] , A.M. José[2], J.C.D. Lopes[2]

[1]Departamento de Bioquímica e Imunologia, Bioinformática, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, Belo Horizonte – M.G., 31270-901, Brazil, Tel: +55 31 34095765, FAX: +55 31 34095700

[2]Chemoinformatics Group, NEQUIM, Departamento de Química, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, Belo Horizonte – M.G., 31270-901, Brazil, Tel: +55 31 34095765, FAX: +55 31 34095700

[*]Corresponding author at all stages of refereeing and publication

E-mail addresses:

SADA: sergiodealencar@gmail.com
ECS: edu.campos.santos@gmail.com
AMJ: andrellym@gmail.com
JCDL: jlopes.ufmg@gmail.com

Abstract

The presence of nsSNPs in genes encoding drug targets, or drug metabolizing enzymes has been increasingly associated with drug response and diseases. We have developed TargetSNPdb, a database server that contains computational predictions of the structural and functional impact of nsSNPs in protein coding genes, including drug target and drug metabolizing enzyme encoding genes. The analysis results obtained from several computational tools (such as SIFT, PolyPhen, AutoDock, and GROMACS) relevant to the study of the impact of amino acid residue substitutions were integrated to existent information records from the literature and genetic association databases, enabling the combination of results from a variety of different approaches to evaluate the impact of nsSNPs on protein function. Potential applications of TargetSNPdb include the priorization of nsSNPs for association and experimental studies. TargetSNPdb is available at http://nequim.qui.ufmg.br/targetsnp/.

Introduction

Single nucleotide polymorphisms (SNPs) constitute the most frequent type of sequence variation in humans, making up about 90% of all human genetic variation. Currently, there are almost 24 million human SNPs listed in publicly accessible databases, of which over 210,000 are located within protein coding sequences [dbSNP Build:131]. A fraction of these coding SNPs which alter the encoded amino acid sequence are known as non-synonymous SNPs (nsSNPs) [Sachidanandam et al., 2001].

The presence of nsSNPs in genes coding drug targets, or drug metabolizing enzymes, can cause structural variations in the active site of these proteins and, as a result, could affect drug interaction or destabilize the complex formed [Rajasekaran et al., 2008]. Also, changes in stability, which could be caused by a reduction in hydrophobic area, overpacking, backbone strain, or loss of electrostatic interactions, may affect a protein's folding rate and increase its susceptibility to proteolysis, resulting in reduced concentration of the native protein, and diseases [Wang et al., 2001; Yue et al., 2005; Karchin et al., 2005]. Therefore, nsSNPs are critical to understand the efficiency and toxicity of drugs.

The use of Bioinformatics and Chemoinformatics computational tools to analyze available sequence and structure data of proteins can contribute to increase prediction efficiency of the impact caused by nsSNPs on protein coding genes [Kapetanovic, 2008]. Several studies have shown that the impact caused by the substitution of amino acid residues on protein structures can be predicted by using both a sequence homology based tool (SIFT) and a structural homology based method (PolyPhen) [Rajasekaran et al., 2007; Doss et al., 2008; Doss et al., 2008b; Rajasekaran et al., 2009], and that molecular docking can be useful in predicting possible changes in ligand interaction energies between native and variant drug targets [Purohit et al., 2008].

Hence, the rapid accumulation of new data of human nsSNPs and drug target (and metabolizing enzyme) protein sequence and structure, together with computational analysis results, is opening the way to improve understanding of the relationships between genotype, drug response, and disease. However, at present, relevant nsSNP and protein target information are scattered across many databases, and the computational prediction of the impact of nsSNPs on drug targets is limited to a few receptors [Bigler et al., 2007; Liu et al., 2009], creating new challenges for linking genetic variation with drug response variation.

We propose a database to collect, analyze and integrate as much as possible of the molecular level data relevant to the mechanisms that link nsSNP records to drug related information.

TargetSNPdb is a Bioinformatics database that describes nsSNP records data, frequency information, nsSNP prediction of impact results, molecular docking and stability comparisons between native and mutant structures, association studies from the literature, and mapping of nsSNP positions in drug target and drug metabolizing enzyme structures.

Materials and Methods

Database setup

TargetSNPdb was implemented in MySQL, version 5.1.45 (http://www.mysql.com/), a freely available relational database management system (RDBMS), and its graphical CGI interface was programmed in PHP, version 5.2.8 (http://php.net), using the ADOdb, version 5.11 (http://adodb.sourceforge.net), a open source database abstraction library for PHP. The software DBDesigner, version 4.0.5.6 (http://www.fabforce.net/dbdesigner4) was used to model the data (Figure 1). The database is maintained on a DELL PowerEdge server using Ubuntu Linux, version 8.04.2 (http://www.ubuntu.com).

Contents of TargetSNPdb

nsSNP data

Information about human nsSNP records was obtained from dbSNP build 131 [dbSNP Build:131], a resource at the National Center of Biotechnology Information that catalogs SNPs [Sherry et al., 2001]. The following limits were used: Organism (*Homo sapiens*), Function Class (coding non-synonymous missense), and SNP Class (SNP). All redundant nsSNP records which have been merged to existent nsSNP records were removed. Population frequency data of nsSNP records was obtained from the International HapMap Project Biomart site [Thorisson et al., 2005] using the following parameters: Schema (rel22_NCBI_Build36), Database (HapMap_rel22), Dataset (All Populations), and filtering only nsSNPs and alleles with a frequency [>=] 0.01.

Prediction of the impact of nsSNPs on protein function

The SIFT algorithm predicts whether an amino acid substitution affects protein function based on sequence homology among related genes and domains over evolutionary time, and the physical-chemical properties of the amino acid residues [Ng and Henikoff, 2001; Ng and Henikoff, 2002; Ng and Henikoff, 2006]. SIFT takes a query sequence and uses multiple alignment information to predict

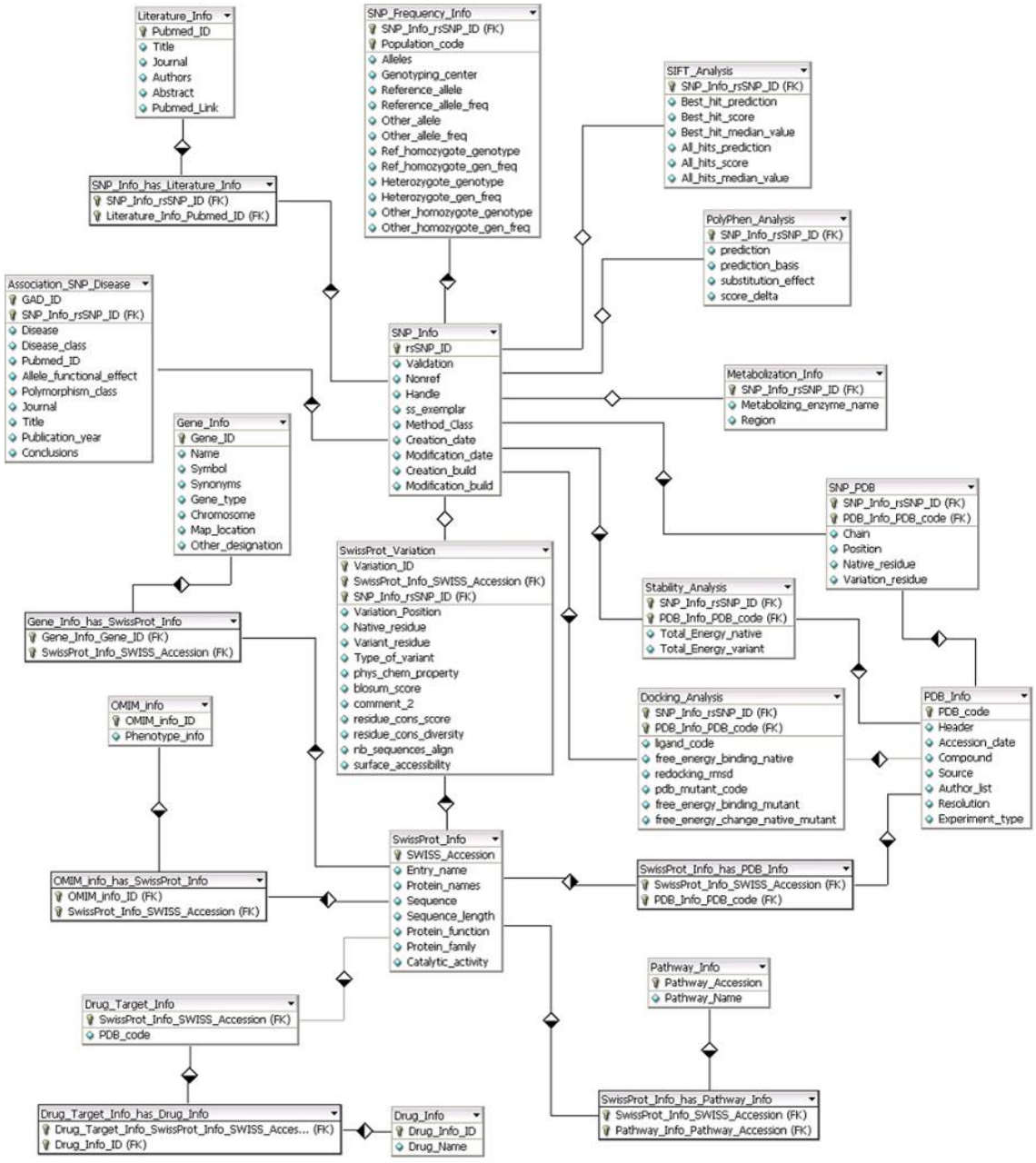


Figure 1. Data model schema showing the relational structure of TargetSNPdb, and all the tables and their relationships. A line with an empty diamond represents a one-to-one relationship while a half-filled diamond represents a one-to-many relationship. Primary keys are indicated with a key.

tolerated and deleterious substitutions for a position of interest in the query sequence (http://sift.jcvi.org) [Ng and Henikoff, 2003]. It is a multistep procedure that, given a protein sequence, (1) searches for similar sequences, (2) chooses closely related sequences that may share similar function, (3) obtains the multiple alignment of these chosen sequences, and (4) calculates normalized probabilities for a chosen substitution in a given position in the alignment. Substitutions at each position with normalized probabilities less than a tolerance index of 0.05 are predicted to be intolerant or deleterious; those greater than or equal to 0.05 are predicted to be tolerated [Ng and Henikoff, 2001; Ng and Henikoff, 2006].

Sequence conservation and the nature of the amino acid residues involved in a substitution are also incorporated by PolyPhen, but it also values the location of the substitution within known structures and structural features of the protein available in the annotated database SwissProt [Ramensky et al., 2002; Sunyaev et al., 2001]. Based on a query protein sequence, PolyPhen searches for related 3D protein structures, multiple alignments of homologous sequences and amino acid contact information in several protein structure databases, calculates position-specific independent counts (PSIC) scores the original residue and the nsSNP, and then computes the PSIC scores difference of the two residues. The higher a PSIC score difference, the higher functional impact a particular amino acid substitution is likely to have. A PSIC score difference of 1.5 and above is considered to be damaging. Publicly available pre-computed SIFT [http://sift.jcvi.org/www/SIFTing_databases.html] and PolyPhen [http://genetics.bwh.harvard.edu/pph/data/index.html] predictions of human nsSNPs from dbSNP were incorporated into TargetSNPdb.

Association of nsSNP records with diseases or literature records

Information about disease associated nsSNP records described in the Genetic Association Database, and nsSNPs records linked to PubMed entries were included in TargetSNPdb [Becker et al., 2004].

Protein data

Protein structural and sequence data along with annotations of function, pathway, family and disease association were obtained from the PDB, SwissProt, PANTHER and OMIM databases [Bermanet al., 2000; Gasteiger et al., 2001; Thomas et al., 2003; McKusick et al., 1998]. Additional sequence and structure information about the location of variant amino acid residues in the SwissProt

sequence was obtained from the SwissProt Variant Pages [Yip et al., 2004].

Drug related information

All information related to drugs (drug entries, drug targets, and drug metabolizing enzymes) was obtained from the DrugBank [Wishart et al., 2008], KEGG [Kanehisa et al., 2010], and TTD databases [Zhu et al., 2009].

Protein Side Chain Modeling

We retrieved from the Protein Data Bank all native three dimensional crystal structures available which were coded by genes which contained nsSNPs [Berman et al., 2000]. Information about positions of the nsSNPs on the PDB native structures was obtained from the coliSNP database [Kono et al., 2008]. Amino acid residue substitutions corresponding to nsSNPs in the native proteins were performed using the software SCWRL version 4, one of most accurate programs of protein side-chain modeling [Krivov et al., 2009].

Stability Analysis

In order to evaluate and compare the stability of native and modelled mutant structures generated with SCWRL4, energy minimization of the modelled 3D structures were done using the GROMACS software version 4.0 [Hess et al., 2008]. The algorithms used for energy minimization were steepest descent (6000 steps). The stability change value was calculated as the Potential Energy change (in Kcal/mol) between the native and variant protein structures using the GROMOS G53a6 force field [Oostenbrink et al., 2004].

Molecular Docking Analysis

All ligands crystallized in complex with drug targets which were coded by genes containing nsSNPs were selected for docking studies. Molecular docking calculations were carried out using the public software AutoDock 4.0 [Morris et al., 2009]. Before the docking process, grid maps representing the interaction energies between the various ligand atom types and the amino acid residue atoms in the receptor active site were calculated with the AutoGrid package of AutoDock. The center of the grid was defined as the center of the receptor active site, with points spaced at 0.375 Å.

Using the AutoDockTools (ADT) [Morris et al., 2009] package, polar hydrogen atoms were

added geometrically to the protein structures, and partial atomic charges were calculated using the Gasteiger-Marsili method. ADT was also used to assign the number of torsions and to add polar hydrogen atoms to each of the ligand structures.

Docking experiments were done using the Lamarckian genetic algorithm for the global search, and the Solis-Wets algorithm for the subsequent local optimization. The actual population comprised 150 individuals. We set the maximum number of energy evaluations accordingly to the number of degrees of freedom of the ligands studied (ranging from 1-6 million energy evaluations), the maximum number of generations to 270,000 and the number of runs to 100. A maximal mutation rate of 0.02, an elitism of 1, a crossover rate of 0.8 and a local search rate of 0.06 were used. Default values were used for all remaining parameters.

Results and Discussion

TargetSNPdb can be accessed through a web-based interface, which was constructed using php scripts to communicate with the MySQL database. The interface was designed to offer a variety of searching options: SNP RS Number, Gene symbol, SwissProt AC, PDB Code, Protein Name, Drug Target Name, Drug Name, Metabolizing Enzyme Name, Pathway Name, and OMIM Phenotype Info Figure 2A). Full list of drug target names, drug names, drug metabolizing enzyme names, pathway names, and OMIM Phenotyping Info names are also provided in the TargetSNPdb main web-page for facilitating the search of particular entries.

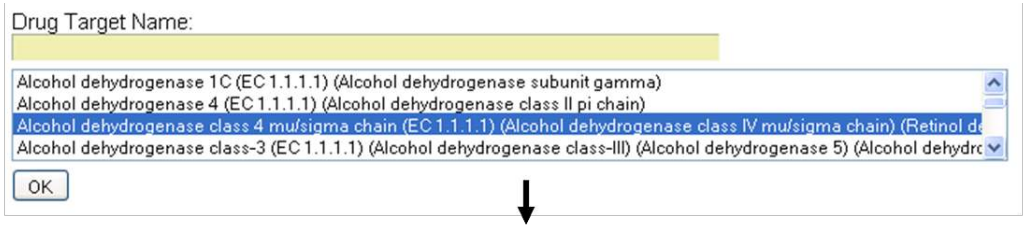
The search is case insensitive, and incomplete form of names (or characters) can be used in all search fields. For instance, the input of ‘acetyl’ finds entries with drug target name composed of characters ‘acetyl’ such as ‘Acetylcholine’ and ‘Acetyl-CoA carboxylase 2’. The wild character ‘%’ can also be used in a search to allow for more flexibility. For example, the input of ‘tyrosine%kinase’ in the protein name search field finds entries whose protein name contains both ‘tyrosine’ and ‘kinase’, such as ‘Tyrosine-protein kinase Lck’. The character ‘%’ here represents a string of arbitrary characters of any length.

The result of each search is displayed as a table, in which each column corresponds to information relevant to the search chosen by the user, such as a search by drug target name “dehydrogenase class 4 mu/sigma chain” (Figure 2B). In this example, all the drug target names that satisfy the search criteria are listed along with its SwissProt AC, SwissProt Variation ID, and SNP RS Number. More detailed information about the variation contained in the protein can be obtained by

A


SNP RS Number (e.g. 4531, 2020914):
OK
Gene symbol (e.g. CDH2, PCK1):
OK
SwissProt AC (e.g. P52630, O95477):
OK
PDB (e.g. 10GS, 1A22):
OK
Drug Target Name:
type above or select here
1-aminocyclopropane-1-carboxylate synthase-like protein 1 (ACC synthase-like protein)
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-3 (EC 3.1.4.11) (Phosphoinositide phospholipase C delt
10-formyltetrahydrofolate dehydrogenase (10-FTHFDH) (EC 1.5.1.6) (Aldehyde dehydrogenase family 1 member L1)
OK
Drug Metabolizing Enzyme Name:
type above or select here
adenosine_A2a_receptor
adrenergic_alpha_1A_receptor
adrenergic_beta-1_receptor
OK
Pathway Name:
type above or select here
2-arachidonoylglycerol biosynthesis
5-Hydroxytryptamine biosynthesis
5-Hydroxytryptamine degredation
OK
Drug:
Drug Name:
type above or select here
(+)-Irinotecan
(-)-3PPP, Maryland
(1'r,2's)-9-(2-Hydroxy-3'-Keto-Cyclopenten-1-Yl)Adenine
OK
Omim info:
Omim info:
type above and/or select here
2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency (MHBD deficiency)
3-alpha-hydroxyacyl-CoA dehydrogenase deficiency (HADH deficiency)
3-hydroxy-3-methylglutaryl-CoA lyase deficiency (HMG-CoA lyase deficiency)
OK

B



Drug Target Name(s)	SwissProt AC	SwissProt Variant ID	SNP RS Number
Alcohol dehydrogenase class 4 mu/sigma chain (EC 1.1.1.1) (Alcohol dehydrogenase class IV mu/sigma chain) (Retinol dehydrogenase) (Gastric alcohol dehydrogenase)	P40394[T]	VAR_024364	1573496

Protein Information

Protein Name	Alcohol dehydrogenase class 4 mu/sigma chain (EC 1.1.1.1) (Alcohol dehydrogenase class IV mu/sigma chain) (Retinol dehydrogenase) (Gastric alcohol dehydrogenase)
SwissProt AC	P40394
Drug Target?	Yes
Sequence Length	374
Sequence	MGTAGKVIKC KAAVLWEQKQ PFSIEEIEVA PPKTKEVRIK ILATGICRTD DHVIKGTMVS KFPVIVGHEA TGIVESIGEG VTTVKPGDKV IPLFLPQCRE CNACRNPDGN LCIRSDITGR GVLADGTTRF TCKGKPVHHF MNTSTFTEYT VVDESSVAKI DDAAPPEKVC LIGCGFSTGY GAAVKTGKVK PGSTCVVF GGVGLSVIMG CKSAGASRII GIDLNKDKFE KAMAVGATEC ISPKDSTKPI SEVLSEMTGN NVGYTFEVIG HLETMIDALA SCHMNYGTSV VVGVPPSAKM LTYDPMLLFT GRTWKGCVFG GLKSRDDVPK LVTEFLAKKF DLDQLITHVL PFKKISEGFE LLNSGQSIRT VLTF
Function	Could function in retinol oxidation for the synthesis of retinoic acid, a hormone important for cellular differentiation. Medium-chain (octanol) and aromatic (m-nitrobenzaldehyde) compounds are the best substrates. Ethanol is not a good substrate but at the high ethanol concentrations reached in the digest plays a role in the ethanol oxidation and contributes to the first pass ethanol metabolism.
Protein Family	Zinc-containing alcohol dehydrogenase family, Class-IV subfamily
Catalytic Activity	CATALYTIC ACTIVITY: An alcohol + NAD(+) = an aldehyde or ketone + NADH.
Pathway(s)	
Gene Symbol(s)	ADH7
SwissProt Variants(s)	VAR_024364
SNP RS Number(s)	1573496
PDB Code(s)	1AGN 1D1S 1D1T
Drug(s)	NADH Nicotinamide-Adenine Acetate Ion Cacodylate Ion

SwissProt Variant Information

SwissProt Variant ID	VAR_024364
Amino acid position of the variant	80
Native Residue	gly
Variant Residue	ala
Type of Variant	Polymorphism
Physico-Chemical Property	Change from glycine (G) to small size and hydrophobic (A)
BLOSUM score	0
Residue Conservation Score	0.882
Residue Conservation Diversity	76.17
Nb. of Sequences in Alignment	17
Surface Accessibility	The residue is on surface (SAS = 37.1804)
Comment	
SwissProt AC	P40394[T]

SNP Information

SNP RS Number	1573496
Validation	by-cluster,by-frequency,by-hapmap
Build	130
SwissProt AC(s)	P40394[T]
SIFT Analysis	
Prediction	DELETERIOUS
Score	0
Median Value	3.23
PolyPhen Analysis	
Prediction	possibly damaging
Prediction Basis	alignment
Substituition Effect	
Score	1.974
Potential Energy (KJ/mol)	
Native	-9.28267e+06
Variant	-9.64e+06
Molecular Docking Results (Kcal/mol)	
Ligand	NAD
Free Energy of Binding (Native)	-10.59
Free Energy of Binding (Variant)	-10.32
Complementary Information	
Alleles	C/G
Population Code	CEU
Reference Allele	C
Reference Allele Frequency	0,908
Other Allele	G
Other Allele Frequency	0,092
Reference Homozygote Genotype	C/C
Reference Homozygote Genotype Frequency	0,817
Heterozygote Genotype	C/G
Heterozygote Genotype Frequency	0,183
Other Homozygote Genotype	G/G
Other Homozygote Genotype Frequency	

Figure 2. (**A**) A screenshot montage of the TargetSNPdb interface showing several possible search options available for the user. (**B**) Overview of a result returned by querying TargetSNPdb using the drug target name search option (selecting Alcohol dehydrogenase class 4 mu/sigma chain). The blue arrows point to the information contained in each hyperlink shown in the intermediate results page.

clicking the corresponding SwissProt Variation ID or SNP RS Number. The result is displayed in another window, from which one may find information about the location of the variation in the protein structure, protein sequence, protein stability information, physical chemical properties, surface accessibility of the native and variant amino-acid residues, and the computational prediction of the impact of the variation.

In our laboratory, TargetSNPdb is currently being used to search for associations between drug response and diseases. The advantage of combining scores and analysis results produced by different methods, such as SIFT, PolyPhen, optimization, and molecular docking, is that each method uses different algorithms, so that when the results obtained from all these agree, predictions are more trustworthy. Also, if nsSNPs are associated with known drug responses or diseases, these combined predictions might explain the association. Future developments include the integration of a database containing experimentally determined drug affinity data, and updates for newly released dbSNP builds.

Availability

TargetSNPdb can be accessed freely at http://nequim.qui.ufmg.br/targetsnp/.

Acknowledgements

This work was supported by a FAPEMIG fellowship (S.A. de Alencar, and E.C. Santos), a CNPQ fellowship (A.M. José), and a CNPQ grant (J.C.D. Lopes).

References

. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National

Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID:131). Available from: http://www.ncbi.nlm.nih.gov/SNP/

. Sachidanandam R et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928-33.

. Rajasekaran R, Doss CGP, Sudandiradoss C, Ramanathan K, Purohit R, Sethumadhavan R. Effect of deleterious nsSNP on the HER2 receptor based on stability and binding affinity with herceptin: A computational approach. C R Biologies 2008;331:409-417.

. Wang Z, Moult J. SNPs, protein structure, and disease. Hum. Mutat. 2001;17:263-270.

. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005;353:459-473.

. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 2005;21(12):2814-2820.

. Kapetanovic IM. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact. 2008;171(2):165-176.

. Rajasekaran R, Sudandiradoss C, Doss CGP, Sethumadhavan R. Identification and in silico analysis of functional SNPs of the BRCA1 gene. Genomics 2007;90:447-452.

. Doss CGP, Sudandiradoss C, Rajasekaran R, Purohit R, Ramanathan K, Sethumadhavan R. Identification and structural comparison of deleterious mutations in nsSNPs of ABL1 gene in chronic myeloid leukemia: A bio-informatics study. Journal of Biomedical Informatics 2008;41:607-612.

. Doss CGP, Rajasekaran R, Sudandiradoss C, Ramanathan K, Purohit R, Sethumadhavan R. A novel computational and structural analysis of nsSNPs in CFTR gene. Genomic Med 2008b;2:23-32.

. Rajasekaran R, Sethumadhavan R. In Silico Identification of Significant Detrimental Missense Mutations of EGFR and Their Effect with 4-Anilinoquinazoline-Based Drugs. Appl Biochem Biotechnol. 2009;160(6):1723-1733.

. Purohit R, Rajasekaran R, Sudandiradoss C, George Priya Doss C, Ramanathan K, Rao S. Studies on flexibility and binding affinity of Asp25 of HIV-1 protease mutants. Int J Biol Macromol. 2008;42(4):386-391.

. Bigler J, Sibert JG, Poole EM, Carlson CS, Potter JD, Ulrich CM. Polymorphisms predicted to alter function in prostaglandin E2 synthase and prostaglandin E2 receptors. Pharmacogenet Genomics 2007;17(3):221-227.

. Liu YH, Li CG, Zhou SF. Prediction of deleterious functional effects of non-synonymous single nucleotide polymorphisms in human nuclear receptor genes using a bioinformatics approach. Drug Metab Lett. 2009;3(4):242-286.

. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308-311.

. Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. Genome Research 2005;15:1591-1593.

. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. Genome Res. 2001;11:863-874.

. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Research 2002;12(3):436-46.

. Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu. Rev. Genomics Hum. Genet. 2006;7:61–80.

. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-3814.

. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucl Acids Res. 2002;30:3894–3900.

. Sunyaev S, Lathe W 3rd, Bork P. Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms. Curr Opin Struct Biol. 2001;11(1):125-30.

. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431-432.

. Berman, HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235-242.

. Gasteiger E, Jung E, Bairoch A. SWISS-PROT: Connecting biological knowledge via a protein database Curr. Issues Mol. Biol. 2001;3:47-55.

. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K,

Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129-2141.

. McKusick VA. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. 12th ed. Baltimore: Johns Hopkins University Press; 1998.

. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. The Swiss-Prot Variant Page and the ModSNP Database: A Resource for Sequence and Structure information on Human Protein Variants. Hum. Mutat. 2004;23:464-470.

. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36:D901-906.

. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.; KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38,:D355-D360.

. Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 2009; 38:D787-D791.

. Kono H, Yuasa T, Nishiue S, Yura K. coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res. 2008;36:D409-413.

. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009;77(4):778-795.

. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput. 2008;4:435-447.

. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comp. Chem. 2004;25:1656-1676.

. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools: Automated docking with selective receptor flexibility. J. Comput. Chem. 2009;30(16):2785-2791.

6. CONCLUSÕES

- A precisão de vários métodos de modelagem molecular de cadeias laterais de resíduos de aminoácidos foi comparada, mostrando que o programa SCWRL 4 apresentou a melhor performance em geral.
- Para maximizar a precisão de cálculo utilizando o software AutoDock 4.0 em estudos de afinidade de ligação, foi demonstrado que o parâmetro ideal referente ao número de avaliações de energia depende do número de graus de liberdade do ligante estudado.
- Foi demonstrado que, através da modelagem molecular, seguida da minimização da estrutura mutante e do *docking* molecular utilizando o software AutoDock 4.0, é possível detectar substituições de resíduos de aminoácidos que afetam diretamente a afinidade entre um receptor em um ligante.
- Utilizando os métodos de Bioinformática e Quimionformática descritos neste trabalho, foi analisado o impacto funcional e estrutural de nsSNPs presentes no gene *IGF1R*. Vários SNPs analisados, dentre eles seis nsSNPs identificados como deletérios tanto pelo SIFT quanto pelo PolyPhen podem ter efeito nas células afetadas, e um deles (rs61740868) causou uma alteração desfavorável da energia conformacional da proteína, decorrente da substituição de um resíduo de arginina para uma cisteína na superfície da proteína.
- A importância da integração de diversas fontes de informação relevantes ao estudo do impacto de substituições de resíduos de aminoácidos na estrutura protéica foi demonstrada através da construção de um banco de dados relacional, o TargetSNPdb.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. Balasubramanian S, Xia Y, Freinkman E, Gerstein M. Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. Nucleic Acids Res. 2005;33(5):1710–1721.

2. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005;21(10):2185–2190.

3. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215-33.

4. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799-816.

5. Board PG, Pierce K, Coggan M. Expression of functional coagulation factor XIII in Escherichia coli. Thromb Haemost. 1990;63(2):235-40.

6. Borém A, Santos FR. Entendendo a biotecnologia. Viçosa; 2008.

7. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. J Mol Biol. 1997;267(5):1268–1282.

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

9. Canutescu AA, Shelenkov AA., Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003;12(9):2001–2014.

10. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999;22(3):231-238.

11. Celko J. Jow Celko's Data and Databases: Concepts in Practice. Morgan Kaufmann; 1999.

12. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol. 2001;307(2):683-706.

13. Chattopadhyay, K., Bhatia, S., Fiser, A., Almo, S.C., and Nathenson, S.G. Structural basis of inducible costimulator ligand costimulatory function: Determination of the cell surface oligomeric state

and functional mapping of the receptor binding site of the protein. J. Immunol. 2006;177: 3920–3929.

14. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5(4):823–826.

15. Chowbay B, Zhou S, Lee EJ. An interethnic comparison of polymorphisms of the genes encoding drug-metabolizing enzymes and drug transporters: experience in Singapore. Drug Metab Rev. 2005;37(2):327-378.

16. Cramer CJ. Essentials of Computational Chemistry: Theories and Models. Second Edition. Wiley; 2004

17. De Cristofaro R, Carotti A, Akhavan S, Palla R, Peyvandi F, Altomare C, Mannucci PM. The natural mutation by deletion of Lys9 in the thrombin A-chain affects the pKa value of catalytic residues, the overall enzyme's stability and conformational transitions linked to Na+ binding. FEBS J. 2006;273(1):159-69.

18. Dill KA, Fiebig KM, Chan HS. Cooperativity in protein-folding kinetics. Proc Natl Acad Sci U S A. 1993;90(5):1942-6.

19. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? Curr Opin Struct Biol. 2007;17(3):342-6.

20. Dipple KM, McCabe ER. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. Am J Hum Genet. 2000;66(6):1729-35.

21. DuBois P. MySQL. Quarta Edição. Addison-Wesley Professional; 2008.

22. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol. 1993;230(2):543–574.

23. Dunbrack RL Jr, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol. 1994;1(5):334–340.

24. Dunbrack RL, Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.

25. Eisenmenger F, Argos P, Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modelling. J Mol Biol. 1993;231(3):849–860.

26. Elles LM, Uhlenbeck OC. Mutation of the arginine finger in the active site of Escherichia coli DbpA abolishes ATPase and helicase activity and confers a dominant slow growth phenotype. Nucleic Acids Res. 2008;36(1):41-50.

27. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ. SNP

association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet. 2001;17(7):407-413.

28. Ewing TJA. Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput-Aided Molec. Design. 2001;15:411-428.

29. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol. 2002;315(4):771-786.

30. Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet. 2005;1(4):e56.

31. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15(1):R57-66.

32. Feyfant E, Sali A, Fiser A. Modeling mutations in protein structures. Protein Sci. 2007;16(9):2030-2041.

33. Fiser A. Protein structure modeling in the proteomics era. Expert Rev Proteomics. 2004;1(1):97–110.

34. Gibas C, Jambeck P. Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia. Editora Campus; 2002.

35. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985;28(7):849-857.

36. Goto J, Kataoka R, Muta H, Hirayama N. ASEDock-docking based on alpha spheres and excluded volumes. J Chem Inf Model. 2008;48(3):583-90.

37. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Introdução à Genética. Sexta Edição. Guanabara Koogan; 1998.

38. Hanemann CO, D'Urso D, Gabreëls-Festen AA, Müller HW. Mutation-dependent alteration in cellular distribution of peripheral myelin protein 22 in nerve biopsies from Charcot-Marie-Tooth type 1A. Brain. 2000;123(Pt5):1001-6.

39. Hardt M, Laine RA. Mutation of active site residues in the chitin-binding domain ChBDChiA1 from chitinase A1 of Bacillus circulans alters substrate specificity: use of a green fluorescent protein binding assay. Arch Biochem Biophys. 2004;426(2):286-97.

40. Hartman JL 4th, Garvik B, Hartwell L. Principles for the buffering of genetic variation. Science. 2001;291(5506):1001-4.

41. Hedgecoe AM. Terminology and the construction of scientific disciplines: The case of pharmacogenomics. Science, Technology & Human Values. 2003;28(4):513-537.

42. Holm L, Sander C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins. 1992;14(2):213–223.

43. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. J Comput Chem. 2007;28(6):1145-1152.

44. Hwang JK, Liao WF. Side-chain prediction by neural networks and simulated annealing optimization. Protein Eng. 1995;8(4):363–370.

45. Ingelman-Sundberg M. Pharmacogenetics: an opportunity for a safer and more efficient pharmacotherapy. J Intern Med. 2001;250(3):186-200.

46. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.

47. Israelachvili J. Intermolecular and Surface Forces. Segunda Edição. Academica Press; 1992.

48. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol. 2002;320(3):597–608.

49. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J. Med. Chem. 2003;46(4):499–511.

50. Jain T, Cerutti DS, McCammon JA. Configurational-bias sampling technique for predicting side-chain conformations in proteins. Protein Sci. 2006;15(9):2029–2039.

51. Janin J, Wodak S. Conformation of amino acid side-chains in proteins. J Mol Biol. 1978;125(3):357–386.

52. Johnson GC, Todd JA. Strategies in complex disease mapping. Curr Opin Genet Dev. 2000;10(3):330–334.

53. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 1997;267(3):727–48.

54. Jones R, Ruas M, Gregory F, Moulin S, Delia D, Manoukian S, Rowe J, Brookes S, Peters G. A CDKN2A mutation in familial melanoma that abrogates binding of p16INK4a to CDK4 but not CDK6. Cancer Res. 2007;67(19):9134-41.

55. José AM, Almeida V, de Alencar SA, Lopes JCD. NEQUIM Contact System - Protein-Ligand and Protein-Protein contact fingerprint generation and comparison. 2008. 4th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting), Salvador.

56. Kalow W, Meyer UA, Tyndale RF. Pharmacogenomics. Segunda Edição. Taylor & Francis Group; 2005.

57. Kalow W. Pharmacogenetics: Heredity and the response to drugs. W.B. Saunders; 1962.

58. Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform. 2007;8(5):333-46.

59. Kapetanovic IM. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact. 2008;171(2):165-176.

60. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding nonsynonymous SNPs based on multiple information sources. Bioinformatics. 2005;21(12):2814–2820.

61. Kariya Y, Tsubota Y, Hirosaki T, Mizushima H, Puzon-McLaughlin W, Takada Y, Miyazaki K. Differential regulation of cellular adhesion and migration by recombinant laminin-5 forms with partial deletion or mutation within the G3 domain of alpha3 chain. J Cell Biochem. 2003;88(3):506-20.

62. Kiewitz C, Tummler B. Similar profile of single nucleotide substitution types in bacteria and human genetic disease. Genome Letters. 2002;1:111-114.

63. Kirk BW, Feinsod M, Favis R, Kliman RM, Barany F. Single nucleotide polymorphism seeking long term association with complex disease. Nucleic Acids Res. 2002;30(15): 3295-3311.

64. Koehl P, Delarue M. The native sequence determines side-chain packing in a protein, but does optimal side-chain packing determine the native sequence? Pac Symp Biocomp. 1997;198-209.

65. Kolb P, Irwin JJ. Docking screens: right for the right reasons? Curr Top Med Chem. 2009;9(9):755-770.

66. Koukouritaki SB, Poch MT, Henderson MC, Siddens LK, Krueger SK, VanDyke JE, Williams DE, Pajewski NM, Wang T, Hines RN. Identification and functional analysis of common human flavin-containing monooxygenase 3 genetic variants. J Pharmacol Exp Ther. 2007;320(1):266-73.

67. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 1999;37:228–241.

68. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009;77(4):778-795.

69. Krumbholz M, Koehler K, Huebner A. Cellular localization of 17 natural mutant variants of ALADIN protein in triple A syndrome - shedding light on an unexpected splice mutation. Biochem Cell Biol. 2006;84(2):243-9.

70. Kwa LG, Wegmann D, Brügger B, Wieland FT, Wanner G, Braun P. Mutation of a single residue, beta-glutamate-20, alters protein-lipid interactions of light harvesting complex II. Mol Microbiol. 2008;67(1):63-77.

71. Ladurner, A.G. and Fersht, A.R. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. J. Mol. Biol. 1997;273: 330–337.

72. Lai E, Riley J, Purvis I, Roses A. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics. 1998;54(1):31-38.

73. Lander ES et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

74. Laskowski RA, Thornton JM. Understanding the molecular machinery of genetics through 3D structures. Nat Rev Genet. 2008;9(2):141-151.

75. Lasters I, Desmet J. The fuzzy-end elimination theorem: Correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem. Protein Eng. 1993;6(7):717–722.

76. Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature. 1991;352(6334):448–451.

77. Leite M. O DNA. Publifolha. Primeira edição; 2003.

78. Lewin B. Genes VII. Oxford University Press; 2000.

79. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA 3rd. Complete mutagenesis of the HIV-1 protease. Nature. 1989;340(6232):397–400.

80. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. J Mol Biol. 2001;307(1):429–445.

81. Matthews, B.W. Studies on protein stability with T4 lysozyme. Adv. Protein Chem. 1995;46: 249–278.

82. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. J. Computational Chemistry. 1998;19:1639-1662.

83. Morris GM, Goodsell DS, Huey R, Hart WE, Halliday S, Belew R, Olson AJ. AutoDock Version 3.0.5 User's Guide. http://autodock.scripps.edu/faqs-help/manual/autodock-3-user-s-guide. 2001.

84. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;156(1):297-304.

85. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-874.

86. Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu Rev Genomics Hum Genet. 2006;7:61–80.

87. Nussbaum RL, McInnes RR, Willard HF. Thompson & Thompson: Genética Médica. Sexta Edição. Editora Guanabara Koogan; 2002.

88. Ode H, Matsuyama S, Hata M, Neya S, Kakizawa J, Sugiura W, Hoshino T. Computational characterization of structural role of the non-active site mutation M36I of human immunodeficiency virus type 1 protease. J Mol Biol. 2007;370(3):598-607.

89. Ollila S, Sarantaus L, Kariola R, Chan P, Hampel H, Grabowski M, Macrae F, Kohonen-Corish M, Gerdes A-M, Peltoma¨ki P, Mangold E, de La Chapelle A, Greenblatt M, Nyström M. Pathogenicity of MSH2 missense mutations is typically associated with impaired repair capability of the mutated protein. Gastroenterology. 2006;131(5):1408–1417.

90. Ortiz MA, Light J, Maki RA, Assa-Munt N. Mutation analysis of the Pip interaction domain reveals critical residues for protein-protein interactions. Proc Natl Acad Sci U S A. 1999;96(6):2740-5.

91. Otzen, D.E. and Fersht, A.R. Analysis of protein-protein interactions by mutagenesis: Direct versus indirect effects. Protein Eng. 1999;12: 41–45.

92. Perrot P. A to Z of Thermodynamics. Oxford University Press; 1998.

93. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci. 2004;13(3):735–751.

94. Petrella RJ, Lazaridis T, Karplus M. Protein sidechain conformer prediction: A test of the energy function. Fold Des. 1998;3(5):353–377.

95. Phillips C. Online resources for SNP analysis: a review and route map. Mol Biotechnol. 2007;35(1):65-97.

96. Pidoux AL, Allshire RC. The role of heterochromatin in centromere function. Philos Trans R Soc Lond B Biol Sci. 2005;360(1455):569-79.

97. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol. 1987;193(4):775–791.

98. Pricl S, Fermeglia M, Ferrone M, Tamborini E. T315I-mutated Bcr-Abl in chronic myeloid leukemia and imatinib: insights from a computational study. Mol Cancer Ther. 2005;4(8):1167-1174.

99. Raevaara TE, Korhonen MK, Lohi H, Hampel H, Lynch E, Lonnqvist KE, Holinski-Feder E, Sutter

C, McKinnon W, Duraisamy S, Gerdes AM, Peltomaki P, Kohonen-Corish M, Mangold E, Macrae F, Greenblatt M, de la Chapelle A, Nyström M. Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. Gastroenterology. 2005;129(2):537–549.

100. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894–3900.

101. Rignall TR, Baker JO, McCarter SL, Adney WS, Vinzant TB, Decker SR, Himmel ME. Effect of single active-site cleft mutation on product specificity in a thermostable bacterial cellulase. Appl Biochem Biotechnol. 2002;98:383-94.

102. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000;405(6788):847-856.

103. Rogers DJ, Tanimoto TT. A computer program for classifying plants. Science. 1960;132(3434):1115-1118.

104. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinformatics. 2009;25(11):1431-2.

105. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234(3):779–815.

106. Schellack G. Farmacologia Uma Abordagem Didática. Editora Fundamento; 2005.

107. Schork NJ, Fallin D, Lanchbury JS. Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet. 2000;58(4):250-264.

108. Shirley BA, Stanssens P, Hahn U, Pace CN. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. Biochemistry. 1992;31(3):725-32.

109. Smith MB, Lamb ML, Tirado-Rives J, Jorgensen WL, Michejda CJ, Ruby SK, Smith RH Jr. Monte Carlo calculations on HIV-1 reverse transcriptase complexed with the non-nucleoside inhibitor 8-Cl TIBO: contribution of the L100I and Y181C variants to protein stability and biological activity. Protein Eng. 2000;13(6):413-421.

110. Song ES, Daily A, Fried MG, Juliano MA, Juliano L, Hersh LB. Mutation of active site residues of insulin-degrading enzyme alters allosteric interactions. J Biol Chem. 2005;280(18):17701-6.

111. Stevanin G, Hahn V, Lohmann E, Bouslam N, Gouttard M, Soumphonphakdy C, Welter ML, Ollagnon-Roman E, Lemainque A, Ruberg M, Brice A, Durr A. Mutation in the catalytic domain of protein kinase C gamma and extension of the phenotype associated with spinocerebellar ataxia type 14. Arch Neurol. 2004;61(8):1242-8.

112. Stouten PFW, Frömmel C, Nakamura H, Sander C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. Mol Simul. 1993;10(2):97-120.

113. Stryer L. Biochemistry. Quarta Edição. W. H. Freeman and Company; 1999.

114. Suarez-Kurtz G. Farmacogenômica: A genética dos medicamentos. Ciência Hoje. 2004; 208(35):20-27.

115. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 2000;16(5):198-200.

116. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 1999;12(5):387-94.

117. Sunyaev SR, Lathe WC 3rd, Ramensky VE, Bork P. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet. 2000;16(8):335-337.

118. Takamiya O, Seta M, Tanaka K, Ishida F. Human factor VII deficiency caused by S339C mutation located adjacent to the specificity pocket of the catalytic domain. Clin Lab Haematol. 2002;24(4):233-8.

119. Tang KE, Dill KA. Native protein fluctuations: the conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. J Biomol Struct Dyn. 1998;16(2):397-411.

120. Tanimoto TT. IBM Internal Report. 1957.

121. The Human Genome. Nature. 2001;409(6822):745-964.

122. The Human Genome. Science. 2001b;291(5507):1145-1434.

123. Tiede S, Cantz M, Spranger J, Braulke T. Missense mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTA) in a patient with mucolipidosis II induces changes in the size and cellular distribution of GNPTG. Hum Mutat. 2006;27(8):830-1.

124. Torkamani A, Schork NJ. Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics. 2007;90(1):49-58.

125. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn. 1991;8(6):1267–1289.

126. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE. Fine-scale structural variation of the human genome. Nat Genet. 2005;37(7):727-32.

127. Ung MU, Lu B, McCammon JA. E230Q mutation of the catalytic subunit of cAMP-dependent protein kinase affects local structure and the binding of peptide inhibitor. Biopolymers. 2006;81(6):428-39.

128. Venkatesan RN, Treuting PM, Fuller ED, Goldsby RE, Norwood TH, Gooley TA, Ladiges WC, Preston BD, Loeb LA. Mutation at the polymerase active site of mouse DNA polymerase delta increases genomic instability and accelerates tumorigenesis. Mol Cell Biol. 2007;27(21):7669-82.

129. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: methodologies and updates. J Med Chem. 2005;48(12):4111-4119.

130. Watson JD, Crick FHC. A Structure for Deoxyribose Nucleic Acid. Nature 1953;171:737-738.

131. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc. 1984;106(3):765–784.

132. Weinshilboum R. Inheritance and drug response. N Engl J Med. 2003;348(6):529-537.

133. Wolfsberg TG, McEntyre J, Schuler GD. Guide to the draft human genome. Nature. 2001;409(6822):824-6.

134. Wright AF. Nature Encyclopedia of the Human Genome. Volume 2:959–968. Nature Publishing Group; 2003.

135. Wright JD, Lim C. Mechanism of DNA-binding loss upon single-point mutation in p53. J Biosci. 2007;32(5):827-39.

136. Wu G, Fiser A, ter Kuile B, Sali A, Müller M. Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci. 1999;96(11):6285–6290.

137. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics. 2004;83(6):970–979.

138. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol. 2001;311(2):421–430.

139. Yamada Y, Banno Y, Yoshida H, Kikuchi R, Akao Y, Murate T, Nozawa Y. Catalytic inactivation of human phospholipase D2 by a naturally occurring Gly901Asp mutation. Arch Med Res. 2006;37(6):696-9.

140. Young D. Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems. First Edition. Wiley-Interscience; 2001.

141. van Wijk R, Rijksen G, Huizinga EG, Nieuwenhuis HK, van Solinge WW. HK Utrecht: missense mutation in the active site of human hexokinase associated with hexokinase deficiency and severe nonspherocytic hemolytic anemia. Blood. 2003;101(1):345-7.

8.1 Definição dos ângulos diedros χ1 e χ2 referentes às cadeias laterais dos resíduos de aminoácidos estudados

Cadeia Lateral	Eixo	Átomos usados para definir ângulo χ1
Arg	CA-CB	N-CA-CB-CG
Asn	CA-CB	N-CA-CB-CG
Asp	CA-CB	N-CA-CB-CG
Cys	CA-CB	N-CA-CB-SG
Gln	CA-CB	N-CA-CB-CG
Glu	CA-CB	N-CA-CB-CG
His	CA-CB	N-CA-CB-CG
Ile	CA-CB	N-CA-CB-CG1
Leu	CA-CB	N-CA-CB-CG
Lys	CA-CB	N-CA-CB-CG
Met	CA-CB	N-CA-CB-CG
Phe	CA-CB	N-CA-CB-CG
Pro	CA-CB	N-CA-CB-CG
Ser	CA-CB	N-CA-CB-OG
Thr	CA-CB	N-CA-CB-OG1
Trp	CA-CB	N-CA-CB-CG
Tyr	CA-CB	N-CA-CB-CG
Val	CA-CB	N-CA-CB-CG1

Cadeia Lateral	Eixo	Átomos usados para definir ângulo χ2
Arg	CB-CG	CA-CB-CG-CD
Asn	CB-CG	CA-CB-CG-OD1
Asp	CB-CG	CA-CB-CG-OD1
Gln	CB-CG	CA-CB-CG-CD
Glu	CB-CG	CA-CB-CG-CD
His	CB-CG	CA-CB-CG-ND1
Ile	CB-CG1	CA-CB-CG1-CD
Leu	CB-CG	CA-CB-CG-CD1
Lys	CB-CG	CA-CB-CG-CD
Met	CB-CG	CA-CB-CG-SD
Phe	CB-CG	CA-CB-CG-CD1
Pro	CB-CG	CA-CB-CG-CD
Trp	CB-CG	CA-CB-CG-CD1
Tyr	CB-CG	CA-CB-CG-CD1

8.2 Lista de estruturas obtidas do banco de dados PDB utilizadas no estudo de avaliação da precisão de vários métodos de modelagem molecular de cadeias laterais de resíduos de aminoácidos

Mutante Cristalizado (PDB ids)	Template (PDB ids)	Chain	Position	AA	Surface Accessibility
133l	1jsf	A	115	His	4
134l	1jsf	A	115	Glu	4
1a40	1ixh	A	197	Trp	9
1a4v	1b9o	A	45	Asn	0
1a6g	1a6m	A	122	Asn	5
1a6m	1a6g	A	122	Asp	5
1a7d	1a7e	A	103	Leu	9
1a7e	1a7d	A	103	Asn	9
1abe	8abp	A	107	Met	9
1abf	8abp	A	107	Met	9
1b0y	1cku	A	42	Gln	6
1b4t	1jcv	A	48	Cys	9
1b6q	1rop	A	31	Pro	6
1b7l	1jsf	A	32	Leu	9
1b7n	1jsf	A	35	Leu	7
1b7o	1jsf	A	37	Gln	3
1b8r	4cpv	A	102	Trp	9
1b9o	1a4v	A	45	Asp	0
1b9o	1hml	A	45	Asp	0
1bcx	1xnb	A	172	Cys	8
1bn8	2bsp	A	279	Arg	6
1bpq	1une	A	56	Met	4
1c5h	1xnb	A	35	Asp	7
1c5i	1xnb	A	35	Asp	7
1ceh	1une	A	99	Asn	9
1cj7	1jsf	A	11	Val	5
1cj9	1jsf	A	40	Val	9
1ckd	1jsf	A	43	Val	5
1clu	1rvd	A	12	Pro	6
1czk	1ofv	A	100	Asn	9
1czr	1ofv	A	90	Asn	7
1d3w	7fd1	A	15	Glu	5
1d6q	1jsf	A	102	Glu	4
1det	1i0v	A	25	Gln	5
1dmm	1opy	A	56	Phe	9
1e4c	1fua	P	71	Gln	9
1eq4	1jsf	A	7	Gln	5
1eq5	1jsf	A	102	Asn	4
1eqe	1jsf	A	120	Asn	1
1ert	1erv	A	73	Cys	4
1erv	1ert	A	73	Ser	4
1f5b	7fd1	A	2	His	5

1f5c	7fd1	A	25	His	9
1f98	3pyp	A	50	Val	8
1f9i	3pyp	A	42	Phe	9
1fdd	7fd1	A	15	Asn	5
1fla	5nul	A	57	Asp	3
1fua	1e4c	A	71	Ser	9
1fvx	5nul	A	57	Asn	3
1g02	1i0v	A	16	Ser	9
1g3o	7fd1	A	19	Glu	6
1g6b	7fd1	A	47	Ser	4
1gaz	1jsf	A	2	Ile	3
1gb0	1jsf	A	2	Leu	9
1gb2	1jsf	A	2	Met	9
1gb3	1jsf	A	2	Phe	2
1gb6	1jsf	A	74	Ile	3
1gb7	1jsf	A	74	Leu	3
1gb8	1jsf	A	74	Met	3
1gb9	1jsf	A	74	Phe	3
1gbw	1jsf	A	110	Ile	2
1gbx	1jsf	A	110	Leu	2
1gby	1jsf	A	110	Met	2
1gbz	1jsf	A	110	Phe	2
1gf8	1jsf	A	2	Ser	7
1gf9	1jsf	A	2	Tyr	7
1gfa	1jsf	A	2	Asp	9
1gfe	1jsf	A	2	Asn	3
1gfg	1jsf	A	2	Arg	9
1gfh	1jsf	A	74	Tyr	3
1gfj	1jsf	A	74	Asp	3
1gfk	1jsf	A	74	Asn	3
1gfr	1jsf	A	74	Arg	3
1gft	1jsf	A	110	Tyr	2
1gfu	1jsf	A	110	Asp	2
1gfv	1jsf	A	110	Asn	2
1hem	3lzt	A	91	Thr	9
1hen	1hep	A	40	Thr	9
1heo	3lzt	A	55	Val	9
1hep	1hen	A	40	Ser	9
1hep	1heq	A	55	Val	9
1heq	1hep	A	55	Ile	9
1her	3lzt	A	40	Ser	9
1hml	1b9o	A	45	Asn	0
1hnj	1ebl	A	233	Leu	8
1i0v	1bir	A	100	Phe	8
1i0v	1det	A	25	Lys	5
1i0v	1g02	A	16	Val	9
1i0v	1hyf	A	16	Val	9
1i0v	1hz1	A	16	Val	9
1i0v	1lra	A	58	Glu	8
1i0v	1rgk	A	46	Glu	7
1i0v	1rls	A	25	Lys	5

1i0v	1rn1	A	25	Lys	5
1i0v	2aae	A	40	His	5
1i0v	2hoh	A	9	Asn	4
1i0v	3hoh	A	93	Thr	4
1i0v	4bir	A	92	His	6
1i0v	5bir	A	92	His	6
1i0v	7rnt	A	45	Tyr	0
1icn	1ifc	A	106	Gln	8
1inu	1jsf	A	110	Arg	2
1ixg	1ixh	A	141	Asp	9
1ixh	1ixg	A	141	Thr	9
1ixh	1ixi	A	56	Asp	9
1ixh	1pbp	A	141	Thr	9
1ixh	1qui	A	137	Asp	7
1ixh	1quj	A	137	Asp	7
1ixh	1quk	A	137	Asp	7
1ixh	1qul	A	137	Asp	7
1ixi	1ixh	A	56	Asn	9
1jai	1ctq	A	12	Pro	3
1jcv	1b4t	A	48	His	9
1jka	1jsf	A	35	Asp	7
1jkc	1jsf	A	109	Phe	8
1jsf	133l	A	115	Arg	4
1jsf	134l	A	115	Arg	4
1jsf	1b5u	A	24	Ser	4
1jsf	1b7n	A	35	Glu	7
1jsf	1b7r	A	58	Gln	8
1jsf	1cj6	A	11	Thr	5
1jsf	1cj7	A	11	Thr	5
1jsf	1cj8	A	40	Thr	9
1jsf	1cj9	A	40	Thr	9
1jsf	1ckc	A	43	Thr	5
1jsf	1ckd	A	43	Thr	5
1jsf	1ckf	A	52	Thr	8
1jsf	1d6q	A	102	Asp	4
1jsf	1di3	A	50	Arg	3
1jsf	1eq4	A	7	Glu	3
1jsf	1eq5	A	102	Asp	4
1jsf	1eqe	A	120	Asp	1
1jsf	1gdx	A	21	Arg	5
1jsf	1ge0	A	38	Tyr	8
1jsf	1ge1	A	58	Gln	8
1jsf	1ge3	A	118	Asn	0
1jsf	1ge4	A	118	Asn	0
1jsf	1gfh	A	74	Val	3
1jsf	1gfj	A	74	Val	3
1jsf	1gfk	A	74	Val	3
1jsf	1gfr	A	74	Val	3
1jsf	1gft	A	110	Val	2
1jsf	1hnl	A	77	Cys	9
1jsf	1inu	A	110	Val	2

1jsf	1jka	A	35	Glu	7
1jsf	1jkb	A	35	Glu	7
1jsf	1jkc	A	109	Trp	8
1jsf	1jkd	A	109	Trp	8
1jsf	1laa	A	53	Asp	7
1jsf	1lhh	A	110	Val	2
1jsf	1lhi	A	71	Pro	4
1jsf	1lhj	A	103	Pro	1
1jsf	1lhk	A	91	Asp	4
1jsf	1loz	A	56	Ile	9
1jsf	1lyy	A	67	Asp	8
1jsf	1lz4	A	77	Cys	9
1jsf	1oua	A	56	Ile	9
1jsf	1oub	A	100	Val	9
1jsf	1ouc	A	110	Val	2
1jsf	1oud	A	121	Val	8
1jsf	1oue	A	125	Val	7
1jsf	1ouh	A	74	Val	3
1jsf	1oui	A	93	Val	9
1jsf	1ouj	A	99	Val	9
1jsf	1tay	A	63	Tyr	2
1jsf	1tby	A	63	Tyr	2
1jsf	1tcy	A	63	Tyr	2
1jsf	1tdy	A	63	Tyr	2
1jsf	1wqm	A	124	Tyr	8
1jsf	1wqn	A	20	Tyr	6
1jsf	1wqo	A	38	Tyr	8
1jsf	1wqp	A	45	Tyr	3
1jsf	1wqq	A	54	Tyr	8
1jsf	1wqr	A	63	Tyr	2
1jsf	1yam	A	106	Ile	9
1jsf	1yan	A	23	Ile	8
1jsf	1yao	A	56	Ile	9
1jsf	1yap	A	59	Ile	9
1jsf	1yaq	A	89	Ile	9
1jsf	207l	A	77	Cys	9
1jsf	2hea	A	106	Ile	9
1jsf	2heb	A	23	Ile	8
1jsf	2hec	A	56	Ile	9
1jsf	2hed	A	59	Ile	9
1jsf	2hee	A	59	Ile	9
1jsf	2hef	A	89	Ile	9
1jsf	2meb	A	56	Ile	9
1jsf	2med	A	59	Ile	9
1jsf	2mee	A	59	Ile	9
1jsf	2mef	A	59	Ile	9
1jsf	2meg	A	59	Ile	9
1jsf	2meh	A	59	Ile	9
1jsf	2mei	A	59	Ile	9
1kvc	2rn2	A	134	Asn	4
1kvw	1une	A	48	Gln	8

1kvy	1une	A	49	Glu	6
1kxw	3lzt	A	27	Asp	8
1kxy	3lzt	A	18	Asn	6
1l36	1l73	A	127	Asp	0
1l36	1l74	A	133	Leu	9
1l50	1l51	A	149	Cys	9
1l51	1l50	A	149	Ile	9
1l70	1l36	A	128	Glu	7
1l71	1l36	A	132	Asn	7
1laa	1jsf	A	53	Glu	7
1lav	2rn2	A	74	Leu	9
1law	2rn2	A	74	Ile	9
1lhh	1jsf	A	110	Pro	2
1lhk	1jsf	A	91	Pro	4
1lhl	1jsf	A	47	Pro	2
1loz	1jsf	A	56	Thr	9
1lsy	3lzt	A	52	Ser	6
1lyy	1jsf	A	67	His	8
1lzd	3lzt	A	62	Tyr	5
1lze	3lzt	A	62	Tyr	5
1lzg	3lzt	A	62	Phe	5
1mun	1muy	A	138	Asn	4
1muy	1mun	A	138	Asp	4
1ofv	1czh	A	58	Asn	0
1ofv	1czk	A	100	Asp	9
1ofv	1czo	A	58	Asn	0
1ofv	1czr	A	90	Asp	7
1ofv	1d03	A	58	Asn	0
1oua	1jsf	A	56	Thr	9
1pbp	1ixh	A	141	Asp	9
1qjd	1e39	A	365	His	9
1qke	3ebx	A	26	Asn	7
1quk	1ixh	A	137	Asn	7
1qul	1ixh	A	137	Thr	7
1ra9	1dhi	A	27	Asp	7
1ra9	1dra	A	27	Asp	7
1ra9	1drb	A	27	Asp	7
1ra9	2drc	A	22	Trp	8
1ra9	4dfr	A	154	Glu	7
1rbr	2rn2	A	62	Pro	2
1rbu	2rn2	A	95	Asn	2
1rdb	2rn2	A	48	Gln	8
1rgk	1i0v	A	46	Gln	7
1rls	1i0v	A	25	Gln	5
1rvd	1clu	A	12	Val	6
1tby	1jsf	A	63	Leu	2
1tcy	1jsf	A	63	Phe	2
1tdy	1jsf	A	63	Trp	2
1thv	1thw	A	46	Asn	2
1thw	1thv	A	46	Lys	2
1tys	1axw	A	146	Ser	6

1udb	1udc	A	131	Asn	7
1udb	2udp	A	131	Asn	7
1udc	1udb	A	131	Gln	7
1uid	3lzt	A	15	Phe	6
1uif	3lzt	A	15	Val	6
1une	1bpq	A	56	Lys	4
1une	1ceh	A	99	Asp	9
1une	1kvw	A	48	His	8
1une	1kvx	A	99	Asp	9
1une	1kvy	A	49	Asp	6
1vqb	1vqg	A	47	Ile	9
1vqc	1vqf	A	47	Phe	9
1vqd	1vqf	A	47	Leu	9
1vqe	1vqf	A	47	Met	9
1vqf	1vqc	A	47	Val	9
1vqf	1vqd	A	47	Val	9
1vqf	1vqe	A	47	Val	9
1vqf	1vqi	A	35	Ile	8
1vqf	1vqj	A	47	Val	9
1vqg	1vqa	A	35	Val	8
1vqg	1vqb	A	47	Leu	9
1vqg	1vqh	A	47	Leu	9
1vqh	1vqg	A	47	Met	9
1vqi	1vqf	A	35	Val	8
1vqj	1vqf	A	47	Ile	9
1wqm	1jsf	A	124	Phe	8
1wqn	1jsf	A	20	Phe	6
1wqo	1jsf	A	38	Phe	8
1wqp	1jsf	A	45	Phe	3
1wqq	1jsf	A	54	Phe	8
1wqr	1jsf	A	63	Phe	2
1xnb	1bcx	A	172	Glu	8
1xnb	1c5h	A	35	Asn	7
1xnb	1c5i	A	35	Asn	7
1xnb	2bvv	A	69	Tyr	9
1yam	1jsf	A	106	Val	9
1yan	1jsf	A	23	Val	8
1yao	1jsf	A	56	Val	9
1yap	1jsf	A	59	Val	9
1yaq	1jsf	A	89	Val	9
219l	237l	A	149	Val	9
2aae	1i0v	A	40	Lys	5
2acu	1ads	A	48	His	8
2bsp	1bn8	A	279	Lys	6
2bvv	1xnb	A	69	Phe	9
2meb	1jsf	A	56	Leu	9
2med	1jsf	A	59	Phe	9
2mee	1jsf	A	59	Leu	9
2mef	1jsf	A	59	Met	9
2meg	1jsf	A	59	Ser	9
2meh	1jsf	A	59	Thr	9

2mei	1jsf	A	59	Tyr	9
2mnr	1mdl	A	164	Lys	6
2ovo	1ppf	A	18	Met	0
2rn2	1kva	A	134	Asp	4
2rn2	1kvb	A	134	Asp	4
2rn2	1kvc	A	134	Asp	4
2rn2	1lav	A	74	Val	9
2rn2	1law	A	74	Val	9
2rn2	1rbr	A	62	His	2
2rn2	1rbs	A	62	His	2
2rn2	1rbt	A	95	Lys	2
2rn2	1rbu	A	95	Lys	2
2rn2	1rbv	A	95	Lys	2
2rn2	1rdb	A	48	Glu	8
3ebx	1qkd	A	26	His	7
3ebx	1qke	A	26	His	7
3lzt	1hem	A	91	Ser	9
3lzt	1heo	A	55	Ile	9
3lzt	1her	A	40	Thr	9
3lzt	1kxw	A	27	Asn	8
3lzt	1lsy	A	52	Asp	6
3lzt	1lzd	A	62	Trp	5
3lzt	1lze	A	62	Trp	5
3lzt	1lzg	A	62	Trp	5
3lzt	1uic	A	15	His	6
3lzt	1uid	A	15	His	6
3lzt	1uie	A	15	His	6
3lzt	1uif	A	15	His	6
3pyp	1f98	A	50	Thr	8
4bir	1i0v	A	92	Gln	6
4cpv	1b8r	A	102	Phe	9
4enl	1one	A	84	Ser	6
4nll	5nul	A	57	Asp	3
5abp	8abp	A	107	Met	9
5pti	1fan	A	45	Phe	8
5pti	8pti	A	35	Tyr	8
6paz	8paz	A	80	Ile	6
7fd1	1d3w	A	15	Asp	5
7fd1	1f5b	A	2	Phe	5
7fd1	1f5c	A	25	Phe	9
7fd1	1fd2	A	20	Cys	8
7fd1	1fdd	A	15	Asp	5
7fd1	1g3o	A	19	Val	6
7fd1	1g6b	A	47	Pro	4
7fd1	2fd2	A	24	Cys	9
7rnt	1i0v	A	45	Trp	0
821p	1ctq	A	12	Pro	3
8paz	4paz	A	80	Pro	6
8paz	5paz	A	80	Pro	6
8paz	6paz	A	80	Pro	6

8.3 Dados experimentais de afinidade de ligação (p*Ki*) obtidos da base de dados PDBBind

Código PDB	Proteína	Res (Å)	Ligante	p*ki*
1a69	PURINE NUCLEOSIDE PHOSPHORYLASE	2,10	AGF	5,3
1afk	RIBONUCLEASE A	1,70	PAP	6,62
1ai5	PENICILLIN AMIDOHYDROLASE	2,36	MNP	3,72
1ajp	PENICILLIN AMIDOHYDROLASE	2,31	OMD	2,23
1alw	CALPAIN	2,03	ISA	6,52
1b74	GLUTAMATE RACEMASE	2,30	D-GLUTAMINE	1,3
1bhx	ALPHA THROMBIN	2,30	R56	6,84
1br6	RICIN	2,30	PTEROIC ACID	3,22
1c4u	THROMBIN	2,10	IH1	10,37
1c5o	HUMAN ALPHA THROMBIN	1,90	O-SULFO-L-TYROSINE	3,49
1c5z	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1,85	BENZAMIDINE	4,01
1c88	PROTEIN-TYROSINE PHOSPHATASE 1B	1,80	OTA	5,29
1cbx	CARBOXYPEPTIDASE A	2,00	L-BENZYLSUCCINIC ACID	6,35
1ctt	CYTIDINE DEAMINASE	2,20	DHZ	4,52
1ctu	CYTIDINE DEAMINASE	2,30	ZEB	11,92
1dfo	SERINE HYDROXYMETHYLTRANSFERASE	2,40	5-FORMYL-6-HYDROFOLIC ACID	6,7
1dud	DEOXYURIDINE 5'-TRIPHOSPHATE NUCLEOTIDOHYDROLASE	2,30	DUD	4,82
1e1x	CYCLIN-DEPENDENT PROTEIN KINASE 2	1,85	NW1	5,89
1e66	ACETYLCHOLINESTERASE	2,1	HUX	9,89
1ec9	GLUCARATE DEHYDRATASE	2,00	XYLAROHYDROXAMATE	3,1
1egh	METHYLGLYOXAL SYNTHASE	2,00	2-PHOSPHOGLYCOLIC ACID	5,7
1eix	OROTIDINE 5'-MONOPHOSPHATE DECARBOXYLASE	2,50	BMQ	11,06
1f3e	QUEUINE TRNA-RIBOSYLTRANSFERASE	1,85	3,5-DIAMINOPHTHALHYDRAZIDE	6,7
1f4f	THYMIDYLATE SYNTHASE	2,00	TP3	4,62
1f4g	THYMIDYLATE SYNTHASE	1,75	TP4	6,48
1f57	CARBOXYPEPTIDASE A	1,75	D-CYSTEINE	5,64
1fjs	COAGULATION FACTOR XA	1,92	Z34	9,96
1fki	FK506 BINDING PROTEIN	2,20	SB1	7
1fkw	ADENOSINE DEAMINASE	2,40	PURINE RIBOSIDE	5,05
1fm9	RETINOIC ACID RECEPTOR RXR-ALPHA	2,10	570	9
1fv0	PHOSPHOLIPASE A2	1,70	9-HYDROXY ARISTOLOCHIC ACID	5,93
1g32	PROTHROMBIN	1,90	R11	6,11
1g3e	BETA-TRYPSIN	1,80	109	5,38

1gah	GLUCOAMYLASE-471	2,00	ALPHA-ACARBOSE	12
1gai	GLUCOAMYLASE-471	1,70	DIHYDRO-ACARBOSE	8
1gcz	MACROPHAGE MIGRATION INHIBITORY FACTOR	1,90	YZ9	5,13
1gi8	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1,75	BMZ	5,05
1gja	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1,56	135	5,42
1gjc	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1,73	130	6,35
1gpk	ACETYLCHOLINESTERASE	2,10	HUPERAINE A	5,37
1gyy	HYPOTHETICAL PROTEIN YDCE	1,35	FHC	3,64
1h1s	CELL DIVISION PROTEIN KINASE 2	2,00	4SP	8,22
1h23	ACETYLCHOLINESTERASE	2,15	E12	8,35
1hfs	STROMELYSIN-1	1,70	L04	8,7
1hii	HIV-2 PROTEASE	2,30	C20	7,28
1hlk	BETA-LACTAMASE, TYPE II	2,50	113	5
1hqg	ARGINASE 1	2,00	ORNITHINE	3
1hsh	HIV-1 PROTEASE	1,90	MK1	9,42
1i00	THYMIDYLATE SYNTHASE	2,50	TOMUDEX	6,34
1i5r	TYPE 1 17 BETA-HYDROXYSTEROID DEHYDROGENASE	1,60	HYC	8,52
1ik4	METHYLGLYOXAL SYNTHASE	2,00	PHOSPHOGLYCOLOHYDROXAMIC ACID	7,41
1j01	BETA-1,4-XYLANASE	2,00	XIL	6,47
1j14	TRYPSIN II, ANIONIC	2,40	BENZAMIDINE	4,49
1j17	TRYPSIN II, ANIONIC	2,00	ZEN	5,22
1j4r	FK506-BINDING PROTEIN	1,80	1	7,72
1jcx	2-DEHYDRO-3-DEOXYPHOSPHOOCTONATE ALDOLASE	1,80	PAI	5,15
1jqd	HISTAMINE N-METHYLTRANSFERASE	2,28	S-ADENOSYL-L-HOMOCYSTEINE	5,16
1jys	MTA/SAH NUCLEOSIDASE	1,90	ADENINE	3,52
1k1y	4-ALPHA-GLUCANOTRANSFERASE	2,40	ALPHA-ACARBOSE	3,22
1kv5	TRIOSEPHOSPHATE ISOMERASE, GLYCOSOMAL	1,65	DTT	4,22
1lox	15-LIPOXYGENASE	2,40	RS7	5,52
1lrt	INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE	2,20	BOG	5,64
1m0n	2,2-DIALKYLGLYCINE DECARBOXYLASE	2,20	HCP	2,22
1m2p	CASEIN KINASE II, ALPHA CHAIN	2,00	HNA	6,11
1m9n	AICAR TRANSFORMYLASE-IMP CYCLOHYDROLASE	1,93	AMZ	6,92
1meu	HIV-1 PROTEASE	1,90	DMP	6,1
1mfi	MACROPHAGE MIGRATION INHIBITORY FACTOR	1,80	FHC	5,59
1mmp	GELATINASE A	2,30	RSS	6,07
1mmr	MATRILYSIN	2,40	SRS	5,4
1moq	GLUCOSAMINE 6-PHOSPHATE SYNTHASE	1,57	GLUCOSAMINE 6-PHOSPHATE	3,46

1mrs	THYMIDYLATE KINASE	2,00	5HU	3,96
1n2v	QUEUINE TRNA-RIBOSYLTRANSFERASE	2,10	BDI	4,08
1n3i	PURINE NUCLEOSIDE PHOSPHORYLASE	1,90	DIH	8,89
1n4h	NUCLEAR RECEPTOR ROR-BETA	2,10	RETINOIC ACID	6,55
1n51	XAA-PRO AMINOPEPTIDASE	2,30	ATN	4,85
1n5r	ACETYLCHOLINESTERASE	2,25	ALPHA-L-FUCOSE	5,66
1nc1	MTA/SAH NUCLEOSIDASE	2,00	MTH	6,12
1ndv	ADENOSINE DEAMINASE	2,30	FRO	5,92
1nhu	HEPATITIS C VIRUS NS5B RNA-DEPENDENT RNA	2,00	153	5,66
1njs	PHOSPHORIBOSYLGLYCINAMIDE FORMYLTRANSFERASE	1,98	KEU	7,82
1nm6	THROMBIN	1,80	L86	10,05
1nny	PROTEIN-TYROSINE PHOSPHATASE	2,40	515	7,66
1no6	PROTEIN-TYROSINE PHOSPHATASE	2,40	794	4,41
1nvr	SERINE/THREONINE-PROTEIN KINASE CHK1	1,80	STAUROSPORINE	8,11
1o0f	RIBONUCLEASE PANCREATIC	1,50	ADENOSINE-3'-5'-DIPHOSPHATE	5,3
1o2j	BETA-TRYPSIN	1,65	656	6,92
1o3h	BETA-TRYPSIN	1,53	907	7,3
1o86	ANGIOTENSIN CONVERTING ENZYME	2,0	LPR	9,57
1owh	UROKINASE-TYPE PLASMINOGEN ACTIVATOR	1,61	239	7,4
1pb8	N-METHYL-D-ASPARTATE RECEPTOR SUBUNIT 1	1,45	D-SERINE	5,15
1pb9	N-METHYL-D-ASPARTATE RECEPTOR SUBUNIT 1	1,60	4AX	3,62
1pbq	N-METHYL-D-ASPARTATE RECEPTOR SUBUNIT 1	1,90	DK1	6,27
1pkx	BIFUNCTIONAL PURINE BIOSYNTHESIS PROTEIN PURH	1,90	XANTHOSINE-5'-MONOPHOSPHATE	6,92
1pr5	PURINE NUCLEOSIDE PHOSPHORYLASE	2,50	TBN	3,92
1pro	HIV-1 PROTEASE	1,80	A88	11,3
1pxo	CELL DIVISION PROTEIN KINASE 2	1,96	CK7	8,7
1pzp	BETA-LACTAMASE TEM	1,45	FTA	3,31
1q65	QUEUINE TRNA-RIBOSYLTRANSFERASE	2,10	BHB	5,46
1q84	ACETYLCHOLINESTERASE	2,45	TZ4	11,05
1qan	ERMC' METHYLTRANSFERASE	2,40	S-ADENOSYL-L-HOMOCYSTEINE	4,48
1qbq	FPT ALPHA-SUBUNIT	2,40	HFP	8,3
1qbv	THROMBIN	1,80	PPX	5,39
1qhc	RIBONUCLEASE A	1,70	PUA	7,57
1qin	LACTOYLGLUTATHIONE LYASE	2,00	GIP	8
1qq9	AMINOPEPTIDASE	1,53	METHIONINE	2,06
1r1h	NEPRILYSIN	1,95	BIR	8,92
1rdl	MANNOSE-BINDING PROTEIN-C	1,70	O1-METHYL-MANNOSE	2,24
1re8	CAMP-DEPENDENT PROTEIN KINASE	2,10	BD2	9,52

1rej	CAMP-DEPENDENT PROTEIN KINASE	2,20	B1L	8,3
1rql	PHOSPHONOACETALDEHYDE HYDROLASE	2,40	VINYLSULPHONIC ACID	2,75
1siv	SIV PROTEASE	2,50	PSI	8,08
1ssq	SERINE ACETYLTRANSFERASE	1,85	CYSTEINE	6
1t4v	PROTHROMBIN	2,00	14A	7,68
1ta2	THROMBIN	2,30	176	8,52
1tcw	SIV PROTEASE	2,40	IM1	6,02
1tkb	TRANSKETOLASE	2,30	N1T	8
1trd	TRIOSEPHOSPHATE ISOMERASE	2,50	PHOSPHOGLYCOLOHYDROXAMIC ACID	5,4
1uj5	RIBOSE 5-PHOSPHATE ISOMERASE	2,00	RIBULOSE-5-PHOSPHATE	3,05
1uou	THYMIDINE PHOSPHORYLASE	2,11	CMU	7,7
1upf	URACIL PHOSPHORIBOSYLTRANSFERASE	2,30	5-FLUOROURACIL	4,6
1uwt	BETA-GALACTOSIDASE	1,95	GTL	5,97
1uz1	BETA-GLUCOSIDASE A	2,0	IFL	6,89
1uz4	MAN5A	1,71	IFL	3,4
1v2l	TRYPSIN	1,60	BENZAMIDINE	4,29
1v48	PURINE NUCLEOSIDE PHOSPHORYLASE	2,20	HA1	7,8
1vfn	PURINE-NUCLEOSIDE PHOSPHORYLASE	2,15	HYPOXANTHINE	5,6
1wcq	SIALIDASE	2,1	DAN	6,26
1wvj	IONOTROPIC GLUTAMATE RECEPTOR 2	1,75	IBC	6,73
1x1z	OROTIDINE 5'-PHOSPHATE DECARBOXYLASE	1,45	BMP	11,06
1x8j	RETINOL DEHYDRATASE	2,35	ADENOSINE-3'-5'-DIPHOSPHATE	6,96
1x8t	3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE	1,90	RC1	7,8
1xff	GLUCOSAMINE--FRUCTOSE-6-PHOSPHATE	1,80	GLUTAMIC ACID	4,82
1xgi	BETA-LACTAMASE	1,96	NST	4,85
1xgj	BETA-LACTAMASE	1,97	HTC	6
1y1m	GLUTAMATE [NMDA] RECEPTOR SUBUNIT ZETA 1	1,80	AC5	1,82
1yds	C-AMP-DEPENDENT PROTEIN KINASE	2,20	IQS	5,92
1yqy	LETHAL FACTOR	2,30	915	7,62
1z1r	POL POLYPROTEIN	1,85	HBH	9,22
1z4n	BETA-PHOSPHOGLUCOMUTASE	1,97	GL1	4,52
1zpa	POL POLYPROTEIN	2,02	A83	8,4
1zs0	NEUTROPHIL COLLAGENASE	1,56	EIN	6,15
1zvx	NEUTROPHIL COLLAGENASE	1,87	FIN	9,22
2afw	GLUTAMINYL-PEPTIDE CYCLOTRANSFERASE	1,56	AHN	4,77
2aou	HISTAMINE N-METHYLTRANSFERASE	2,30	S-HYDROXYCYSTEINE	7,73
2arm	PHOSPHOLIPASE A2 VRV-PL-VIIIA	1,23	OIN	8,13
2b07	TYROSINE-PROTEIN PHOSPHATASE, NON-RECEPTOR	2,10	598	6,43

	TYPE			
2b7d	COAGULATION FACTOR VII	2,24	C1B	8,7
2boh	COAGULATION FACTOR XA	2,2	IIA	8,52
2bvd	ENDOGLUCANASE H	1,6	ISX	6
2bz6	BLOOD COAGULATION FACTOR VIIA	1,6	346	7,09
2bza	TRYPSIN	1,90	BENZYLAMINE	2,8
2c02	NONSECRETORY RIBONUCLEASE	2,0	ADENOSINE-5'-DIPHOSPHATE	4,04
2c3j	SERINE/THREONINE-PROTEIN KINASE CHK1	2,1	DEBROMOHYMENIALDISINE	6,18
2c3l	SERINE/THREONINE-PROTEIN KINASE CHK1	2,35	IDZ	5,07
2ceq	BETA-GALACTOSIDASE	2,14	GLUCOIMIDAZOLE	7,28
2d1n	COLLAGENASE 3	2,37	SM-25453	8,15
2dua	PHOSPHONOPYRUVATE HYDROLASE	2,00	XYLOPYRANOSE	4,77
2erz	CAMP-DEPENDENT PROTEIN KINASE, ALPHA-CATALYTIC	2,20	TPO	5,66
2f7p	ALPHA-MANNOSIDASE II	1,28	2SK	6,6
2fai	ESTROGEN RECEPTOR	2,10	459	6,24
2fdp	BETA-SECRETASE 1	2,50	FRP	7,59
2ff1	IAG-NUCLEOSIDE HYDROLASE	2,07	IMH	8,21
2fzc	ASPARTATE CARBAMOYLTRANSFERASE CATALYTIC CHAIN	2,10	CTP	2,7
2g8r	RIBONUCLEASE PANCREATIC	1,70	N3E	3,99
2gst	GLUTATHIONE S-TRANSFERASE	1,80	GPS	6,07
2gv6	SUPPRESSOR OF TUMORIGENICITY 14	2,10	730	7,34
2gvj	NICOTINAMIDE PHOSPHORIBOSYLTRANSFERASE	2,10	DGB	9,52
2gvv	PHOSPHOTRIESTERASE	1,73	DI9	3,9
2hdq	BETA-LACTAMASE	2,10	C21	1,4
2hh5	CATHEPSIN S	1,80	GNQ	7,49
2i0g	ESTROGEN RECEPTOR BETA	2,50	I0G	9,72
2ihq	ANDROGEN RECEPTOR	2,00	LG7	8,49
2iuz	CHITINASE	1,95	D1H	5,55
2j47	GLUCOSAMINIDASE	1,98	GDV	5,41
2j4i	COAGULATION FACTOR X	1,8	GSJ	9
2qwb	NEURAMINIDASE	2,00	BETA-D-MANNOSE	2,74
2qwd	NEURAMINIDASE	2,00	4AM	4,85
2sim	SIALIDASE	1,60	DAN	3,42
2usn	STROMELYSIN-1	2,20	IN8	6,51
3pcb	PROTOCATECHUATE 3,4-DIOXYGENASE	2,19	3-HYDROXYBENZOIC ACID	2,4
4tln	THERMOLYSIN	2,30	LNO	3,72
6fiv	RETROPEPSIN	1,90	3TL	8,08
7std	SCYTALONE DEHYDRATASE	1,80	CRP	10,72
830c	MMP-13	1,60	RS1	9,28