UNIVERSIDADE DE SÃO PAULO

FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

**Estudos de modelagem molecular e relação estrutura atividade
da oncoproteína hnRNP K e ligantes**

Vinicius Barreto da Silva

**Ribeirão Preto
2007**

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

**Estudos de modelagem molecular e relação estrutura atividade
da oncoproteína hnRNP K e ligantes**

Vinicius Barreto da Silva

Ribeirão Preto
2007

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS DE RIBEIRÃO PRETO

**Estudos de modelagem molecular e relação estrutura atividade
da oncoproteína hnRNP K e ligantes**

Dissertação de mestrado apresentada
ao Programa de Pós-Graduação em
Ciências Farmacêuticas para obtenção
do Título de Mestre em Ciências
Farmacêuticas.

Área de Concentração: Física Biológica

Orientado: Vinicius Barreto da Silva

Orientador: Carlos Henrique Tomich
de Paula da Silva

Ribeirão Preto
2007

SILVA, V. B	Estudos de modelagem molecular e relação estrutura atividade da oncoproteína hnRNP K e ligantes		MESTRADO FCFRP/USP 2008
--------------------	--	--	--

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

SILVA, VINICIUS BARRETO

Estudos de modelagem molecular e relação estrutura atividade da oncoproteína hnRNP K e ligantes. Ribeirão Preto, 2008.

129p.; il, 30cm

Dissertação de Mestrado, apresentada à Faculdade de Ciências Farmacêuticas de Ribeirão Preto/USP - Área de concentração: Física Biológica.

Orientador: SILVA, CARLOS HENRIQUE TOMICH DE PAULA

1. Câncer. 2. hnRNP K. 3. Modelagem molecular. 4. Planejamento racional de fármacos.

Folha de Aprovação

Vinicius Barreto da Silva

Estudos de modelagem molecular e relação estrutura atividade da oncoproteína hnRNP K e ligantes.

Dissertação de Mestrado apresentada
ao Programa de Pós-Graduação em
Ciências Farmacêuticas para obtenção
do Título de Mestre em Ciências
Farmacêuticas.

Área de Concentração: Física Biológica

Orientador: Carlos Henrique Tomich de
Paula da Silva

Aprovado em:

Banca Examinadora

Prof(a). Dr(a). _____

Instituição: _____ Assinatura: _____

Prof(a). Dr(a). _____

Instituição: _____ Assinatura: _____

Prof(a). Dr(a). _____

Instituição: _____ Assinatura: _____

“Não aceiteis o que é de hábito como coisa natural, pois em tempo de desordem sangrenta, de confusão organizada, de arbitrariedade consciente, de humanidade desumanizada, nada deve parecer natural, nada deve parecer impossível de mudar”.

Bertold Brecht (1898-1956). Escritor e dramaturgo alemão, além de grande teórico teatral.

“ Feliz aquele que transfere o que sabe e aprende o que ensina”

“O que vale na vida não é o ponto de partida e sim a caminhada. Caminhando e semeando, no fim terás o que colher”

Cora Coralina (1889-1985). Pseudônimo da grande poetisa do Estado de Goiás, Ana Lins do Guimarães Peixoto Brêtas.

“Aqui tem um bando de louco
Louco por ti Corinthians
Aqueles que acham que é pouco
Eu vivo por ti Corinthians
Eu canto até ficar rouco
Eu canto pra te empurrar
Vamos vamos meu timão
Vamos timão
Não para de lutar”

Canto eternizado pela torcida do Corinthians

Dedico este trabalho a toda minha família, especialmente meus pais, Marcio Barreto e Martha Beatriz, que se esforçaram ao extremo para que eu pudesse ter uma educação de qualidade, e à minha noiva Naira Tainá.

AGRADECIMENTOS

Primeiramente à Deus, por ter me abençoado e me dado saúde, ânimo e vontade de trabalhar.

Ao Prof. Dr. Carlos Tomich, pela brilhante orientação que recebi durante o desenvolvimento da dissertação, além da grande amizade construída durante este período.

À toda minha família, pelo carinho, compreensão, sacrifício, crédito e confiança depositados em mim.

À Naira Tainá, pelo amor, carinho e dedicação no dia a dia, que me ajudaram bastante durante esta caminhada.

À Profa. Dra. Andréia Machado Leopoldino pela colaboração, com a qual pretendo, em breve, trabalhar em conjunto em novos projetos.

Aos colegas do Laboratório Computacional de Química Farmacêutica, Adriana, Josy, bin e xita, com os quais compartilhei momentos de trabalho, descontração e alegria.

Aos técnicos dos laboratórios de Química Farmacêutica, Luis Otávio e Claudia, pelo convívio nas aulas práticas das turmas de graduação e disposição para ajudar nos entraves burocráticos.

Aos vigilantes da Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Henrique, Luciano, Silvio, Clóvis, Sérgio, Antônio, Paulo, Paulão, Lima e Gilmar, pelo convívio e pelas agradáveis conversas nos vários finais de semana que tive que esperar a chuva passar para poder ir embora para casa.

À Profa. Dra. Ivone Carvalho pela colaboração no trabalho e nas publicações.

Aos colegas de pós-graduação da FCFRP, Lilian, Vanessa, Adriane, Pedro, Peterson, Daniel, Margareth, Luciano, Maristela, Flávio, Warley, Denise, Julierme, Michelle, Fernanda, Gaby, Neri e Willian pelo trabalho em conjunto e a amizade cultivada neste período.

Ao Prof. Dr. Edemilson Cardoso pelas dicas e proveitosas discussões na época de graduação, que me incentivaram a buscar a FCFRP como reduto para o desenvolvimento deste trabalho.

À CAPES pela bolsa de estudos concedida, primordial no desenvolvimento deste trabalho.

À Ana, funcionária da Seção de Pós-graduação, sempre prestativa e disposta quando precisei da sua ajuda nas questões burocráticas.

Ao Zé Maria, funcionário da FCFRP, que me acolheu muito bem quando cheguei a Ribeirão Preto.

À Faculdade de Ciências Farmacêuticas de Ribeirão Preto, pela infra-estrutura oferecida, e a todos os seus docentes e funcionários pelo convívio diário.

RESUMO

O projeto Genoma Câncer brasileiro (Projeto Genoma Humano do Câncer - PGHC), financiado pela FAPESP e pelo Instituto Ludwig de Pesquisa sobre o câncer, buscou identificar os genes expressos nos tipos mais comuns de câncer no Brasil. Tal projeto conseguiu identificar aproximadamente um milhão de seqüências de genes de tumores freqüentes no Brasil. A contribuição brasileira foi maior para tumores de cabeça e pescoço, mama e cólon. Uma das iniciativas mais recentes e estimuladas pelo Projeto Genoma Humano do Câncer é o projeto Genoma Clínico, o qual visa desenvolver novas formas de diagnóstico e tratamento do câncer através do estudo de genes expressos. A partir da análise molecular de tecidos saudáveis e neoplásicos em diferentes estágios, é possível identificar marcadores relacionados aos estágios de câncer, permitindo escolhas de terapias mais adequadas e eficientes. A proteína hnRNP K foi identificada como um desses marcadores, em neoplasias da região da cabeça e pescoço, sendo objetivo deste estudo a aplicação de técnicas de bioinformática e modelagem molecular no planejamento de candidatos a fármacos antineoplásicos contra a atividade da proteína. A proteína hnRNP K apresenta diversas funções e é encontrada nos mais diversos compartimentos celulares, interferindo, basicamente, no sistema de expressão gênica. Essa proteína apresenta 3 domínios KH, os quais são responsáveis por sua ligação com DNA e RNA. Os modelos dos domínios KH foram construídos através da estratégia de modelagem molecular por homologia estrutural. Após “screening” em bases de dados virtuais de compostos com propriedades “drug-like”, 15 compostos com potencial de interação com o domínio KH3 foram selecionados. Os modos de ligação para cada um dos compostos no sítio ligante do domínio KH3 foram sugeridos e os resultados comparados com os campos de interação molecular gerados para vários grupos químicos de prova diferentes. Simulações de dinâmica molecular foram realizadas com o intuito de avaliar a estabilidade dos compostos selecionados, que também foram avaliados quanto à presença de grupamentos toxicofóricos em sua estrutura.

ABSTRACT

The Brazilian Project “Genoma Câncer” (PGHC) supported by FAPESP and the Ludwig Institute for Cancer Research, intended to identify the genes involved in the most common cases of cancer in Brazil. In this project about a million of gene sequences were identified. The major contribution was made in breast, colorectal and head and neck cancer. The results obtained stimulate the creation of another project, called “Genoma Clínico”, which intend to develop new trends in treatments and diagnosis of cancer based on the study of genes. Analyzing healthy and neoplastic tissues in different stages, it is possible to identify molecular markers related to the prognosis of cancer, allowing the use of more adequate therapies. The hnRNP K protein was identified as a molecular marker in head and neck cancer, where the objective of this work lies in the application of bioinformatics and molecular modeling strategies to plan antineoplastic drug candidates that could act against hnRNP K protein. The hnRNP K protein is encountered in all cellular compartments and act, basically, in the gene expression pathways. Its structure is composed by three KH domains that mediate interactions with DNA and RNA. Models of KH domains were built by homology modeling. After the virtual screening simulations performed with drug-like compounds databases, 15 compounds were selected as potential ligands of KH3 domain of hnRNP K. The binding modes suggested for these compounds, by docking simulations, were compared with molecular interaction field data generated for different chemical probes. Molecular dynamics simulations were performed to evaluate the stability of the binding modes suggested. The molecular structure of the potential ligands were also evaluated to identify toxicophoric groups.

LISTA DE FIGURAS

- Figura 1.** Ilustração do caráter modular e da disposição dos domínios presentes na proteína hnRNP K. Além dos domínios KH, que se ligam a ácidos nucleicos, pode-se destacar também a presença de domínios responsáveis por interações com outras proteínas, como GRGG “box”, um domínio de ligação a motivos SH3 e um domínio de ligação a proteínas quinases. A isoforma α da proteína apresenta 464 resíduos de aminoácidos enquanto que a isoforma β apresenta 463. **10**
- Figura 2.** (A) Arranjo estrutural típico de um domínio KH. (B) Representação do domínio KH3 da proteína hnRNP K em complexo com o oligonucleotídeo de ssDNA TCCCT (T1, C2, C3, C4, T5). (C) Representação da superfície de potencial eletrostático do domínio KH3 formando uma fenda com um centro hidrofóbico que acomoda o oligonucleotídeo TCCCT. **11**
- Figura 3.** Modelo de atuação da proteína hnRNP K, funcionando como uma plataforma para integrar sinais das cascatas de quinases para um sítio de processos dirigidos ao RNA. **19**
- Figura 4.** Verificação da sobreposição dos resíduos para realização de possíveis correções no alinhamento. **35**
- Figura 5.** Modelagem Molecular por Satisfação de Restrições Espaciais. Inicialmente, as estruturas tridimensionais (‘3D’) conhecidas são alinhadas com a sequência-alvo (‘SEQ’). A seguir, parâmetros espaciais, tais como distâncias C α -C α , ligações de hidrogênio e torções, são transferidos do molde para o alvo. Com isso, várias restrições espaciais são extraídas. Com a aplicação do campo de força, o modelo é então obtido satisfazendo-se, ao máximo possível, todas essas restrições. **42**
- Figura 6.** Gráfico de Ramachandran do modelo do domínio KH1, gerado pelo software **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). Os resíduos de glicina (7 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu Ca não apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados. **64**
- Figura 7.** Representação dos gráficos de cinco propriedades estruturais da cadeia principal. Os valores do modelo do domínio KH1 são marcados por quadrados e comparados com estruturas bem definidas com resolução estrutural similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média. **65**
- Figura 8.** Representação do perfil 3D do modelo do domínio KH1. Os valores dos 10 primeiros resíduos de cada extremidade são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore. **67**
- Figura 9.** Gráfico de Ramachandran do modelo do domínio KH2, gerado pelo “software” **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). Os resíduos de glicina (6 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu Ca não apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados. **68**
- Figura 10.** Representação das propriedades estruturais da cadeia principal. Os valores do modelo do domínio KH2 são marcados por quadrados e comparados com estruturas bem definidas com resolução similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média. **69**
- Figura 11.** Localização dos resíduos (em amarelo) com baixo índice da qualidade de contato no modelo do domínio KH2. **71**

Figura 12. Representação do perfil 3D do modelo do domínio KH2. Os valores dos 10 resíduos mais próximos de ambas as extremidades são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore. 71

Figura 13. Gráfico de Ramachandran do modelo do domínio KH3, gerado pelo “software” **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). 72 Os resíduos de glicina (10 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu C α não apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados.

Figura 14. Representação das propriedades estruturais da cadeia principal. Os valores obtidos do modelo do domínio KH3 (isoforma a) são marcados por quadrados e comparados com estruturas bem definidas com resolução similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média. 73

Figura 15. Localização do resíduo LEU 45 (em amarelo) no modelo 3 do domínio KH3 (isoforma a) 75

Figura 16. Representação do perfil 3D do modelo da isoforma a do domínio KH3. Os valores dos 10 resíduos mais próximos de ambas extremidades são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore. 75

Figura 17. Estrutura do domínio KH3 (código PDB 1J5K) em complexo com a sequência oligonucleotídica TCCCT. Os átomos de carbono dos resíduos de aminoácidos do sítio ligante da proteína estão indicados em verde e os da sequência nucleotídica em amarelo. 77

Figura 18. Bases de dados e suas respectivas subcoleções de compostos utilizadas nas simulações de “screening” virtual. 78

Figura 19. Fórmula estrutural dos compostos da base de dados Ilibdiverse que apresentaram maior “escore” nas simulações de “screening” virtual. 79

Figura 20. Fórmula estrutural dos compostos da base de dados IResearch Library que apresentaram maior “escore” nas simulações de “screening” virtual. 79

Figura 21. Fórmula estrutural dos compostos da base de dados Chembridge que apresentaram maior “escore” nas simulações de “screening” virtual e suas respectivas subcoleções de compostos. 80

Figura 22. Orientações de melhor escore dos compostos selecionados nas simulações de “screening” virtual com o sítio ligante do domínio KH3. 82

Figura 23. Orientação dos compostos 1 e 14 no sítio ligante do domínio KH3, representados por A e B, respectivamente. A orientação do composto 1 (átomos de carbono em azul) é mostrada em comparação com a orientação do oligonucleotídeo TCCCT (carbonos em magenta) no complexo depositado no PDB (código 1J5K). As regiões circuladas destacam os átomos de oxigênio carbonílico dos grupamentos amida dos compostos em torno do resíduo de R59 da proteína. 83

Figura 24. Orientação dos compostos 3, 12, 13 e 10 no sítio ligante do domínio KH3, representados por A, B, C e D, respectivamente. A orientação do composto 3 (átomos de carbono em amarelo) é mostrada em comparação com a orientação do oligonucleotídeo TCCCT (carbonos em magenta) do complexo depositado no PDB (código 1J5K). As regiões circuladas destacam os átomos de oxigênio dos compostos que interagem com o resíduo de R59 do domínio KH3. Em D a linha tracejada representa uma interação entre o composto 10 e o resíduo de K31. 84

Figura 25. Orientações de melhor escore dos compostos 9 e 15 no sítio ligante do domínio KH3, representados por A e B, respectivamente. As regiões circuladas destacam os átomos de oxigênio carboxílico e/ou carbonílico dos compostos em torno do resíduo de R59 da proteína. 85

- Figura 26.** Orientações de melhor escore dos compostos 2 e 4 em comparação com a orientação do oligonucleotídeo TCCCT. (A) composto 2, com destaque para o posicionamento do anel tiazol próximo ao resíduo de R40. (B) composto 4 ao redor do resíduo de R59, com destaque para a posição dos átomos de nitrogênio que quase se sobrepõem aos átomos de nitrogênio da citosina 2 da sequência oligonucleotídica TCCCT. **86**
- Figura 27.** Orientações de melhor escore dos compostos 5, 8 e 11, representados por A, B e C, respectivamente. As regiões circuladas destacam os átomos de oxigênio carbonílico dos compostos que interagem com o resíduo R59. **87**
- Figura 28.** Orientação de maior escore do composto 6 no sítio ligante do domínio KH3. **88**
- Figura 29.** Orientação de melhor escore do composto 7 (19(R)-hidroxiprostaglandinaF2a) no sítio ligante do domínio KH3. Em círculos estão destacados um grupamento hidroxila próximo a R59 e um grupamento carboxilato próximo a R40. **88**
- Figura 30.** Orientações do oligonucleotídeo TCCC (A) e dos compostos 3 (B) e 6 (C) no sítio ligante do domínio KH3. As superfícies representam os sítios virtuais de interação hidrofóbica. As regiões dos ligantes mais próximas dos sítios hidrofóbicos são evidenciadas por círculos. **91**
- Figura 31.** Orientações do oligonucleotídeo TCCC (A) e dos compostos 1 (B), 3 (C), 5 (D), 7 (E) e 8 (F), pertencentes às bases de dados IResearch Library e Ilibdiverse, no sítio ligante do domínio KH3. As superfícies representam os sítios virtuais de interação. As regiões dos ligantes mais próximas dos sítios são evidenciadas por círculos. **92**
- Figura 32.** Orientações dos compostos 9 (A), 10 (B), 11 (C), 12 (D), 13 (E), 14 (F) e 15 (G), pertencentes à base de dados Chembridge, no sítio ligante do domínio KH3. As superfícies representam os sítios virtuais que favorecem interações polares. As regiões favoráveis dos ligantes mais próximas dos sítios são evidenciadas por círculos. **93**
- Figura 33.** Orientação do composto 7 no sítio ligante do domínio KH3. Os elementos coloridos em ciano representam os átomos de oxigênio das moléculas de água que foram adicionadas ao sistema. **95**
- Figura 34.** Gráficos da energia total em função do tempo de simulação dos 15 compostos selecionados nas simulações de “screening” virtual complexados ao domínio KH3. **96**
- Figura 35.** Gráficos referentes à variação dos valores de RMSD em função do tempo de simulação dos 15 compostos selecionados nas simulações de “screening” virtual em complexo com o domínio KH3. **99**
- Figura 36.** Avaliação da estabilidade da interação (indicada por uma linha tracejada) do resíduo de R59 com o elemento C2 da tétrade oligonucleotídica, através do cálculo dos valores de RMSD em função do tempo de simulação. **101**
- Figura 37.** Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 1 e R59. (B) composto 14 e R59. **102**
- Figura 38.** Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 3 e R59. (B) composto 10 e R59. (C) composto 12 e R59. (D) composto 13 e R59. **103**
- Figura 39.** Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 9 e R59. (B) composto 15 e R59. **104**
- Figura 40.** Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 2 e R40. (B) composto 4 e R59. **105**

- Figura 41.** Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 5 e R40. (B) composto 8 e R59. (C) composto 11 e R59. 106
- Figura 42.** Avaliação da estabilidade da interação (indicada por uma linha tracejada) do resíduo de R59 com uma hidroxila do composto 7, através do cálculo dos valores de RMSD em função do tempo 107
- Figura 43.** Subestruturas de amidas aromáticas responsáveis pelos alertas tóxicos gerados para os compostos 1, 2, 3, 9, 10, 11 e 12. 108
- Figura 44.** Subestruturas de fenóis e precursores e do grupamento hidrazida, responsáveis pelos alertas de hipersensibilidade cutânea gerados para os compostos 1, 5 e 6. 109
- Figura 45.** Subestrutura básica de um anidrido ácido presente nos anéis oxazina dos compostos 3, 10, 12 e 13. 109
- Figura 46.** Anel de pirimidina do composto 4, responsável pelo alerta de toxicidade gerado para o composto 4. 110
- Figura 47.** Éster de cianohidrina presente no composto 5, responsável pelo alerta de toxicidade gerado. 111
- Figura 48.** Precursor de anilina presente na estrutura do composto 14, responsável pelo alerta de toxicidade gerado. 111
- Figura 49.** Diarilcetona presente no composto 15, responsável pelo alerta de fototoxicidade gerado. 112

LISTA DE TABELAS

Tabela 1. Parceiros moleculares da proteína hnRNP K nos diversos processos de expressão gênica e na transdução de sinais.	14
Tabela 2. Exemplos do envolvimento da proteína hnRNP K de mamíferos em múltiplos processos de expressão gênica.	15
Tabela 3. Estrutura primária do domínio KH3 das isoformas a e b da proteína hnRNP K. Em destaque estão os resíduos de aminoácidos diferentes (região C-terminal) entre as duas isoformas.	30
Tabela 4. Descrição das seqüências selecionadas na busca com o BLAST , e seus respectivos códigos PDB, com os valores de identidade seqüencial obtidos. Para as estruturas resolvidas por cristalografia de raios-X é indicada a resolução, e para as estruturas resolvidas por ressonância magnética nuclear é indicado RMN.	59
Tabela 5. Comparação entre os valores de identidade seqüencial obtidos pelos “softwares” BLAST e Multalign .	61
Tabela 6. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH1, na ordem requisitada pelo “software” AMPS . Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI, 4 – 2AXY e 5 – seqüência alvo KH1.	62
Tabela 7. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH2, na ordem requisitada pelo “software” AMPS . Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI, 4 – 1WVN, 5 – 2AXY e 6 – seqüência alvo KH2.	62
Tabela 8. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH3 (isoforma a), na ordem requisitada pelo “software” AMPS . Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI e 4 – seqüência alvo da isoforma a do domínio KH3.	62
Tabela 9. Valores dos índices da qualidade de contato para todos os resíduos do modelo do domínio KH1 e o índice total do modelo.	66
Tabela 10. Valores dos índices da qualidade de contato para todos os resíduos do modelo do domínio KH2 e o índice do modelo.	70
Tabela 11. Valores dos índices da qualidade de contato para todos os resíduos do modelo do domínio KH3 (isoforma a) e o índice do modelo.	74
Tabela 12. Nome IUPAC e valores obtidos pela função Goldscore nas simulações de “docking” flexível dos quinze compostos selecionados nas simulações de “screening” virtual.	81
Tabela 13. Propriedades físico-químicas relacionadas à Regra dos Cinco dos 15 compostos selecionados nas simulações de “screening” virtual.	89

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
A	Alanina
ADMET	Absorção, distribuição, metabolismo, excreção e toxicidade
AMPS	Alignment of Multiple Pair Segments
BHE	Barreira hematoencefálica
BLAST	Basic Local Alignment Search Tool
C	Cisteína
D	Aspartato
DEREK	Deductive Estimation of Risk from Existing Knowledge
dsDNA	Fita dupla de DNA
E	Glutamato
EGFR	Receptor do fator de crescimento epidermal
F	Fenilalanina
G	Glicina
GPCR	Receptores acoplados à proteína G
H	Histidina
I	Isoleucina
IUPAC	União Internacional de Química Pura e Aplicada
K	Lisina
L	Leucina
M	Metionina
MIF	Campos de interação molecular
MSP	Maximal Pair Segments
N	Asparagina
P	Prolina
PGCH	Projeto Genoma do Câncer Humano
PDB	Banco de dados de proteína
Q	Glutamina
R	Arginina
RMN	Ressonância Magnética Nuclear
RMSD	Raiz do desvio médio quadrático

S	Serina
ssDNA	Fita simples de DNA
T	Treonina
W	Triptofano
Y	Tirosina
V	Valina

SUMÁRIO

Resumo	i
Abstract	ii
Lista de Figuras	iii
Lista de Tabelas	vii
Lista de Abreviaturas e Siglas	viii
1. INTRODUÇÃO	1
1.1. Genoma câncer	1
1.2. Câncer: aspectos gerais	2
1.3. Câncer de cabeça e pescoço	5
1.4. Proteína hnRNP K	6
1.5. Modelo para atuação da proteína hnRNP K	17
1.6. Família de proteínas hnRNP e câncer	19
1.7. Planejamento racional de fármacos <i>in silico</i>	23
2. OBJETIVOS	28
3. MÉTODOS	30
3.1. Identificação, expressão, purificação e validação do marcador protéico	30
3.2. Alinhamento de seqüências	31
3.2.1. Alinhamento global	32
3.2.2. Alinhamento local	32
3.2.3. Alinhamento múltiplo	33
3.3. BLAST	33
3.4. AMPS	34
3.5. Refinamento do alinhamento	35
3.6. Modelagem molecular por homologia estrutural	36
3.6.1. Modelagem por homologia dos domínios KH da proteína hnRNP K	41
3.7. Validação dos modelos	43
3.8. “docking” molecular	45
3.8.1. Simulações de “screening” virtual	47
3.8.2. Modelagem dos compostos selecionados	48
3.8.3. “Rescore”	48
3.9. Determinação dos potenciais de interação molecular fármaco- receptor	49
3.9.1. Potenciais eletrostáticos moleculares	49
3.9.2. Campos de interação molecular	50
3.9.2.1. Almond	52
3.10. Predições ADMET	52
3.10.1. DEREK	54
3.11. Dinâmica molecular	54
3.11.1. Simulações de dinâmica molecular	57
4. RESULTADOS E DISCUSSÃO	59
4.1. Busca de seqüências homólogas	59
4.2. Alinhamento múltiplo	60
4.3. Construção dos modelos	63
4.4. Validação dos modelos	63
4.4.1. Domínio KH1	63

4.4.2. Domínio KH2	67
4.4.3. Domínio KH3 (isoforma a)	72
4.5. Análise do complexo KH3-ssDNA	76
4.6. “Screening” virtual	77
4.7. Propriedades físico-químicas	89
4.8. Campos de interação molecular	90
4.9. Dinâmica molecular	94
4.9.1. Estabilidade energética	95
4.9.2. Estabilidade conformacional	98
4.9.3. Estabilidade das interações com R40 e R59	100
4.10. Predição de toxicidade	107
4.10.1. Amidas e aminas aromáticas	107
4.10.2. Fenóis, precursores fenólicos, hidrazidas e análogos de anidrido ácido	108
4.10.3. Pirimidina	110
4.10.4. Nitrila	110
4.10.5. Precursores de anilina	111
4.10.6. Diarilcetona	112
5. CONCLUSÕES	113
6. REFERÊNCIAS BIBLIOGRÁFICAS	117

1. INTRODUÇÃO

1.1. Genoma Câncer

O Projeto Genoma do Câncer Humano (PGCH), financiado pela FAPESP e pelo Instituto Ludwig de Pesquisa sobre o câncer, buscou identificar os genes expressos nos tipos mais comuns de câncer no Brasil. A fase de seqüenciamento foi finalizada em 2001 e seu sucesso serviu de estímulo para que outras iniciativas fossem apoiadas, tais como o “Human Transcript Validation Initiative”, e a bioinformática recebesse um grande impulso no país. O PGCH começou em abril de 1999 e conseguiu identificar, em menos de um ano, um milhão de seqüências de genes de tumores freqüentes no Brasil (REVISTA PESQUISA FAPESP, 2000). A contribuição brasileira foi maior para tumores de cabeça e pescoço, mama e cólon (intestino), e é qualitativamente importante porque, diferentemente de outros projetos, a estratégia utilizada (ORESTES) analisou prioritariamente a parte central dos genes, onde está concentrada a informação relevante para a síntese de proteínas (DUNHAN et al., 1999; de SOUZA et al., 2000).

Uma das iniciativas mais recentes e estimuladas pelo PGCH é o Projeto Genoma Clínico, o qual visa o desenvolvimento de novas formas de diagnóstico e tratamento do câncer a partir do estudo de genes expressos. Este projeto envolve oncologistas, cirurgiões e pesquisadores paulistas na análise dos genes expressos em quatro tipos de manifestação do câncer: as doenças linfoproliferativas, tumores gastrintestinais, tumores neurológicos e de cabeça e pescoço (REIS et al., 2005).

Sua meta inclui a análise da expressão gênica em neoplasias humanas e a identificação de diferenças nos perfis de expressão que possam estar relacionadas aos parâmetros clínicos e o comportamento biológico do câncer. A partir da análise molecular de tecidos saudáveis e neoplásicos em diferentes estágios, é possível identificar marcadores relacionados com as fases iniciais da transformação maligna e marcadores de prognóstico, que aumentam as chances de previsão da evolução do tumor, permitindo escolhas de terapias mais adequadas e eficientes (DUNHAN et al., 1999). A identificação desses marcadores é essencial, porque auxiliam o diagnóstico precoce e o possível sucesso do tratamento do câncer. Dentre os marcadores de câncer de cabeça e pescoço identificados pode-se destacar as proteínas hnRNP K, ZRF1, SET e MARK3.

O conhecimento gerado por pesquisas sobre a função de genes que participam do processo de gênese tumoral tem permitido o desenvolvimento de fármacos e estudos clínicos correspondentes em diferentes neoplasias. Um exemplo é o fármaco antineoplásico erlotinibe (Tarceva®), utilizado em casos de câncer de pulmão. Esse fármaco inibe especificamente a enzima tirosina quinase do EGFR (do inglês, “epidermal growth factor receptor”), bloqueando a cascata de sinais que é desencadeada pelo receptor e ligante (de Bono; Rowinski, 2002; GRIDELLI et al., 2007).

Na visão de Andrew Simpson – um inglês que reside há mais de 12 anos no Brasil e que esteve à frente de projetos de peso da ciência nacional, tais como o sequenciamento do genoma da bactéria *Xylella fastidiosa* e o PGCH– desde a descoberta da estrutura molecular do DNA, nos anos 50, o tratamento contra o câncer não mudou radicalmente, sobretudo no que diz respeito à descoberta de fármacos contra a doença. Na opinião do pesquisador, o Brasil deveria elaborar mais projetos que visem o desenvolvimento de fármacos contra o câncer, ainda que os valores necessários para essa empreitada pareçam elevados. Segundo ele, o país pode e deve ter essa ambição, e acrescenta que não se pode esquecer que a verba investida no desenvolvimento de um fármaco não é gasta de uma só vez, mas ao longo de vários anos (REVISTA PESQUISA FAPESP, 2002).

1.2. Câncer: aspectos gerais

O câncer é uma doença quase sempre associada ao estigma de mortalidade e dor. Na verdade, a palavra câncer de origem latina *cancer*, significando “caranguejo”, provavelmente foi empregada inicialmente em analogia ao modo de crescimento infiltrante, o que pode ser comparado às pernas do crustáceo, que as introduz na areia ou lama para se fixar e dificultar sua remoção (ALMEIDA et al., 2005).

Atualmente, a definição científica de câncer refere-se ao termo neoplasia, especificamente aos tumores malignos, como sendo uma doença caracterizada pelo crescimento descontrolado de células transformadas. Existem quase 200 tipos que correspondem aos vários sistemas de células do corpo, os quais se diferenciam pela capacidade de invadir tecidos e órgãos, vizinhos ou distantes (ALMEIDA et al., 2005).

Em pesquisa realizada pela Organização Mundial da Saúde, o câncer é a terceira causa de óbitos no mundo, com 12%, levando a óbito cerca de 6,0 milhões de pessoas

por ano. Atualmente, é a segunda causa de mortes por doença no Brasil, estimando-se em 2002, 337.535 casos novos e 122.600 óbitos (ALMEIDA et al., 2005).

O câncer é uma doença caracterizada pela multiplicação e propagação descontrolada de formas anômalas das próprias células do organismo. O câncer é, em grande parte, uma doença que acomete os grupos etários mais avançados, e, com os progressos na saúde pública e na ciência médica, um número grande de indivíduos atinge a idade em que se tornam mais sujeitos a desenvolver câncer (RANG; DALE; RITTER, 2001).

As células cancerosas manifestam, em graus variados, algumas características que as distinguem das células normais, como: proliferação descontrolada, capacidade de desdiferenciação e perda de função, poder de invasão e capacidade de formar metástase (RANG; DALE; RITTER, 2001).

Os fatores de risco do câncer podem ser encontrados no meio ambiente ou podem ser hereditários. A maioria dos casos (cerca de 80%) está relacionada ao meio ambiente, onde encontramos um grande número de fatores de risco. As mudanças provocadas no meio ambiente, pelo próprio homem, além dos hábitos e estilos de vida adotados podem determinar a indução de diferentes tipos de câncer (ALMEIDA et al., 2005).

As alterações que geram as neoplasias podem ocorrer em genes especiais denominados proto-oncogenes. Os proto-oncogenes são genes que normalmente controlam a apoptose, a divisão e a diferenciação celulares, podendo ser convertidos em oncogenes, responsáveis pela malignização das células normais, pela ação de agentes carcinogênicos (ALMEIDA et al., 2005; RANG; DALE; RITTER, 2001).

As células normais contêm genes que têm a capacidade de suprimir alterações malignas, denominados genes supressores tumorais ou antioncogenes. Atualmente, existem evidências de que a ocorrência de mutações nestes genes está envolvida no desenvolvimento de vários tipos de câncer. A perda de função dos genes supressores tumorais pode se constituir em um dos eventos críticos no processo de carcinogênese (RANG; DALE; RITTER, 2001).

A proliferação de células cancerosas não é controlada pelos processos que normalmente regulam a divisão celular e o crescimento dos tecidos. Este aspecto, mais do que sua velocidade de proliferação, as distingue das células normais, ou seja, a

proliferação de células cancerosas não esta sujeita aos processos reguladores normais do organismo (RANG; DALE; RITTER, 2001).

A inativação de genes supressores tumorais e a transformação de proto-oncogenes em oncogenes podem conferir autonomia de crescimento a determinada célula, resultando em proliferação descontrolada ao produzir alterações nos seguintes níveis: fatores de crescimento e seus receptores, as vias de fatores de crescimento (transdutores citosólicos e nucleares), reguladores positivos do ciclo celular (ciclina e quinases dependentes de ciclina), reguladores negativos do ciclo celular (p53, Rb e inibidores das quinases dependentes de ciclina), mecanismos de apoptose (morte celular programada), expressão da telomerase e em vasos sanguíneos locais (RANG; DALE; RITTER, 2001).

O processo de carcinogênese (Figura 1), ou seja, de formação de câncer, em geral, dá-se lentamente, podendo levar vários anos para que uma célula cancerosa origine um tumor detectável (ALMEIDA et al., 2005). Esse processo passa por vários estágios antes de chegar ao tumor:

- Estágio de iniciação: É o primeiro estágio da carcinogênese. Nele as células sofrem o efeito de um agente oncoiniciador, que provoca modificações em alguns de seus genes. Nesta fase, as células encontram-se geneticamente alteradas, porém ainda não é possível se detectar um tumor clinicamente. Alguns exemplos de substâncias químicas carcinogênicas são: sulfato de dimetila, metilnitrossurêia, cloreto de vinila, aflatoxinas, dimetilnitrosoamina, benzopireno, dentre outras.

- Estágio de promoção: As células geneticamente alteradas sofrem o efeito dos agentes cancerígenos classificados como oncopromotores. A célula iniciada é transformada em célula maligna, de forma lenta e gradual. Para que ocorra essa transformação, é necessário um longo e continuado contato com o agente cancerígeno promotor. A suspensão do contato muitas vezes interrompe o processo nesse estágio.

- Estágio de progressão: É o terceiro e último estágio, e caracteriza-se pela multiplicação descontrolada, sendo este um processo irreversível. O câncer já está instalado, evoluindo até o surgimento das primeiras manifestações clínicas da doença.

Os fatores que promovem a iniciação ou progressão da carcinogênese são chamados de agentes carcinogênicos. O tabaco, por exemplo, é um agente carcinogênico completo, pois possui componentes que atuam nos três estágios da carcinogênese (ALMEIDA et al., 2005).

Existem três tipos principais de tratamento para o câncer: cirurgia, radioterapia e quimioterapia. Mais recentemente tem-se usado a terapia de fotorradiação com derivados hematoporfíricos (HTP) e a imunoterapia, sendo que o objetivo de cada um destes tratamentos é erradicar o câncer, normalmente por meio de terapia combinada, em que é associado mais do que um tipo específico de tratamento (ALMEIDA et al., 2005).

A quimioterapia do câncer apresenta um entrave crítico, pois a células cancerosas e as células normais, por serem tão semelhantes em inúmeros aspectos, dificultam a identificação de diferenças bioquímicas gerais e exploráveis entre elas (RANG; DALE; RITTER, 2001). O objetivo primário da quimioterapia é destruir as células neoplásicas, preservando as normais. Entretanto, a maioria dos agentes quimioterápicos atua de forma não-específica, lesando tanto células malignas quanto normais, particularmente as células de rápido crescimento, como as gastrointestinais, capilares e as do sistema imunológico. Isto explica a maior parte dos efeitos colaterais da quimioterapia: náuseas, perda de pêlos e susceptibilidade maior às infecções. Porém, o organismo recupera-se destes inconvenientes após o tratamento, e o uso clínico desses fármacos exige que os benefícios sejam confrontados com a toxicidade, na procura de um índice terapêutico favorável (ALMEIDA et al., 2005). Uma das alternativas é o estudo de genes expressos na identificação de alvos moleculares mais relevantes e específicos, que possam ter uma relação mais profícua com as células neoplásicas, diferenciando-as das células normais do organismo.

1.3. Câncer de cabeça e pescoço

O câncer de cabeça e pescoço é um termo associado a um grupo de doenças que acometem os tecidos dessas regiões, cada qual apresentando suas características particulares. Defeitos na base do crânio, indicativos de carcinoma nasofaríngeo (um tipo de câncer dessa natureza), foram descritos no Egito há pelos menos 5000 anos atrás (McGURK; GOODGER, 2000). O câncer de cabeça e pescoço era considerado

incomum há alguns anos, mas dados recentes sugerem que estes números estão crescendo devido ao elevado consumo de álcool e tabaco por parte da humanidade. Em 1998, 6863 casos de câncer de cabeça e pescoço foram relatados na Inglaterra e no País de Gales. Os sítios de desenvolvimento mais comuns destes casos foram a laringe e a cavidade oral (BRADLEY; ZUTSHI; NUTTING, 2005).

O consumo de álcool e tabaco são os dois principais fatores de risco para o desenvolvimento de câncer de cabeça e pescoço. Alguns tipos de vírus ou até de certas inflamações crônicas, também, podem estar envolvidos com o aparecimento deste tipo de câncer, embora não sejam tão evidentes quanto ao consumo exacerbado de tabaco e álcool (GOLDENBERG, 2004).

O tratamento dos tipos de câncer de cabeça e pescoço varia de acordo com o local acometido e o estágio de desenvolvimento dos mesmos, bem como do estado físico do paciente. A excisão cirúrgica do tumor e a radioterapia são as ferramentas mais comumente empregadas nos estágios iniciais. Radioterapia e quimio-radioterapia têm sido ferramentas extensivamente usadas em pacientes que sofrem de recorrência e nos casos mais complicados (MARCU; DOORN; OLVER, 2003).

A quimioterapia utilizada como ferramenta isolada de tratamento, normalmente, não é eficaz, mostrando a necessidade de associação com radioterapia. Os fármacos mais empregados são: cisplatina, doxorrubicina, fluoruracil, vincristina, vimblastina, bleomicina e metotrexato (ALMEIDA et al., 2005; MARCU; DOORN; OLVER, 2003). Estes fármacos são inespecíficos e não conseguem distinguir células tumorais de células normais, apresentando, dessa forma, vários efeitos indesejáveis ao organismo. Isso ocorre pelo fato de não atuarem contra um alvo molecular representativo deste tipo de câncer, o que contribui sobremaneira para a obtenção de resultados terapêuticos pobres.

1.4 Proteína hnRNP K

As proteínas da família hnRNP (do inglês, “heterogeneous nuclear ribonucleoprotein”) foram primeiramente caracterizadas como proteínas que se ligam a transcritos da RNA polimerase II, formando partículas hnRNP. Inicialmente, imaginava-se que o complexo era composto de 6 proteínas, mas investigações subseqüentes identificaram mais proteínas envolvidas. Um total de 19 genes hnRNP já foram identificados. Entretanto, o número total de membros da família permanece sem

determinação precisa, uma vez que, a cada dia, membros menos abundantes do complexo tem sido caracterizados (CARPENTER, 2006).

As proteínas que se ligam ao RNA estão intensamente envolvidas no seu processamento pós-transcricional, constituindo-se em peças chaves no *exon-intron* “splicing”, poliadenilação, exportação nuclear, controle traducional, estabilização/degradação e edição de sua sequência. Em geral, estes fenômenos são dirigidos pela presença de sequências específicas de ácidos nucleicos encontradas no RNA. O recrutamento e a agregação dos multicomponentes que processam os RNAs envolvem o reconhecimento, a nível molecular, destas sequências pelas ribonucleoproteínas (RBPs) [MUSUNURU; DARNELL, 2004].

A especificidade das interações do tipo proteína-RNA apresenta-se como o centro da regulação das atividades celulares. As interações do tipo proteína-RNA desempenham um importante papel na expressão gênica e em outros processos celulares. A diversidade de processos dirigidos ao RNA não poderia ter emergido sem a evolução da seletividade desse tipo de interações. Existem poucos motivos de ligação ao RNA bem descritos, incluindo o domínio RNP (ribonucleoprotein), RGG boxes, zinc fingers e o domínio KH (K homology), embora o repertório de RNAs seja amplo (PAZIEWSKA et al., 2004).

Entre os motivos de ligação ao RNA, que já tenham sido descritos na literatura, o domínio KH é um dos encontrados com maior frequência, presente em inúmeras proteínas. Originalmente identificada na proteína hnRNP K, os domínios KH contêm cerca de 70 aminoácidos que se enovelam em um motivo conservado $\beta\alpha\beta\beta\alpha$, incluindo um “loop” invariável GXXG entre a primeira e a segunda α -hélice, e um “loop” de comprimento variável entre a segunda e terceira folha β (MUSUNURU; DARNELL, 2004).

As proteínas que apresentam domínios KH incluem as proteínas Nova, implicadas na regulação do “splicing” de pré-mRNA; as proteínas hnRNP E e hnRNP K, implicadas, principalmente, na estabilização do mRNA e controle transcricional e traducional; a proteína ZBP-1, envolvida na localização subcelular de mRNA; e a proteína FMRP, envolvida na regulação traducional (MUSUNURU; DARNELL, 2004).

Algumas das proteínas que possuem domínios KH mostram a capacidade de interagir com DNA de fita simples (ssDNA). Entre estas se destacam as proteínas hnRNP K e DDP1. hnRNP K também mostra capacidade de se ligar a DNA de fita

dupla (dsDNA). Estas proteínas apresentam vários domínios KH em sua estrutura, entretanto, não está evidente o papel de vários domínios KH em uma única proteína (MUSUNURU; DARNELL, 2004; BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

A proteína hnRNP K foi descoberta como um componente do conjunto hnRNP, de onde seu nome é derivado. A proteína hnRNP K é codificada por um gene localizado, em humanos, no cromossomo 9. Este gene é responsável pela produção de, pelo menos, cinco proteínas resultantes de transcritos processados de maneira alternativa. Embora a proteína hnRNP K tenha relação estrutural com outras quatro proteínas que se ligam a elementos nucleotídicos ricos em citosina, como hnRNP E1, hnRNP E2, α CP-3 e α CP-4, há apenas um *locus* gênico para hnRNP K humana (GeneID: 3190). A característica mais conservada evolutivamente pela hnRNP K é sua capacidade de se ligar ao RNA através de domínios KH, que está presente em organismos bem distantes evolutivamente, como mamíferos e bactérias. Esta característica conservada reflete um papel fundamental da hnRNP K em processos envolvendo RNA (BOMSZTYK et al., 1997; BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

A proteína hnRNP K apresenta três domínios KH dispostos de maneira assimétrica, em que os domínios KH1 e KH2 estão mais próximos da região N-terminal e o domínio KH3 da região C-terminal.. Estes 3 domínios KH são quase completamente conservados entre *Xenopus laevis* (espécie de sapo) e mamíferos. Domínios KH também são encontrados em proteínas que se ligam a RNA em espécies como: *Escherichia coli* e *Saccharomyces cerevisiae*. O primeiro domínio KH da hnRNP K humana e o domínio KH da hnRNP K de *Halobacterium halobium* apresentam 36% de identidade e 54% de similaridade, sendo maior que a observada entre o primeiro e o segundo domínio KH da própria hnRNP K humana. A sequência mais conservada com o consenso VIGXXGXXI se encontra na região central do domínio estrutural. Uma única substituição de aminoácido (I304N) nesta sequência do consenso afeta as propriedades de ligação da proteína FMR1 com o RNA e causa a mais comum desordem de retardo mental hereditária em humanos, a síndrome do X frágil (PAZIEWSKA et al., 2004; BOMSZTYK et al., 1997; GRISHIN, 2001).

Sidiqi et al. (2005), realizaram um alinhamento conjunto das estruturas dos domínios KH de várias proteínas para observar os resíduos conservados mais

importantes na interação com oligonucleotídeos. Em particular, foi observado que o motivo GXXG, bem como os resíduos da folha $\beta 2$ promoviam a principal superfície de contato. Destes, I20, I21, I28 e I41 mostraram ser resíduos altamente conservados com relação à hidrofobicidade e ao volume, ao passo que G18, G22 e G25 integram o sítio de ligação de oligonucleotídeos. Resíduos de arginina conservados, especialmente R23 e R51, também mostraram estar envolvidos nas interações com oligonucleotídeos. Vale ressaltar que a numeração dos resíduos supracitados corresponde aos do domínio KH3 da proteína α CP1. Dessa forma, os mesmos resíduos conservados de outras proteínas, como a hnRNP K, podem apresentar uma numeração distinta.

A hnRNP K é uma proteína de caráter modular (Figura 1), que apresenta 463 resíduos de aminoácidos. As interações com os nucleotídeos são mediadas pelos seus três domínios KH. Os domínios KH1 (resíduos 32-112) e KH2 (resíduos 142-217) estão localizados na região amino-terminal da proteína, separados por um “linker” de 30 resíduos de aminoácidos, espaço este que é essencialmente o mesmo encontrado entre os domínios KH3 e KH4 da proteína FBP (do inglês, “Fuse Binding Protein”), também envolvida em etapas da expressão gênica. O domínio KH3 (resíduos 389-459) da hnRNP K é isolado dos outros dois e está localizado na região carboxi-terminal da proteína. Os 172 resíduos de aminoácidos que separam os domínios KH2 e KH3 (genericamente conhecido como domínio KI), onde estão localizados outros domínios, como GRGG “box”, SH3 e um domínio de ligação a quinases, estão envolvidos em interações do tipo proteína-proteína com múltiplos parceiros moleculares, dentre eles: outros fatores de transcrição, como *TATA* “binding protein” e vários “zinc fingers”, bem como proteínas envolvidas em diversas vias de transdução de sinais, como tirosina e serina/treonina quinases e a proto-oncoproteína Vav (BRADDOCK et al., 2002).

Figura 1. Ilustração do caráter modular e da disposição dos domínios presentes na proteína hnRNP K. Além dos domínios KH, que se ligam a ácidos nucleicos, pode-se destacar também a presença de domínios responsáveis por interações com outras proteínas, como GRGG “box”, um domínio de ligação a motivos SH3 e um domínio de ligação a proteínas quinases. A isoforma a da proteína apresenta 464 resíduos de aminoácidos enquanto que a isoforma b apresenta 463.

O domínio KI não é encontrado nas outras proteínas que se ligam a elementos ricos em citosina. Este domínio é responsável por muitas das interações conhecidas da hnRNP K com outras proteínas. O domínio KI contém sítios ligantes ricos em prolina, como RXXPXXP e PXXPXR, responsáveis por interações com domínios SH3, como o domínio SH3 da proteína quinase da família Src (BOMSZTYK et al., 1997).

Inúmeros estudos têm sido realizados com o intuito de explorar a ligação ao RNA e/ou DNA de proteínas que contêm domínios KH. A maioria destes estudos foi realizada através de testes *in vitro* (PAZIEWSKA et al., 2004). Dejgaard e Leffers (1996), sugeriram que a ligação da hnRNP K a elementos nucleotídicos ricos em citosina é mediada pelo terceiro domínio KH. Similarmente, Ito, Sato e Endo (1994), evidenciaram que a proteína hnRNP K se liga a fitas simples de DNA ricas em citosina através de sua região carboxi-terminal, exatamente onde se encontra o domínio KH3. Siomi et al. (1994), sugerem que todos os três domínios KH, da proteína hnRNP K, têm um importante papel na ligação a oligonucleotídeos sob condições limitadas (NaCl na concentração de 1M). Mas, os mesmos afirmam que, em condições fisiológicas, nenhuma conclusão poderia ser feita acerca da relativa contribuição de cada domínio KH na ligação a RNA em hnRNP K.

O domínio KH3 tem mostrado se ligar a ácidos nucleicos como um domínio isolado, embora com menor afinidade quando comparado com a proteína na sua forma íntegra (PAZIEWSKA et al., 2004). Estruturas de complexos entre o domínio KH3 da hnRNP K e fitas simples de DNA ou RNA têm sido resolvidas por ressonância magnética nuclear e cristalografia de raios-X. O domínio KH3 da hnRNP K apresenta três folhas β antiparalelas (resíduos 14-21, 45-50 e 58-65) que dão suporte a três α -hélices (resíduos 23-29, 34-42, 67-83), que se encontram no arranjo típico dos domínios

KH (Figura 2A), com a seguinte configuração estrutural: $\beta 1-\alpha 1-\alpha 2-\beta 2-\beta 3-\alpha 3$. A face externa das folhas β antiparalelas é composta de resíduos de aminoácidos hidrofílicos, com exceção do resíduo I60. Já a face interna das folhas β é composta de resíduos de aminoácidos hidrofóbicos, com exceção do resíduo T16 que é acessível ao solvente. As α -hélices anfifílicas se empilham na face hidrofóbica das folhas β , formando um centro de característica hidrofóbica. O motivo invariável GXXG (resíduos 30-33) está localizado em um *loop* curto que conecta as α -hélices 1 e 2, ao passo que o *loop* variável (resíduos 51-57) está localizado entre as folhas β 2 e 3. Um resíduo de glicina altamente conservado, G44, localizado no final da α -hélice 2 permite estericamente a formação de um “turn” entre as α -hélices 2 e 3. O resíduo G65 no final da folha β 3 também é altamente conservado e tem papel similar entre a folha β 3 e a α -hélice 3 (BACKE et al., 2005; BABER, et al., 1999).

Figura 2. (A) Arranjo estrutural típico de um domínio KH. (B) Representação do domínio KH3 da proteína hnRNP K em complexo com o oligonucleotídeo de ssDNA TCCCT (T1, C2, C3, C4, T5). (C)

Representação da superfície de potencial eletrostático do domínio KH3 formando uma fenda com um centro hidrofóbico que acomoda o oligonucleotídeo TCCCT.

Alguns estudos mostram que a cadeia polipeptídica do domínio KH3 forma uma fenda estreita e alongada, localizada na superfície da proteína, responsável pelo reconhecimento específico de oligonucleotídeos ricos em citosina (Figuras 2B e 2C), como 5'd-TATTCCCT, 5'd-CTCCCC e 5'd-TTCCCCTCCCCATTT. Os oligonucleotídeos de ssDNA se localizam nesta fenda, que apresenta caráter predominantemente hidrofóbico (resíduos I29, I36, I47, I49), justamente entre o motivo invariável GXXG e o loop variável. Através de estudos de modelagem molecular foi revelado que a ligação do domínio KH3 com cognatos de RNA apresentam uma configuração similar. O sítio de reconhecimento molecular dos oligonucleotídeos mostrou que esta fenda está localizada na superfície dos domínios KH, e por ser relativamente estreita favorece a ligação de oligonucleotídeos ricos em bases nitrogenadas pirimidínicas, dificultando o acesso de oligonucleotídeos ricos em purinas. A fenda é cercada, principalmente, de resíduos de aminoácidos com carga positiva (K31, K37, R40, K48 e R59), sendo que as extremidades contêm outros resíduos hidrofílicos que apresentam grupos hidroxila ou carboxilato (S27, S46 e E51) [BRADDOCK et al., 2002; BACKE et al., 2005; BABER, et al., 1999].

Braddock et al. (2002), que solucionaram a estrutura do complexo entre o domínio KH3 da hnRNP K e o oligonucleotídeo 5'd-TATTCCCT por ressonância magnética nuclear, mostraram que apenas a tétrade TCCC interage com o domínio KH3. Os primeiros dois nucleotídeos da tétrade (TC) mostraram-se capazes de interagir com resíduos de aminoácidos encontrados na α -hélice 1. Logo, foram identificados resíduos de aminoácidos importantes para a ligação com a tétrade oligonucleotídica, destacando-se G26, I29, K31, G32, I36, K37, R40, I49 e R59. As interações entre as bases nitrogenadas e os resíduos de aminoácidos são caracterizadas por uma extensa rede de ligações de hidrogênio, algumas das quais entre os grupamentos metila dos aminoácidos e os átomos de oxigênio e nitrogênio das bases nitrogenadas. Embora de natureza fraca, as ligações de hidrogênio CH---O têm sido observadas com frequência entre proteínas e complexos proteína-DNA. Estas interações CH---O são suplementadas por ligações de hidrogênio típicas entre os grupamentos amida da cadeia principal das proteínas e os átomos de oxigênio das bases nitrogenadas. Estes grupamentos amida das

proteínas também mostraram ser importantes nas interações eletrostáticas com grupamentos fosfatos dos oligonucleotídeos.

Backe et al. (2005), realizaram estudos de ressonância magnética nuclear e cristalografia de raios-X para identificar a estrutura do complexo hnRNP K KH3-ssDNA, e constataram que o domínio KH3 da proteína hnRNP K se liga, especificamente, a seqüências oligonucleotídicas que possuem a tétrade TCCC ou CCCC. Os nucleotídeos da seqüência central, TCCC ou CCCC, em conjunto com moléculas de água, encontram-se envolvidos em uma densa rede de interações, em que se destacam fortes ligações de hidrogênio e interações hidrofóbicas. A região que compreende os resíduos de aminoácidos 26-33, incluindo o motivo invariável GXXG (resíduos 30-33), é considerada crítica para a definição da conformação do DNA e permissão do reconhecimento específico. A região carboxi-terminal da α -hélice 1, de forma especial os resíduos G26 e I29, se comporta estruturalmente como uma cunha, impedindo o empilhamento das bases TCCC. Foram identificados, também, outros resíduos de aminoácidos importantes para o reconhecimento das bases nitrogenadas das seqüências TCCC e CCCC, como S27, G30, G32, G33, R40, E51, R59, Y75, S80, além da interação entre os resíduos K31 e K37 com os grupamentos fosfatos das seqüências oligonucleotídicas.

A função da proteína hnRNP K no complexo hnRNP (*heterogeneous ribonucleoprotein*) ainda não está bem definida. Sabe-se que a proteína hnRNP K é facilmente obtida de extratos nucleares e citoplasmáticos, o que indica uma ampla distribuição intracelular. A hnRNP K se liga a seqüências específicas de RNA, bem como de ssDNA e dsDNA. Seqüências de RNA ricas em grupamentos de citosina, normalmente, se ligam fortemente aos domínios KH da hnRNP K, o que não ocorre com outros homopolímeros de RNA que interagem sutilmente ou simplesmente não interagem com os domínios KH. A afinidade de ligação ao RNA é diminuída quando a proteína se encontra fosforilada (BOMSTYK et al., 1997).

A proteína hnRNP K interage com diversos parceiros moleculares protéicos (Tabela 1). Para a maioria das interações do tipo proteína-proteína, a relevância do ponto de vista funcional ainda é enigmática. Apesar de tudo, a diversidade das interações da hnRNP K supõe que ela esteja envolvida em múltiplos processos que compõem a expressão gênica, como: remodelagem de cromatina, transcrição, *splicing*, tradução e estabilização do mRNA. O envolvimento da hnRNP K no processo de

expressão gênica tem sido demonstrado em muitos estudos. Estes experimentos também demonstram que a proteína hnRNP K pode ativar ou reprimir a expressão gênica (tabela 2) [BOMSZTYK; DENISENKO; OSTROWSKI, 2004].

Tabela 1. Parceiros moleculares da proteína hnRNP K nos diversos processos de expressão gênica e na transdução de sinais.

Processo	Parceiros moleculares
Transdução de sinais	Tirosina quinases: Src, Lyn, Fyn, Lck, Itk Serina-treonina quinases: PKC, ERk1/2, JNK Arginina metiltransferases: PRMT1 Fator de permuta de nucleotídeos: Vav
Expressão gênica: Remodelagem de cromatina	Eed DNA-metiltransferase SAF-B
Expressão gênica: Transcrição	Fatores gerais: TBP, HMGB1 Ativadores: Pura, Sox 10, C/EBP β Repressores: Zik1, Kid1, MZF1
Expressão gênica: “Splicing”	hnRNP: E2, I, K, L, U Fatores de <i>splicing</i> : 9G8, SRp20 Helicase: DDX1 Fatores gerais: YB-1, Sam68
Expressão gênica: Tradução	Elongação: EF-1 α

Tabela 2. Exemplos do envolvimento da proteína hnRNP K de mamíferos em múltiplos processos de expressão gênica.

Processo	Gene
Transcrição: Ativação	c-Myc c-Src
Transcrição: Repressão	Timidina quinase Subunidade $\beta 4$ do receptor de acetilcolina
“Splicing”	B-tropomiosina
Estabilidade do RNA	Renina
Tradução: Ativação	c-Myc
Tradução: Silenciamento	15-lipoxigenase (LOX) Papilomavírus tipo 16 (HPV-16)

A proteína hnRNP K interage *in vivo* e *in vitro* com o fator de remodelamento de cromatina Eed. O fator Eed existe na forma de um complexo com Ezh2, uma metiltransferase de histona (HMT), metilando os resíduos de aminoácidos de H3 a K9 e K27. A relevância funcional destas interações ainda permanece sem um conhecimento profundo. A matriz nuclear é uma estrutura dinâmica implicada na organização de cromatina, replicação de DNA, transcrição e processamento de RNA (“splicing”). A hnRNP K é um componente da matriz nuclear e mostra capacidade de interagir com SAF-B (“scaffold attachment factor-B”), outro componente da matriz nuclear (BOMSZTYK; DENISENKO; OSTROWSKI, 2004). Estes achados sugerem que a proteína hnRNP K apresenta funções relacionadas à cromatina e a matriz nuclear.

Interações específicas da proteína hnRNP K com motivos distintos de DNA já foram observadas, como o elemento CT. Por exemplo, a proteína hnRNP K se liga a seqüências de homopirimidina (CCCC) presentes no elemento CT, que se encontra no promotor *c-myc* P1. O elemento CT corresponde a 4 repetições imperfeitas da seqüência 5’d-CCCTCCCCA de 9 pares de bases. A hiperexpressão de hnRNP K aumenta a atividade do promotor do gene *c-myc*, efeito este que é estimulado quando a hnRNP K se apresenta co-expressa com a proteína TBP (TATA “box-binding protein”). *In vivo*, a

proteína hnRNP K existe em complexo com TBP, e mostra interagir com a mesma *in vitro*. Logo, a indução do gene *c-myc* pode depender da interação de hnRNP K com TBP (BRADDOCK et al., 2002; BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

A proteína hnRNP K se liga e ativa o promotor *c-src* em cooperação com o fator de transcrição Sp1. Em contraste com a sua atuação nos genes *c-myc* e *c-src*, a proteína hnRNP K reprime o promotor do gene da timidina quinase através do elemento CT. A respeito de repressão transcricional mediada pela hnRNP K, pode-se destacar ainda a interação da hnRNP K com o repressor transcricional Zik1. Dessa forma, fica evidente que a hnRNP K apresenta um papel pleiotrópico na transcrição, refletindo sua associação, de caráter heterogêneo, em complexos ativadores e repressores transcricionais (BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

Os mecanismos de ação da proteína hnRNP K são melhores conhecidos no processo de tradução. Uma das primeiras pistas de que a proteína hnRNP K poderia estar envolvida no processo de tradução tem origem na observação de que a hnRNP K se liga ao fator EF-1 α (“translation elongation factor” - 1 α). Subseqüentemente, vários estudos promoveram maiores esclarecimentos a respeito do papel da hnRNP K na regulação da tradução. Como um exemplo clássico da sua influencia no processo de tradução, a proteína hnRNP K atua no citoplasma reprimindo a maturação de eritrócitos, silenciando a tradução da 15-lipoxigenase (LOX) por se ligar ao elemento DICE (“differentiation control element”), que constitui-se em uma seqüência de repetições ricas em bases nitrogenadas CU encontrada na região 3’ UTR do mRNA da LOX. O silenciamento ocorre na iniciação da tradução, em que a hnRNP K em conjunto com hnRNP E1/2 estão ligadas ao elemento 3’ UTR DICE bloqueando o recrutamento da subunidade ribossômica 60S e a conseqüente formação do componente traducional ribossômico 80S. Entre o segundo e o terceiro domínios KH da hnRNP K encontra-se um cluster de ligação a motivos SH3 (Figura 1). A proteína hnRNP K se liga seletivamente a estes motivos SH3 das tirosina quinases, como: Src, Fyn, Lyn e Lck. A fosforilação de resíduos de tirosina da hnRNP K, principalmente por Src, mostram-se capazes de diminuir a afinidade da mesma por ácidos nucléicos *in vitro* e de reprimir o silenciamento do mRNA da LOX. Estas observações sugerem que a família Src de proteína quinases é um fator responsável pelo controle traducional dependente da hnRNP K em resposta a sinais extracelulares (BOMSZTYK; DENISENKO; OSTROWSKI, 2004; OSTARECK et al., 1997; BACKE et al., 2005).

Outra maneira pela qual a hnRNP K pode regular o processo de tradução é a fosforilação de seus resíduos S284 e S353 mediada pela ERK, que promove um acúmulo citoplasmático da proteína hnRNP K, fato responsável por um aumento da repressão da tradução do mRNA da LOX. Em contraste ao silenciamento da tradução da LOX, a hnRNP K, em conjunto com hnRNP E1/2, estimula a tradução do mRNA do gene c-myc. Logo, como na transcrição, os efeitos da proteína hnRNP K no processo de tradução são pleiotrópicos (BOMSZTYK; DENISENKO; OSTROWSKI, 2004; BACKE et al., 2005).

Recentemente, Huth et al. (2004) realizaram um “screening” virtual em bases de dados em busca de ligantes da proteína FBP (“FUSE Binding Protein”). Essa proteína também liga DNA através de domínios KH. Dos compostos planejados *in silico* para inibir o sítio ligante de DNA da FBP, pelo menos 5 novos protótipos foram selecionados e testados com a proteína. Para avaliar a especificidade do inibidor os autores usaram a proteína hnRNP K. Os resultados dos ensaios da proteína com DNA, em presença e ausência de ligantes competitivos, demonstraram uma atividade média dos compostos na faixa micromolar de IC_{50} e K_D para FBP. Tomonaga e Levens (1995) descrevem um método de ensaio da proteína hnRNP K com oligonucleotídeos, em presença ou não de ligantes. Esses e outros recentes estudos justificam a hnRNP K, que é superexpressa em células tumorais de câncer de cabeça e pescoço (LEOPOLDINO et al., 2007) e para a qual ainda não existem ligantes específicos descritos, como um atrativo alvo terapêutico em câncer. Análises preliminares com respeito à alta identidade sequencial da hnRNP K com outra homóloga contendo estrutura resolvida (domínio KH3, códigos PDB 1ZZI e 1J5K) justificam a construção do referido modelo e seu uso para planejamento de ligantes.

1.5. Modelo para atuação da proteína hnRNP K

O envolvimento da proteína hnRNP K de mamíferos em múltiplos processos celulares sugere que há um grande número de mecanismos de expressão gênica que se utilizam da hnRNP K para integração de sinais. Vários estudos sugerem um modelo, no qual a hnRNP K atua como uma “plataforma de ancoragem”, permitindo que a proteína Lck, da família Src de proteínas quinases, interaja com um membro de outra cascata de quinases, a proteína PKC, para controlar o fator de tradução EF-1 α (Figura 3). Em

resposta a um sinal extracelular, Lck fosforila resíduos de tirosina da hnRNP K gerando sítios de interação SH2, que em conjunto com os sítios SH3, recrutam Lck para hnRNP K. A enzima PKC é induzida por outro sinal extracelular, através da formação de diacilglicerol. Enquanto permanece ligada ao RNA, as interações diretas de hnRNP K com a enzima PKC permanecem impossibilitadas. A ligação da enzima Lck a hnRNP K aumenta sua atividade, resultando em fosforilação adicional de resíduos de tirosina da hnRNP K, causando a dissociação da mesma ao RNA e permitindo a interação com a enzima PKC. Depois que PKC se encontra ligada a hnRNP K a atividade da mesma é induzida por fosforilação de seus resíduos de tirosina mediada pela Lck que também se encontra ligada a hnRNP K. A enzima PKC ativada não só fosforila o resíduo S302 da hnRNP K como também fosforila resíduos de outros efetores que estejam ligados à hnRNP K ou no microambiente à sua volta. Por exemplo, EF-1 α se liga a hnRNP K, além de ser um substrato da enzima PKC. A fosforilação de EF-1 α mediada pela PKC poderia ocorrer com a integração da proteína hnRNP K. A fosforilação de EF-1 α poderia contribuir para a ativação do sistema de tradução. A defosforilação mediada pelas tirosinas fosfatases permite que a hnRNP K retome suas interações com o RNA. Logo, a retomada das interações da hnRNP K com o RNA desloca a proteína PKC e retorna o sistema de tradução ao estado inicial. Este cenário ilustra como a proteína hnRNP K poderia integrar duas vias, uma da cascata da Lck e outra da cascata da PKC (BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

Este modelo apresentado não incluiu a contribuição da cascata da proteína MAPK, que também tem como alvo a proteína hnRNP K. Embora alguns dos passos apresentados não estejam plenamente confirmados, este modelo ilustra a natureza dinâmica dos processos que envolvem a hnRNP K e proteínas similares que se ligam a DNA/RNA, bem como as informações conduzidas pelas vias das tirosinas quinases e PKC, iniciadas por dois estímulos diferentes e integradas pela hnRNP K para gerar uma resposta específica (BOMSZTYK; DENISENKO; OSTROWSKI, 2004).

Figura 3. Modelo de atuação da proteína hnRNP K, funcionando como uma plataforma para integrar sinais das cascatas de quinases para um sítio de processos dirigidos ao RNA.

Vários experimentos ainda sugerem que a proteína hnRNP K desempenhe importante papel como um centro de alojamento para moléculas envolvidas em processos que regulam a transcrição (BOMSZTYK; DENISENKO; OSTROWSKI, 2004). Por exemplo, o fator de transcrição TBP é fosforilado pela ERK1/2 e outras quinases. A proteína hnRNP K interage com ERK1/2, bem como com o fator TBP. Logo, é concebível que hnRNP K possa promover um ambiente favorável para a regulação da fosforilação do fator TBP mediada pela ERK1/2.

1.6. Família de proteínas hnRNP e câncer

Um oncogene pode ser definido como um gene capaz de causar a transformação de células normais em células neoplásicas ou cancerosas. Baseado nesta definição, alguns membros da família hnRNP podem ser considerados oncogenes (CARPENTER et al., 2006).

A proteína hnRNP P2 é uma proteína multifuncional, responsável pela transcrição, “splicing” e transporte do mRNA do núcleo para o citoplasma. Em 90% dos

casos de lipossarcoma mixóide em humanos, translocações gênicas levam à criação de um gene fusionado (hnRNP P2-CHOP), que codifica uma proteína contendo o domínio de ativação transcricional (região N-terminal) da proteína hnRNP P2 e o domínio de ligação ao DNA da proteína CHOP. A proteína CHOP está funcionalmente implicada na eritropoiese (formação de eritócitos), diferenciação de adipócitos, interrupção do crescimento celular e na progressão do ciclo celular da fase G1 para S, funcionando como repressor ou ativador transcricional. A injeção de linhagens celulares com hiperexpressão induzida de hnRNP P2-CHOP em camundongos resulta na formação de tumores. Logo, a fusão dos genes que codificam o domínio de ativação da hnRNP P2 e o domínio de ligação ao DNA da CHOP leva à criação de um potente oncogene (CARPENTER et al., 2006).

O gene hnRNP P2 também está envolvido em outra translocação, a qual resulta em outro tipo de câncer. Nesta translocação, um evento comum em leucemia mielóide, a região de ligação ao DNA da proteína ERG se funde ao domínio N-terminal da hnRNP P2. A hiperexpressão da proteína fusionada hnRNP P2-ERG em linhagens celulares de fibroblastos de camundongos induz a proliferação independente das células, mostrando que pelo menos em culturas celulares a proteína hnRNP P2-ERG é capaz de promover divisão celular. Há evidências de que a atividade da proteína hnRNP P2-ERG seja responsável pela patogenia da leucemia mielóide aguda (CARPENTER et al., 2006).

A proliferação celular é uma importante etapa no desenvolvimento tumoral, sendo que o ciclo celular é regulado por proteínas do gene *c-myc*. Logo, um mecanismo pelo qual as proteínas da classe hnRNP poderiam regular a progressão tumoral é o controle da expressão dessas proteínas. Muitas das hnRNPs têm mostrado regular a expressão do gene *c-myc*, como a hnRNP K que aumenta a transcrição deste gene, assim como a proteína hnRNP C, que ativa a tradução do mRNA *c-myc*. A hnRNP K também se mostra hábil na ativação transcricional do oncogene *c-Src* (CARPENTER et al., 2006).

O fator de iniciação da tradução eIF4E desempenha importante papel na proliferação celular, e sua hiperexpressão ocorre em vários tumores malignos, incluindo em câncer de cabeça e pescoço, mama, cólon, pulmão e vesícula biliar. A proteína hnRNP K tem se mostrado capaz de ligar a um promotor no gene eIF4E, resultando na ativação da transcrição do mesmo. A transcrição do fator eIF4E apresenta-se aumentada

se ambos, *c-myc* e hnRNP K, são co-hiperexpressos, sugerindo que a proteína hnRNP K pode cooperar com outras oncoproteínas para induzir expressão de genes envolvidos no crescimento de células cancerosas (CARPENTER et al., 2006).

A importância das hnRNPs no desenvolvimento tumoral evidencia-se também pelo fato de que o genoma do adenovírus codifica uma proteína (E1B-AP5) relacionada a esta família, possuindo homologia com hnRNP U, a qual possui propriedades para promover o crescimento de células cancerosas. A proteína E1B-AP5 interage com o fator de supressão tumoral p53, sendo que a transcrição de fatores dependentes da proteína p53 é inibida por tal interação. Logo, a proteína E1B-AP5, considerada um membro da família hnRNP, pode promover o desenvolvimento de câncer por se ligar e inibir o fator de supressão tumoral p53 (CARPENTER et al., 2006).

Os defeitos na morte celular programada (apoptose) consistem em um importante mecanismo no desenvolvimento do câncer. Os oncogenes como *c-myc* e E1A que promovem a divisão celular, também se mostram hábeis em inibir o processo de apoptose. A inibição do processo de apoptose facilita a sobrevivência de células geneticamente instáveis, permitindo a seleção de células com características agressivas. Um potente inibidor da apoptose é a proteína XIAP. Em resposta ao estresse celular, a tradução do mRNA do gene XIAP mediada pelo IRES (“internal ribosomal entry site”, que é uma estrutura especializada em recrutar o ribossomo ao mRNA) é estimulada pela ligação de um complexo RNP à estrutura IRES. A proteína hnRNP C1/C2 forma parte deste complexo e os níveis celulares deste membro hnRNP aumentam paralelamente à atividade do XIAP IRES em culturas celulares. Logo, a hiperexpressão de hnRNP C1/C2 aumenta da tradução do mRNA do gene XIAP, indicativo de que hnRNP C1/C2 controla os níveis de expressão celular da proteína XIAP, que é um inibidor do processo de apoptose (CARPENTER et al., 2006).

Outra proteína envolvida no controle da apoptose é a Bcl-x. O transcrito primário do gene *Bcl-x* sofre splicing alternativo e gera duas variantes, Bcl-xs e Bcl-xl. Bcl-xs promove a apoptose, ao passo que Bcl-xl inibe o processo de apoptose. As proteínas hnRNP F e hnRNP H se ligam a uma região no *exon 2*, fato que modula a seleção de Bcl-x 5'. Dessa forma, estes membros da família hnRNP favorecem o “splicing” para o regulador pró-apoptótico Bcl-xs. Logo, várias hnRNPs podem atuar como reguladores positivos ou negativos da apoptose, hnRNP C1/C2 é considerado um

inibidor, ao passo que hnRNP F e hnRNP H funcionam como ativadores do processo (CARPENTER et al., 2006).

As células neoplásicas frequentemente migram do sítio de crescimento inicial do tumor maligno para outros tecidos do organismo, geralmente transportadas pelo sangue ou sistema linfático. Este processo é conhecido como metástase. A modulação do processo de adesão celular é um importante estágio nos eventos de metástase. Para que as células neoplásicas se livrem do seu tecido original para iniciar o processo de metástase, os complexos de adesão devem ser modulados ou destruídos. Estudos que utilizaram linhagens celulares de fibroblastos de pulmão mostraram que os centros de iniciação da disseminação são compostos por várias proteínas, entre elas hnRNP P2, hnRNP K e hnRNP E1. Logo, estas observações promovem evidências de uma conexão entre as hnRNPs e o estágio de iniciação da proliferação celular na metástase (CARPENTER et al., 2006).

A angiogênese (formação de vasos sanguíneos) é uma função celular através da qual as células de tumores malignos sólidos recrutam seu próprio suprimento de sangue. Sem um suprimento sanguíneo adequado o volume dos tumores sólidos é incapaz de aumentar. Os fatores que normalmente estimulam a angiogênese são: o fator de crescimento endotelial vascular (VEGF) e o fator de crescimento de fibroblastos (FGF). Algumas proteínas da família hnRNP têm mostrado regular estes fatores de angiogênese. A proteína hnRNP L interage especificamente com uma região do mRNA do fator VEGF, e esta interação entre proteína-RNA ocorre somente em células que sofrem de hipóxia, fato que acentua a expressão do fator VEGF. Já a proteína hnRNP A1 tem mostrado aumentar a tradução do fator FGF, deixando evidente outra conexão entre as hnRNPs e o processo de angiogênese (CARPENTER et al., 2006).

Comparado tecidos de cólon intestinal saudáveis com tecidos acometidos por câncer colorretal, Carpenter et al (2006) encontraram altos níveis de expressão da proteína hnRNP K nos tecidos com câncer. O envolvimento da proteína hnRNP K também tem sido descrito em outras manifestações clínicas do câncer, como em câncer de pulmão (PINO et al., 2006) e câncer de fígado (OSTROWSKI; BOMSZTYK, 2003). Elevados níveis de expressão da proteína hnRNP K também já foram relatados em câncer de mama, onde a elevada expressão de hnRNP K contribui no aumento da expressão do gene *c-myc*, que por sua vez desencadeia o desenvolvimento deste tipo de câncer (OSTROWSKI; BOMSZTYK, 2003).

1.7. Planejamento racional de fármacos *in silico*

Atualmente, o planejamento racional baseado em estrutura e no mecanismo de ação é a estratégia mais eficiente no desenvolvimento de novos fármacos, capaz de contribuir em todos os estágios do processo, desde a descoberta de protótipos (também conhecidos como “compostos de partida” ou “*lead compounds*”), sua otimização (com respeito à afinidade, especificidade, eficácia e toxicidade), até a elaboração de compostos candidatos a testes clínicos. Esta estratégia é baseada no bloqueio ou estimulação da atividade biológica de macromoléculas, tais como proteínas ou ácidos nucleicos (DNA ou RNA), associadas a diferentes processos patológicos. A informação estrutural do bioreceptor e/ou ligantes permite a descoberta e síntese de compostos com complementaridade estérica, hidrofóbica e eletrostática ao seu sítio de ligação, os quais podem vir a se tornar fármacos. Essa abordagem, em sua essência, caracteriza o planejamento racional de fármacos baseado em estrutura. O que ainda a torna mais atrativa, quando utilizada em proteínas, é o conhecimento de que 78% dos fármacos atuais têm como alvo receptor esse tipo de biomacromolécula (MARSHALL, 2004).

Desde a concepção do alvo biológico até a descoberta de um novo fármaco, um processo que pode levar em média 11 anos ou até mais, a bioinformática, juntamente com a química computacional, vem oferecendo um excelente direcionamento no planejamento racional de fármacos, já com inúmeros casos de sucesso envolvendo o emprego de simulações computacionais (MARSHALL, 2004), citando como exemplo os importantes fármacos: losartan, atorvastatina e celecoxib. Para Manuel Peitsch, da “Novartis Institutes for BioMedical Research”, o processo de descoberta e desenvolvimento de novos fármacos é hoje totalmente dependente da utilização métodos computacionais (PEITSCH, 2004).

A convergência de tecnologias genômicas e o desenvolvimento de fármacos planejados contra alvos moleculares específicos provêm muitas oportunidades para o uso da bioinformática, com a finalidade de se diminuir o “gap” entre conhecimento biológico e terapia clínica. Isso pode ser alcançado, por exemplo, identificando genes que têm propriedades similares a conhecidos alvos, investigando similaridade *pairwise* entre bibliotecas (ou “pool” de bibliotecas) de diferentes origens, tais como as de células normais e as de células tumorais e, ainda, construindo modelos dos alvos receptores

baseados em homologia seqüencial e similaridade estrutural (DESANY; ZHANG, 2004).

Por outro lado, em química computacional, genericamente citada como modelagem molecular, destaca-se o “docking” molecular como um dos métodos mais empregados. Com esse método, são investigadas as possíveis orientações que determinada molécula assume no interior do sítio ligante de um bioreceptor, ou simplesmente entre duas macromoléculas, tal como é o caso da interação entre proteína-proteína ou proteína-DNA, caracterizando o “docking” macromolecular (INSIGHT II USER GUIDE, 2005). Os métodos de “docking”, em geral, envolvem uma função de energia contendo parâmetros eletrostáticos, de van der Waals, de ligações de hidrogênio e, algumas vezes, hidrofóbicos, os quais geram modelos matemáticos que predizem as melhores orientações do ligante, segundo uma lista de escores de energia. As mais recentes versões dos dois programas de maior sucesso em “docking”, **FlexE** e **GOLD**, consideram a flexibilidade do ligante e também de algumas cadeias laterais do sítio receptor. Os mais recentes e promissores métodos de “docking” utilizam informação farmacofórica do sítio receptor para guiar as simulações, tais como o **FlexX-Pharm**. A partir dessa estratégia é possível selecionar por “screening” virtual compostos de bases de dados contendo tipicamente milhares de estruturas, eliminando compostos não promissores antes que eles sejam sintetizados (ALONSO; BLIZNYUK; GREASY et al., 2006).

Métodos de “screening” virtual vêm sendo amplamente empregados na seleção de novos protótipos nos últimos anos. Diversos casos de sucesso com o uso dessa sistemática, tal como a descoberta de isoflavonóides como inibidores não-esteroidais da 5 α -redutase, utilizando “constraint” farmacofórica (BRENK et al., 2003; CHEN et al., 2001). Nessa era pós-genômica, o “screening” virtual complementa os conhecidos métodos experimentais de “screening” em larga escala no processo de descoberta de novos protótipos (KLEBE, 2006). Porém, o sucesso do “screening” dito *in silico*, e em geral das técnicas de “docking”, depende do conhecimento de detalhes estruturais finos do sítio de reconhecimento da biomacromolécula (CARLSON; MASUKAWA; MACCAMMON, 1999).

Em uma outra categoria de “docking”, a qual não envolve “softwares” que realizam “screening” virtual, ou mesmo que tentam somente predizer a orientação de um ligante no interior de um sítio receptor biológico (“docking single”), encontram-se

os métodos conhecidos como “docking build” ou “docking” *de novo*. Dentre os “softwares” que empregam tal método, destaca-se o **LUDI-CAP** (INSIGHT II USER GUIDE, 2005). Inicialmente, o sítio receptor é caracterizado no tocante à sua capacidade de ligar moléculas, utilizando-se para isso grupos funcionais específicos, selecionados pelo “software” a partir de sua própria base de dados. Esses grupamentos servem como “sonda” para a busca, nessa mesma base de dados, por fragmentos que possam interagir satisfatoriamente com os aminoácidos do sítio receptor, gerando uma nova molécula ligante. Esse método é considerado o pioneiro para a otimização *in silico* de protótipos (MARSHALL, 2004). Alguns casos de sucesso envolvendo o uso dessa tecnologia têm sido reportados, tal como o planejamento validado de 10 novos inibidores da Transcriptase Reversa de HIV-1, na faixa micromolar de IC₅₀ (SCHENEIDER; FECHNER, 2005).

A estratégia baseada na hipótese do farmacóforo é a do “análogo ativo” (MARSHALL, 2004). O farmacóforo representa o conjunto de domínios funcionais das moléculas ligantes através dos quais se define os tipos de interação que os ligantes em comum fazem com o sítio receptor. A análise, por métodos computacionais, dos possíveis conjuntos de grupos farmacofóricos associados a cada molécula ativa, permite a derivação do padrão farmacofórico comum ao conjunto de análogos ativos em questão. Dentre os métodos mais robustos e eficientes que envolvem esse tipo de cálculo, destacam-se **DiscoTech** e **GALAHAD** (SYBYL USER GUIDE, 2005).

Um diferente e robusto método de planejamento racional de fármacos, agora direcionado à otimização *in silico* de protótipos, consiste em investigar as condições energéticas entre moléculas as quais se aproximam uma da outra, gerando os campos de interação molecular (“Molecular Interaction Field” - MIF). Os MIFs descrevem a variação da energia de interação entre uma molécula alvo e um grupo químico de prova que se move confinado ao interior de um “grid” 3D, o qual é posicionado de modo a mapear a região de interesse do alvo molecular (o sítio ligante). As diferentes provas que usualmente são testadas refletem as características químicas que deveriam possuir o ligante ideal ou fragmentos de sua estrutura. Os “softwares” de uso mais freqüente que empregam tal método são o **GRID** (GOODFORD, 1985), o **VolSurf** e o **Almond** (SYBYL USER GUIDE, 2005).

O “software” **VolSurf**, adicionalmente, transforma os campos moleculares de seus grupos de prova em descritores, os quais estão associados com as principais forças

de interação entre ligante e receptor, correlacionando-os espacialmente à atividade biológica. Sua sistemática é similar àquela empregada em um dos métodos preditivos de maior sucesso em estudos que relacionam quantitativamente estrutura com atividade (QSAR): o CoMFA (“**C**omparative **M**olecular **F**ield **A**nalysis”), o qual também é utilizado para a otimização de protótipos. Casos de sucesso em planejamento racional com **VolSurf** têm sido relatados na literatura, como o desenvolvimento de potentes inibidores da metaloprotease MMP-8 (CRUCIANI et al., 2003).

Uma das mais recentes e promissoras tecnologias de ponta empregada na otimização *in silico* de protótipos é o método **RACHEL** (“Real-time Automated Combinatorial Heuristic Enhancement of Lead compounds”), implementado no “software” **SYBYL**. **RACHEL** foi especificamente projetado para otimizar compostos com baixa afinidade pelo sítio receptor, e assim o faz utilizando um método combinatório automatizado. O sistema **RACHEL** utiliza uma base de dados de fragmentos químicos para derivar o protótipo com o intuito de substituir regiões de baixa afinidade pelo sítio ativo por componentes químicos que poderiam aumentar o nível de complementaridade ao mesmo. O também recente **EA-Inventor** se vale de um método de *de novo* “design”, em que novos compostos podem ser otimizados a partir de um “scaffold” estrutural básico através da adição de novos grupamentos R, ou até novos “scaffolds” podem ser gerados e explorados na criação de novos protótipos (SYBYL USER GUIDE, 2005).

Para o desenvolvimento de um novo fármaco, também já é possível estimar propriedades farmacocinéticas, bem como propriedades “drug-like” ou “lead-like” de diferentes compostos, selecionando, durante as diversas etapas da modelagem, somente compostos com potencial de se tornarem fármacos. Como exemplo mais simples, a ‘Regra dos 5’ (RO5), de Lipinski, preconiza que os fármacos que apresentam biodisponibilidade por via oral, em geral, seguem, a saber: peso molecular menor ou igual a 500, log *P* menor ou igual a 5, número de grupos doadores de ligações de hidrogênio menor ou igual a 5 e número de grupos aceptores de ligações de hidrogênio menor ou igual a 10 (LIPINSKI et al., 1997). Citações em *CAS SciFinder* do artigo original da RO5, de 1997, excederam 1000 somente no ano de 2004, e continuam crescendo (LIPINSKI, 2004). Uma variação dessa regra é aplicável a protótipos e, além disso, novas regras empíricas vêm sendo descritas, tais como o número de ligações

rotacionáveis em fármacos ser menor do que 8 ou, ainda, a área superficial ser menor ou igual a $140 \text{ \AA}^2$ (MARSHALL, 2004).

Propriedades tais como absorção, distribuição, metabolismo, excreção e toxicidade (ADMET) podem ser preditas, além da utilização de métodos estatísticos, a partir de “screening” em bases de dados contendo essas informações, as quais são computadas para uma grande variedade de compostos (TESTA et al., 2005). Entre os “softwares” mais utilizados para essa finalidade, destacam-se o **MCASE** (SNYDER et al., 2004), **METEOR** (TESTA et al., 2005) e **DEREK** (SANDERSON; EARNSHAW, 1991).

A importância de se preocupar, desde os estágios iniciais do planejamento de um fármaco, com a baixa toxicidade e a alta especificidade, por exemplo, cresce em um momento em que a credibilidade das grandes indústrias farmacêuticas e do FDA (“Food and Drug Administration”) é colocada em xeque após as recentes retiradas de medicamentos do mercado, como o antiinflamatório Vioxx® (Rofecoxib) e o antidepressivo Aropax® (Cloridrato de paroxetina). Eles tiveram suas vendas suspensas pelos efeitos adversos que provocavam a longo prazo. A necessidade de se planejar novos fármacos e inovar aumenta quando nos deparamos com uma recente estatística: dos 415 fármacos aprovados entre 1998 e 2002, apenas 14% eram inovadores e 9% tinham modificações na fórmula, enquanto que os demais (77%) eram “cópias” de outros já existentes (REVISTA ÉPOCA, 2005).

2. OBJETIVOS

Os objetivos gerais deste trabalho envolvem a identificação e proposição de um modo de ligação de potenciais ligantes dos domínios KH da proteína hnRNP K, selecionados por simulações de “screening” virtual, bem como a avaliação da estabilidade dos mesmos no sítio ligante da proteína e da presença de possíveis subestruturas tóxicas.

Os objetivos fundamentais deste projeto são:

- Levantamento bibliográfico de informações da estrutura tridimensional e da função da proteína hnRNP K e seus domínios KH.
- Aplicar técnicas de bioinformática na construção de modelos dos domínios KH da proteína hnRNP K.
- Validar os modelos construídos através da análise de parâmetros estereoquímicos, de contatos atômicos e de ambientes químicos.
- Realizar simulações de “screening” em bases de dados de estruturas virtuais com propriedades “drug-like” para identificação de compostos que possam interagir com o domínio KH3 da proteína hnRNP K.
- Propor um modo de ligação para cada um dos compostos selecionados, utilizando a abordagem de “docking” flexível.
- Comparar os modos de ligação propostos com a orientação da fita simples de DNA no complexo com KH3, com estrutura resolvida e coordenadas depositadas no PDB, bem como com os campos de interação molecular gerados para a estrutura da proteína.

- Realizar simulações de dinâmica molecular para verificar a estabilidade energética, conformacional e das interações dos modos de ligação propostos, para cada um dos compostos selecionados, no sítio ligante do domínio KH3.

- Verificar a presença de grupamentos toxicofóricos na estrutura dos potenciais ligantes selecionados.

3. MÉTODOS

3.1. Identificação, expressão, purificação e validação do marcador protéico

A proteína hnRNP K é um tipo de proteína que sofre “splicing” alternativo. Dessa forma, foram identificadas duas isoformas (a e b), que apresentam uma pequena diferença na região C-terminal, ou seja, esta diferença é somente aplicável à região do domínio KH3 (Tabela 3), permanecendo o restante da proteína, e conseqüentemente os outros domínios, idêntico em relação as duas isoformas. A isoforma a apresenta 463 resíduos de aminoácidos e a isoforma b apresenta 462.

Tabela 3. Estrutura primária do domínio KH3 das isoformas a e b da proteína hnRNP K. Em destaque estão os resíduos de aminoácidos diferentes (região C-terminal) entre as duas isoformas.

Dominio KH3		Estrutura primária
Isoforma a (82 aminoácidos)		LGGPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKIDEPL EGSEDRIITITGTQDQIQNAQYLLQNSVKQY ADVEGF
Isoforma b (81 aminoácidos)		LGGPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKIDEPL EGSEDRIITITGTQDQIQNAQYLLQNSVKQY SGKFF

A expressão da proteína hnRNP K em *E. coli* BL21(DE), purificada segundo protocolos já definidos e otimizados, vem sendo realizada perante a supervisão da Prof. Dra. Andréia Machado Leopoldino, no Laboratório de Bioquímica Clínica da Faculdade de Ciências Farmacêuticas de Ribeirão Preto da Universidade de São Paulo. As análises de validação quantitativa da hnRNP K por PCR em tempo real já foram iniciadas, bem como ensaios de afinidade das duas isoformas com seqüências oligonucleotídicas específicas. Para esse marcador, está sendo inicialmente investigada a sua expressão em tumores de cabeça e pescoço, especialmente em câncer de língua e carcinoma oral. A proteína de interesse vem sendo analisada por SDS-PAGE, eletroforese bidimensional, pI, espectrometria de massas. Os passos seguintes serão os respectivos ensaios de atividade das duas isoformas da proteína, em presença ou não de DNA e, futuramente, com os potenciais ligantes selecionados, e já em processo de aquisição, com o auxílio das ferramentas computacionais abordadas neste trabalho.

3.2. Alinhamento de seqüências

As ferramentas utilizadas para comparação de seqüências, seja de DNA ou proteínas, representam uma estratégia de grande importância na bioinformática. O rápido acesso a estruturas primárias de proteínas, promovido pelo advento das técnicas de seqüenciamento, facilita a identificação de regiões funcional e/ou estruturalmente conservadas em proteínas, justamente com o auxílio de técnicas de alinhamento entre duas ou mais seqüências. Além disso, se homologia é encontrada em relação a uma proteína bem caracterizada do ponto de vista bioquímico e estrutural, várias propriedades e aspectos da estrutura tridimensional podem ser preditos (BARTON; STERNBERG, 1987).

A busca de proteínas homólogas pode representar o primeiro passo na construção de estruturas tridimensionais através da modelagem molecular por homologia estrutural. Nesse caso, o grau de identidade seqüencial, obtido pelo alinhamento a uma ou várias estruturas conhecidas tridimensionalmente, e a predição das estruturas secundárias que os aminoácidos da seqüência assumirão são aspectos primordiais na construção do modelo estrutural (MUNIZ, 2003).

Há vários “softwares” utilizados para a busca de seqüências homólogas em bancos de dados. Nessa busca, são utilizadas certas ferramentas para a avaliação do grau de similaridade entre as seqüências, com o objetivo de distinguir entre similaridades importantes do ponto de vista biológico ou estrutural de similaridades ao acaso e que não representam importância significativa (ALTSCHUL et al., 1990).

Os primeiros esforços para esclarecer se a similaridade estrutural existente entre proteínas ocorria por homologia ou ao acaso foram realizados por Needleman e Wunsch, o que resultou no desenvolvimento de um algoritmo que herda o nome dos mesmos. Variantes desse algoritmo têm sido desenvolvidos independentemente. Esses algoritmos são mais sensíveis em detectar homologia que os de busca em base de dados, mas são mais lentos em encontrar o alinhamento mais adequado. Entretanto, a grande vantagem do algoritmo de Needleman e Wunsch é que a detecção do melhor alinhamento para duas seqüências é garantida (HÖLTJE et al., 2003a). Conseqüentemente, “softwares” baseados neste método, como **Multalign**, **Bestfit** e **Gap**, têm sido amplamente utilizados em comparações de seqüências biológicas.

Ao passo que o algoritmo original de Needleman e Wunsch é capaz de alinhar somente duas seqüências, muitos programas mais recentes lidam com o alinhamento de mais de duas seqüências. O procedimento de alinhamento de múltiplas seqüências é, significativamente, mais difícil de ser realizado do que o alinhamento de seqüências aos pares. Isso ocorre porque o número de alinhamentos possíveis cresce exponencialmente com o número de seqüências a serem comparadas (DEANE; BLUNDELL, 2003). Vários “softwares” têm sido desenvolvidos para gerar uma solução aproximada para este problema, como o **AMPS** (Alignment of Multiple Pair Segments).

Todos os algoritmos estimam os alinhamentos de seqüências usando esquemas que classificam o pareamento de todos os resíduos alinhados. Em geral, estes esquemas contêm classificações para os 210 pares de aminoácidos possíveis, alojados em uma matriz 20 X 20, em que o alinhamento de resíduos idênticos de aminoácidos (por exemplo, Ile versus Ile) e aqueles considerados similares (por exemplo, Ile versus Leu) recebem uma maior pontuação do que pares bem distintos (por exemplo, Ile versus Asp). Vários esquemas diferentes de classificação têm sido desenvolvidos, incluindo avaliação por identidade sequencial, código genético, similaridade química e estrutural (DEANE; BLUNDELL, 2003).

3.2.1. Alinhamento global

O alinhamento global considera a seqüência completa de resíduos de aminoácidos. Nesse tipo de alinhamento, as penalidades tanto para “gaps” de abertura quanto para “gaps” de extensão são bastante elevadas. Logo, não ocorre a formação de blocos durante os alinhamentos, mas sim de pequenas regiões ou alguns poucos “gaps” distribuídos ao longo da seqüência, preservando, dessa forma, o maior número possível de resíduos alinhados. O alinhamento do tipo global é apropriado para seqüências que apresentam alto grau de similaridade em todo o seu comprimento, já que o alinhamento é otimizado em toda a sua extensão (MUNIZ, 2003).

3.2.2. Alinhamento local

Os alinhamentos locais podem ser representados como blocos desprovidos de “gaps”. A formação de blocos é facilitada pela baixa penalidade imposta aos “gaps” de

abertura e de extensão. Logo, uma sequência de resíduos poderá ter uma maior “mobilidade” e deslocar um grande número de resíduos através da inserção ou deleção de “gaps”. O alinhamento do tipo local é apropriado quando as sequências mostram regiões isoladas de similaridade, por exemplo, múltiplos domínios ou repetições (MUNIZ, 2003).

3.2.3. Alinhamento múltiplo

Quando se dispõe de um banco de dados de proteínas, um alinhamento múltiplo sempre é a melhor opção, pois um grande grupo de proteínas será alinhado e as regiões semelhantes se destacarão de forma pronunciada (MUNIZ, 2003).

3.3. BLAST

Foi realizada uma busca de seqüências homólogas para as sequências dos três domínios KH das duas isoformas (a e b), previamente identificadas, da proteína hnRNP K, com o “software” **BLAST** (Basic Local Alignment Search Tool), sito à pagina da internet www.ncbi.nlm.nih.gov/BLAST/. Para a realização da busca o banco de dados selecionado foi o PDB.

BLAST é um método heurístico para encontrar o melhor alinhamento local entre uma dada seqüência e um banco de dados (ALTSCHUL et al., 1990). Um importante aspecto do **BLAST** é o de não permitir “gaps”, e sim múltiplos resultados de alinhamentos para uma mesma seqüência. O algoritmo do **BLAST** lança mão de estatísticas de alinhamentos seqüenciais sem “gaps”, procurando eliminar estatisticamente homologias casuais, podendo ser configurado com parâmetros tais como: penalidade para a introdução de inserções e deleções (“gaps”) e matriz de substituição. As estatísticas mostram a probabilidade de se obter um alinhamento com o menor número possível de “gaps” (MSP – “Maximal Segment Pair”), com um valor mínimo T pré-fixado pelo usuário, dentro de uma margem de corte S ou um valor de E (*E-value*) menor que o máximo especificado (MUNIZ, 2003).

Basicamente, o algoritmo opera em três etapas:

- Para uma dada sequência de N resíduos a ser estudada, ocorrerá a fragmentação em partes de w resíduos, sendo que este valor w será o número de resíduos a ser utilizado durante a busca em um banco de dados (usualmente $w = 3$ no caso de proteínas). Ou seja, é utilizada uma trinca de aminoácidos e um valor máximo T em uma matriz de alinhamento para cada comparação realizada pela trinca de resíduos.
- A busca em um banco de dados é feita utilizando-se w resíduos, na tentativa de se encontrar esses resíduos correspondentes nas outras sequências do banco de dados.
- Se durante os alinhamentos realizados T for alcançado, w é estendida em ambas as direções para gerar um alinhamento ótimo e sem *gaps* ou MSP com valor de no mínimo S ou valor E (*E-value*).

3.4. AMPS

As sequências identificadas como homólogas (sequências-molde) a cada uma das três sequências-alvo dos domínios KH da proteína hnRNP K na busca com o BLAST, com coordenadas disponíveis no PDB, foram posteriormente alinhadas com o “software” **Multalign**, pertencente ao pacote computacional **AMPS** (BARTON; STERNBERG, 1987).

O “software” **Multalign** exibe várias funções, incluindo o alinhamento de sequências ao par, alinhamento múltiplo e avaliação de significância estatística, bem como funções adicionais que permitem a inclusão de graus de penalidade variável aos “gaps” e esquemas de escore específicos. Para a realização de alinhamentos múltiplos, o “software” **Multalign** emprega o método descrito por Barton e Sternberg (1987). Em primeira instância, a comparação entre todas as sequências é realizada aos pares. A informação é, então, utilizada para a construção de um diagrama para a visualização de grupos com resíduos semelhantes. Com a obtenção de um dado grupo de sequências similares, a segunda etapa visa estabelecer a ordem pela qual as sequências devem ser

alinhas (o par mais similar no topo, seguido das sequências menos similares). Dessa maneira, o algoritmo de alinhamento múltiplo é aplicado às sequências. Primeiramente, o par de sequências mais similar é alinhado, então, a próxima sequência mais similar é alinhada ao alinhamento do primeiro par já alinhado, e assim sucessivamente, sempre com a próxima sequência mais similar se alinhando com o alinhamento anterior, de acordo com o número de sequências dispostas. Após esse procedimento, é necessário uma verificação com relação a inserção de possíveis “gaps” nas sequências-molde. Se ocorre a inserção de “gaps” em elementos de estrutura secundária das estruturas-molde, deve ser realizado, então, o chamado “print vertical”. Nesse procedimento, é inserido manualmente nas sequências as regiões das estruturas-molde que apresentam elementos de estrutura secundária. Dessa forma, na próxima aplicação o algoritmo será tendenciado a não abrir “gaps” nessas regiões.

3.5. Refinamento do alinhamento

O alinhamento obtido pelo “software” **Multalign** foi refinado por sobreposição das estruturas-molde, utilizando como ferramenta os recursos visuais disponíveis no “software” **DS VIEWERPro** (Discovery Studio ViewerPro, 2002), no qual foi verificado o alinhamento dos resíduos das sequências primárias em comparação com o alinhamento dos resíduos obtido por sobreposição das estruturas terciárias (Figura 4).

Figura 4. Verificação da sobreposição dos resíduos para realização de possíveis correções no alinhamento.

3.6. Modelagem molecular por homologia estrutural

O mecanismo evolutivo de duplicação gênica, que está associado a mutações, leva a certas divergências ao longo do tempo e, então, à formação de famílias de proteínas correlacionadas, que apresentam seqüências de aminoácidos e estruturas tridimensionais similares. As proteínas que evoluem a partir de um ancestral comum são conhecidas como homólogas. Duas seqüências homólogas podem ser praticamente idênticas, similares em vários aspectos ou até muito diferentes devido a várias mutações. Um conceito importante em modelagem por homologia é o fato de que a similaridade estrutural é, normalmente, mais preservada que a similaridade seqüencial (HÖLTJE et al., 2003a).

As estruturas tridimensionais de proteínas homólogas são altamente conservadas durante o processo de evolução, pois a estrutura é crucial para o desempenho de funções específicas. As maiores divergências entre proteínas homólogas aparecem com mais freqüência em regiões próximas da superfície protéica, ou seja, nos “loops”. Nessas regiões, até mesmo as propriedades físico-químicas dos resíduos de aminoácidos que sofrem mutações costumam ser diferentes. Em geral, os resíduos localizados no interior das proteínas variam com menor freqüência e com menor distinção de propriedades físico-químicas. Habitualmente, um conjunto de resíduos de aminoácidos que compreendem o centro da proteína e os principais elementos de estrutura secundária permanecem altamente conservados dentro de uma família de proteínas homólogas (SILVA; SILVA, 2007).

Para a abordagem do planejamento racional de fármacos baseado em estrutura, informações estruturais a respeito da proteína (alvo terapêutico) são de vital importância. Embora a base de dados das estruturas resolvidas de proteínas no PDB (“Protein Data Bank”) esteja crescendo exponencialmente nos últimos anos, não há ainda dados estruturais para a maioria das proteínas eleitas como atrativos alvos terapêuticos (DEANE; BLUNDELL, 2003).

Nos casos em que a elucidação estrutural do alvo terapêutico não é possível, modelos do alvo macromolecular (proteína) podem ser construídos por comparação da similaridade de seqüências primárias com as de proteínas homólogas (ou outras proteínas similares) com estruturas resolvidas, pois seqüências de aminoácidos podem ser obtidas com maior facilidade. Esse procedimento comparativo para construção de

modelos estruturais é conhecido como modelagem molecular por homologia estrutural ou modelagem comparativa (DEANE; BLUNDELL, 2003).

A execução da estratégia de modelagem por homologia é um processo bem conhecido e documentado. O método baseia-se no conhecimento de que a conformação estrutural de uma proteína é mais conservada que sua seqüência de aminoácidos, e que pequenas mudanças na seqüência, em geral, resultam em sutis modificações na estrutura tridimensional (NAYEEM; SITKOFF; JUNIOR, 2006). Se pelo menos uma seqüência homóloga para qual a estrutura tridimensional esteja disponível é encontrada, o método de escolha para predição da estrutura tridimensional de uma proteína alvo é a própria modelagem comparativa (HÖLTJE et al., 2003a). O resultado é um conjunto de coordenadas, tanto da cadeia principal como das cadeias laterais, dos aminoácidos que compõem a proteína. Embora a modelagem por homologia gere modelos menos precisos do que os métodos experimentais de resolução estrutural, a mesma pode ser de extrema utilidade na proposição e na verificação de hipóteses em biologia molecular (DEANE; BLUNDELL, 2003).

Os fundamentos da modelagem por homologia estão presentes em uma variedade de “softwares”, tanto na esfera comercial quanto na pública. Para os usuários destas ferramentas, uma importante questão é se algum dos softwares disponíveis se distingue dos outros com relação à performance (NAYEEM; SITKOFF; JUNIOR, 2006). Para os pesquisadores da área de química computacional, o interesse na alta performance deste método é enorme, pois, na ausência de estruturas resolvidas de alvos moleculares de interesse terapêutico, há a necessidade de criação de modelos com alto grau de confiança, para a aplicação de simulações de “docking” e “screening” virtual, com o intuito de identificar e otimizar novos protótipos.

Nayeem, Sitkoff e Junior (2006), comparando a precisão de modelos de proteínas de interesse farmacêutico gerados por vários “softwares” disponíveis comercialmente ou de domínio público, verificaram que quando a identidade seqüencial é maior do que 40%, os modelos gerados através da estratégia de modelagem comparativa possuem um mesmo nível de exatidão, não havendo diferenças significativas quando comparados às estruturas cristalográficas. Quando a identidade seqüencial é menor, os resultados tendem a variar, com alguns “softwares” apresentando resultados mais precisos e confiáveis.

O primeiro passo na modelagem por homologia constitui-se na identificação de estruturas tridimensionais conhecidas que possam atuar como uma base estrutural para a modelagem da seqüência-alvo. Esta identificação pode ser realizada levando-se em consideração vários aspectos como: conhecimento estrutural, similaridade da função, expressão pelo mesmo grupo de genes, similaridade seqüencial ou até correlação evolutiva. (DEANE; BLUNDELL, 2003).

Com respeito à faixa aceitável de identidade seqüencial para a modelagem por homologia, é bem conhecido e a literatura descreve como significativo um valor acima de 30 % entre a(s) proteína(s)-molde e aquela que será modelada (proteína-alvo) (VITKUP et al., 2001; D'ALFONSO; TRAMONTANO; LAHM, 2001; SALI, 1998). Esse valor é dependente do número de resíduos da proteína que será modelada (Figura 8), sendo menos crítico quanto maior o comprimento da proteína que alinha com o molde. Considerando o valor de “threshold” de 30%, a modelagem por homologia se torna significativa para proteínas com mais de 60 resíduos de aminoácidos

Gráfico 1. “Threshold” para realização da modelagem por homologia.

Vários exemplos da utilização dessa estratégia estão presentes na literatura (NAYEEM; SITKOFF; JUNIOR, 2006). Ring et al. (1993) identificaram inibidores de serina e cisteína proteases, com base no emprego de modelos moleculares por homologia. Schafferhans e Klebe (2001) utilizaram modelos, gerados por homologia estrutural, para identificar o modo como certos compostos se ligavam às proteínas. Vangrevelinghe et al. (2003) foram capazes de gerar modelos por homologia, que foram aplicados em simulações de “screening” virtual, para identificar potentes inibidores de

uma coleção com 400.000 compostos da Novartis. Um exemplo recente e bastante interessante (Evers; Klabunde, 2005), foi promovido pela aplicação, bem sucedida, de modelos para o “screening” virtual de antagonistas de receptores GPCR (do inglês, “G-protein coupled receptor”).

Enyedy et al. (2001) reportaram o sucesso no planejamento e desenvolvimento de 15 novos inibidores de matriptase, uma serino-protease envolvida em processos de câncer invasivo e metástase. O estudo, que envolveu a aplicação de “screening” virtual, utilizou como receptor um modelo baseado na estrutura (molde) de trombina, com a qual o modelo compartilhava apenas 34% de identidade seqüencial. Ainda relacionado ao câncer, Diller e Li (2003) reportaram o sucesso no planejamento de inibidores de tirosina e serina/treonina quinases. O trabalho compreendeu “screening” virtual de compostos “drug-like”, utilizando modelos construídos por homologia, na faixa de 30 a 70 % de identidade seqüencial com estruturas-molde extraídas do PDB.

As designações exatas de regiões estruturalmente conservadas dentro de uma família de proteínas homólogas é afetada por vários fatores. O procedimento depende do número disponível, no PDB, de proteínas homólogas com estruturas resolvidas. Um melhor resultado pode ser alcançado quando mais de uma estrutura com coordenadas resolvidas está disponível, pois nesta situação várias estruturas podem ser comparadas para determinação das regiões estruturalmente conservadas. Para reconhecer as regiões conservadas as proteínas devem ser sobrepostas entre si. Este procedimento é realizado pela utilização de métodos de ajuste dos quadrados mínimos (do inglês, “least-squares fitting methods”). O principal problema, neste contexto, é a seleção dos correspondentes átomos a serem sobrepostos. Em uma primeira aproximação, as estruturas podem ser sobrepostas utilizando o ajuste do quadrado mínimo dos átomos de carbono- α das proteínas. A sobreposição inicial, então, pode ser otimizada utilizando dados por comparação de pontos localizados em elementos de estrutura secundária que são considerados conservados (HÖLTJE et al., 2003a).

As diferenças significativas entre estruturas de proteínas homólogas ocorrem, preferencialmente, nas regiões de “loop”. Logo, a construção dessas regiões estruturalmente variáveis é uma tarefa muito mais desafiante. Diferenças em relação ao número de aminoácidos, causadas por inserções ou deleções, são situações que dificultam ainda mais o procedimento de modelagem. Vários métodos para geração de “loops” têm sido desenvolvidos e descritos na literatura. Um bom guia para a

modelagem dessas regiões pode ser a estrutura de um segmento de comprimento equivalente de uma proteína homóloga. Investigações das regiões variáveis em proteínas homólogas têm mostrado que, nos casos em que as regiões de “loops” apresentam o mesmo comprimento e aminoácidos com as mesmas características, a conformação de ambas será a mesma. Logo, as coordenadas podem ser transferidas diretamente para o modelo em construção da proteína alvo. Se não existe algum “loop” comparável entre as proteínas, duas outras estratégias podem ser empregadas. As coordenadas das regiões variáveis podem ser construídas, então, a partir de segmentos peptídicos que são encontrados em outras proteínas e que se encaixam corretamente no modelo espacial, ou gerando um segmento de “loop” pela estratégia *de novo*. A primeira abordagem, conhecida por método de “loop search”, procura por segmentos peptídicos em proteínas que reúnem certos critérios geométricos específicos em bancos de dados. A geometria específica para a pesquisa é dada por distâncias e coordenadas, incluindo os resíduos de aminoácidos das regiões de “loop” no modelo. O produto de uma respectiva pesquisa, realizada por “softwares”, é uma coleção de “loops” que satisfazem as recomendações geométricas. Geralmente, os melhores fragmentos são retidos e submetidos a uma melhor avaliação. Critérios adicionais não usados explicitamente durante o procedimento de “loop search”, os quais podem promover uma classificação para determinar a preferência de um fragmento sobre os outros candidatos. Os fragmentos de “loops” encontrados podem ser avaliados em relação à qualidade de encaixe aos resíduos que compreendem aquela região de “loop”, pela determinação da homologia entre as seqüências, avaliação das interações estéricas ou por critérios energéticos (HÖLTJE et al., 2003a).

O método de “loop search” oferece a vantagem de que todos os “loops” encontrados apresentam geometrias aceitáveis e detêm conformações de proteínas conhecidas. Não é garantido que o fragmento escolhido se encaixe adequadamente ao modelo, então, contatos atômicos podem ocorrer. Se isso de fato ocorre, o método de *de novo* “generation” se torna uma ferramenta alternativa. A partir desta abordagem as coordenadas da cadeia polipeptídica de um “loop” podem ser construídas entre dois fragmentos conservados de uma proteína, utilizando valores numéricos gerados randomicamente para todos os ângulos torcionais. Vários algoritmos têm sido desenvolvidos para otimizar a estratégia de busca e avaliação de energia das

conformações geradas. Devido à maior complexidade, o método *de novo* só é utilizado para “loops” com até 7 resíduos de aminoácidos (HÖLTJE et al., 2003a).

A partir do momento que a cadeia polipeptídica principal foi construída, o passo seguinte é a adição das cadeias laterais ao modelo. A predição das conformações das numerosas cadeias laterais é um problema mais complexo que a predição da conformação da cadeia principal. A maioria das cadeias laterais possui um ou vários graus de liberdade e, logo, podem adotar uma variedade de conformações energeticamente viáveis (SILVA; SILVA, 2007).

Um procedimento desenvolvido para examinar a relação entre as posições das cadeias laterais em estruturas homólogas de proteínas globulares, parte da premissa de que cadeias laterais adotam, geralmente, apenas um pequeno número das muitas conformações possíveis. Cadeias laterais com dois ângulos *Chi*, por exemplo, apresentam de 4 a 6 conformações comuns. Todos os rotâmeros observados são combinações de conformações *gauche* e *anti*. A partir destas avaliações estatísticas, coleções de rotâmeros têm sido desenvolvidas. Uma das coleções de rotâmeros mais utilizadas foi criada por Ponder & Richards, a qual contém 67 rotâmeros para 17 aminoácidos. Vários “softwares” de modelagem por homologia usam esta coleção para gerar as cadeias laterais de proteínas homólogas (HÖLTJE et al., 2003a).

A conformação exata de uma cadeia lateral depende, essencialmente, do ambiente encontrado pelo aminoácido na proteína real. No interior da proteína, interações hidrofóbicas são predominantes e resultam em um enovelamento que comprime os resíduos de aminoácidos. O contato com outros resíduos de aminoácidos também pode influenciar as conformações da cadeia lateral. Modificações devem ser aplicadas, por exemplo, quando aminoácidos estão envolvidos em interações específicas, como pontes dissulfeto, pontes salinas, interações eletrostáticas ou ligações de hidrogênio. Variações também ocorrem quando os resíduos de aminoácidos estão localizados na superfície da proteína. As exceções apresentadas devem ser tratadas especificamente em cada caso (HÖLTJE et al., 2003a).

3.6.1. Modelagem por homologia dos domínios KH da proteína hnRNP K

Uma vez obtido o alinhamento entre as seqüências-molde e a seqüências-alvo, o próximo passo foi a construção dos modelos dos três domínios KH da proteína hnRNP

K. O “software” utilizado para tal finalidade foi o **Modeller 9.0a** (SALI; BLUNDELL, 1993). Para a modelagem por homologia de proteínas, como a hnRNP K, o “software” **Modeller** oferece um excelente suporte. O método empregado por este software é a “Modelagem Molecular por Satisfação de Restrições Espaciais” (Figura 5) [SILVA, 1999].

O alinhamento entre as sequências-molde e a sequência-alvo funciona como o “input” do “software”. O “output” gerado foi um conjunto de coordenadas atômicas de 3 modelos 3D para cada uma das três sequências-alvo (domínios KH), contendo todos os átomos das cadeias principal e lateral. A partir do alinhamento com as sequências das estruturas-moldes, o programa calcula várias restrições de distâncias e de ângulos torsionais na sequência-alvo, as quais são parâmetros extras que são adicionados ao campo de força para tendenciar os cálculos. A forma destes parâmetros é obtida empiricamente a partir de uma análise estatística das relações entre muitos pares de estruturas de proteínas homólogas, inseridas em um banco de dados contendo 105 alinhamentos entre 416 proteínas, as quais possuem estruturas 3D conhecidas (SALI; BLUNDELL, 1993).

Figura 5. Modelagem Molecular por Satisfação de Restrições Espaciais. Inicialmente, as estruturas 3D (‘3D’) conhecidas são alinhadas com a sequência-alvo (‘SEQ’). A seguir, parâmetros espaciais, tais como distâncias Ca-Ca, ligações de hidrogênio e torções, são transferidos do molde para o alvo. Com isso,

várias restrições espaciais são extraídas. O modelo é então obtido satisfazendo-se, ao máximo possível, todas essas restrições.

3.7. Validação dos modelos

A partir do momento que um modelo é gerado através da utilização de modelagem por homologia, e subsequente otimizado por técnicas de mecânica ou dinâmica molecular, se torna importante e relevante a avaliação dos níveis de qualidade e confiabilidade do mesmo (HÖLTJE et al., 2003a). Esta é uma tarefa árdua, pois o nível de qualidade de um modelo gerado por homologia estrutural depende de um grande número de propriedades de diferentes graus de organização estrutural, como: exatidão estereoquímica, qualidade do empacotamento e confiabilidade do enovelamento.

Para verificar a qualidade estereoquímica das estruturas dos modelos construídos, a exatidão de parâmetros como comprimento das ligações, ângulos entre ligações, ângulos torcionais e quiralidade dos aminoácidos, precisa ser avaliada. Normalmente, em estruturas 3D de proteínas, o comprimento das ligações e os ângulos formados entre elas estão perto dos valores ideais estabelecidos. Logo, os valores obtidos a partir dos modelos podem ser comparados com os valores da proteína-molde cristalizada para a descoberta de irregularidades estereoquímicas que poderiam revelar uma estrutura inadequada (HÖLTJE et al., 2003a). Para a avaliação da qualidade dos modelos gerados pelo “software” Modeller, foram utilizados os seguintes “softwares”: **Procheck**, **Whatif** e **Verify 3D**.

O “software” **Procheck** (LAKOWSKI; MACATHUR; THORNTON, 1993) avalia diversos parâmetros estereoquímicos, tais como ângulos torcionais da cadeia principal (Φ e Ψ), ângulos torsionais das cadeias laterais (*Chi*), maus contatos (ou impedimentos estéricos), energias das ligações de hidrogênio, planaridade das ligações peptídicas, desvios em relação a geometria tetraédrica dos carbonos- α e outros. Uma qualidade estereoquímica média relativa aos parâmetros avaliados é representada pelo “Fator G”. Os cálculos comparativos baseiam-se em um banco de dados de proteínas que contém estruturas a diferentes níveis de resolução. A rigor, o Fator G é sempre referido, nos resultados, a uma determinada resolução estrutural, na qual existe um valor médio deste parâmetro associado às proteínas do banco de dados.

O interior de proteínas globulares contém cadeias laterais que se encaixam com certa complementaridade. As altas densidades de empacotamento observadas em proteínas são consequência deste fato, o que resulta em segmentos de estrutura secundária muito próximos: hélices contra hélices, hélices contra folhas β e/ou folhas β contra folhas β . O empacotamento do interior das proteínas globulares é a maior contribuição para a estabilidade de toda a conformação. Logo, a qualidade do empacotamento pode ser usada para estimar a confiabilidade do modelo protéico (HÖLTJE et al., 2003a).

Existe uma variedade de métodos que usam uma grande quantidade de informação derivada de estruturas de proteínas resolvidas para estimar a qualidade do empacotamento de modelos de proteínas. Partindo da premissa de que as interações átomo-átomo são as principais determinantes da conformação protéica, Vriend e Sander (1993) desenvolveram um método que checa a qualidade do empacotamento de modelos de proteínas através do cálculo do chamado “índice da qualidade de contato”. Este índice é a medida entre a distribuição dos átomos ao redor de uma cadeia lateral de um aminoácido e as distribuições equivalentes observadas em proteínas com estruturas resolvidas. Por esta razão, foi gerado um banco de dados que contém uma distribuição de probabilidade de contato atômico para todas as cadeias laterais dos aminoácidos. Nesse banco de dados é descrita a probabilidade de um certo átomo ocorrer em uma região particular ao redor de uma cadeia lateral. Os valores de probabilidade são usados para avaliar a qualidade do contato em um modelo. Quanto maior for a correlação entre as distribuições no modelo e as estruturas resolvidas maior será a qualidade do índice.

A qualidade dos contatos atômicos envolvendo os átomos de cada resíduo foi avaliada utilizando-se o módulo **Coarse Packing Quality Control** do “software” **Whatif**, o qual compara a distribuição das posições de átomos em torno de cada resíduo. Um escore menor do que -5,0 para um resíduo significa contatos atômicos ruins ou incomuns, mas não implica, necessariamente, em uma estrutura incorreta. Existe a necessidade, entretanto, de examinar-se o resíduo (VRIEND; SANDER, 1993).

O modelo protéico também pode ser avaliado em relação à qualidade dos ambientes químicos. O “software” **Verify 3D**, utilizado para tal fim, determina os ambientes químicos de cada resíduo do modelo e atribui escores com referência a uma matriz construída a partir de uma análise estatística envolvendo estruturas de proteínas do PDB (LUTHY; BOWIE; EISENBERG, 1992). Nessa matriz estão contidas três

propriedades que cada resíduo apresenta dentro de cada um dos 18 ambientes químicos definidos. Finalmente, o “software” realiza uma promediação na “janela” com o objetivo de detectar regiões de baixa qualidade. A estratégia empregada pelo “software” **Verify 3D** consiste, efetivamente, em medir a compatibilidade entre uma determinada sequência e a estrutura tridimensional de uma proteína (MUNIZ, 2003). O “software” opera, basicamente, em três etapas:

- Resolução da estrutura tridimensional em uma sequência unidimensional dentro de um ambiente. Esses ambientes são classificados de acordo com a área da cadeia lateral imersa na proteína, a fração de área da cadeia lateral exposta a átomos polares e a estrutura secundária local.

- Geração de uma matriz de comparação dependente da posição, conhecida como perfil 3D. O cálculo é realizado de acordo com o ambiente de cada resíduo da sequência, ou seja, a probabilidade de se encontrar cada um dos 20 aminoácidos em cada uma das classes de ambientes, como observado em um banco de dados protéico e suas respectivas sequências, resultando na formação de uma matriz 18x20 (18 ambientes possíveis x 20 aminoácidos).

- Alinhamento da sequência primária com o seu perfil tridimensional. A qualidade do alinhamento relaciona-se com a medida da compatibilidade da sequência com sua estrutura 3D descrita por seu perfil.

3.8. “Docking” molecular

As técnicas de “docking”, desenvolvidas para encontrar a melhor orientação e conformação de um ligante no seu sítio receptor, vêm sendo, há algum tempo, empregadas no processo de planejamento e desenvolvimento de fármacos. A etapa de ligação entre um fármaco e o seu alvo macromolecular protéico é um processo complexo por natureza. Fatores entrópicos e entálpicos influenciam, sobremaneira, nas interações formadas. A flexibilidade do ligante e da proteína, o efeito do ambiente protéico na distribuição de cargas do ligante e as interações que podem ocorrer com as moléculas de água presentes no meio, são aspectos que dificultam ainda mais a descrição detalhada desse processo. A ideia geral contida nas técnicas de “docking” é a de gerar um leque de conformações do complexo ligante-proteína e ordená-las por

escore com base em suas estabilidades (ALONSO; BLIZNYUK; GREARY, 2006; TAYLOR; JEWSBURY; ESSEX, 2002).

Uma das características mais valiosas dos métodos de “docking” é a sua capacidade de reproduzir modos de ligação observados experimentalmente, funcionando até como uma forma de validação dos mesmos. Para realizar um teste desse nível, um ligante é extraído de seu complexo cristalográfico e submetido a simulações com o sítio ligante da proteína. Dessa forma, os modos de ligação obtidos nas simulações são comparados com os respectivos modos de ligação obtidos experimentalmente. Outra possibilidade inerente ao método é a capacidade de sua função de escore de ordenar ligantes de acordo com valores experimentais de atividade. Essa correlação é feita através dos valores de escore obtidos nas simulações e os valores experimentais de atividade, como por exemplo, IC₅₀ (VERDONK et al., 2003).

De maneira geral, os “softwares” de “docking” são formados por uma combinação de dois componentes: um algoritmo de busca e uma função de escore (VERDONK et al., 2003; TAYLOR; JEWSBURY; ESSEX, 2002). O algoritmo é utilizado na busca de possíveis modos de ligação, e permite explorar os graus de liberdade translacional, rotacional e conformacional do ligante, bem como o de ligações rotacionáveis na proteína. A função de escore é aplicada para tentar distinguir os modos de ligação teoricamente mais próximos dos obtidos experimentalmente entre os demais modos de ligação, explorados pelo algoritmo de busca e, dessa forma, ordenar os diferentes modos de ligação apresentados. As funções de escore podem ser estabelecidas de acordo com campos de força de mecânica molecular, parâmetros empíricos de cálculos de energia livre ou até de acordo com parâmetros denominados “knowledge-based”.

Uma das aplicações dos *softwares* de “docking” ocorre em “screening” virtual em bases de dados, situação em que amplas coleções virtuais de compostos são submetidas às simulações de “docking” em um sítio ligante protéico e os respectivos compostos ordenados de acordo com a afinidade pelo alvo macromolecular, sugerida pela função de escore (SCHNEIDER; BÖHM, 2002). A abordagem de “screening” virtual é uma grande contribuinte no processo de busca de ligantes, pois compostos com potencial de interação com o sítio receptor estudado podem ser futuramente investigados com maior precisão e rigor, reduzindo drasticamente o tempo de identificação de novos protótipos quando comparada com as estratégias convencionais.

3.8.1. Simulações de “screening” virtual

As simulações de “screening” virtual foram realizadas com o “software” **GOLD 3.3** (VERDONK et al., 2003) para o domínio KH3 (código PDB 1J5K, complexo da isoforma b com *ssDNA*) em relação a três bases de dados (Ilibdiverse, IResearch Library e Chembridge) de estruturas de moléculas de fármacos, substâncias ativas e/ou moléculas com propriedades “drug-like”. A base de dados Ilibdiverse contém aproximadamente 1.200 estruturas moleculares virtuais de fármacos ou substâncias ativas clássicas. Em relação à base de dados IResearch Library, foi utilizada uma subcoleção de compostos contendo aproximadamente 100.000 estruturas com propriedades “drug-like”. As subcoleções de compostos Diverset, MolecularWeightset, MicroFormats e CNSset, pertencentes à base de dados Chembridge, também foram utilizadas nas simulações de “screening” virtual. Diverset contém aproximadamente 50.000 estruturas de moléculas pequenas com propriedades “drug-like” e abrangendo diversas características farmacofóricas espaciais relevantes para a manutenção de interações com os mais diversos alvos moleculares. A subcoleção MolecularWeightset contém aproximadamente 30.000 estruturas com características “drug-like” que se dispõem em ordem crescente de peso molecular na base de dados. CNSset é composta de estruturas submetidas a diversas análises computacionais, em que existe alta probabilidade de encontrar protótipos com biodisponibilidade por via oral e capacidade de penetrar a barreira hematoencefálica (BHE). Em relação à MicroFormats, a mesma é composta por moléculas “drug-like” com grande diversidade estrutural preparadas em DMSO.

A base metodológica do “software” **GOLD** é a execução de simulações de “docking” flexível utilizando um algoritmo genético. Os parâmetros utilizados nesse algoritmo foram originalmente otimizados em relação a um grupo de 305 estruturas de complexos com coordenadas depositadas no PDB (VERDONK et al., 2003). Dentre os parâmetros disponíveis no “software”, foi utilizada uma população equivalente a 100 confôrmeros, 10.000 operações, 100 mutações e 100 “crossovers”. Os cálculos de “docking” foram realizados dentro de uma esfera de raio de 15 Å, tendo como centro o átomo de carbono delta 1 da cadeia lateral do resíduo de I49. A estrutura com código PDB 1J5K foi resolvida por ressonância magnética nuclear, e as orientações de seus átomos de hidrogênio foram então consideradas para realização das simulações. A

orientação de melhor escore para cada composto foi selecionada através de uma função matemática, implementada no “software” **GOLD**, denominada GoldScore. Com base nessa função, o “software” classifica as orientações das moléculas do banco de dados de acordo com um padrão de afinidade (escore), do ponto de vista de estabilidade energética, em relação ao sítio ligante da proteína. Foi gerada uma orientação para cada molécula das bases de dados utilizadas e, dessa forma, os 50 compostos que apresentaram maior escore para cada coleção ou subcoleção de compostos foram selecionados para investigações mais criteriosas.

3.8.2. Modelagem dos compostos selecionados

A modelagem de compostos que interagem com um alvo macromolecular protéico é similar à modelagem por homologia no tocante ao objetivo de se predizer a estrutura terciária da molécula, além de ser complementar a essa técnica no que diz respeito ao estudo do reconhecimento molecular existente. Diferentes tipos de cálculos teóricos têm sido utilizados em química computacional visando a predição da geometria e ao cálculo das propriedades eletrônicas de moléculas de interesse. Os métodos dividem-se, basicamente, em duas categorias: os empíricos, como mecânica e dinâmica molecular, os quais são baseados no formalismo matemático que advém da mecânica clássica, e os de mecânica quântica, incluindo cálculos semi-empíricos e *ab initio*, onde a resolução da equação de Schrödinger, por métodos aproximados, descreve o comportamento dos elétrons, nos orbitais, ao redor dos núcleos atômicos (HÖLTJE et al., 2003b).

As estruturas dos compostos selecionados nas simulações de “screening” virtual foram extraídas de suas respectivas bases de dados e submetidas ao processo de minimização de energia por mecânica molecular. O campo de força utilizado foi o MMFFem associação ao algoritmo “steepest descent”, implementados no “software” **Spartan v.06** (Spartan User’s Guide, 2006).

3.8.3. “Rescore”

Uma vez modeladas, as estruturas dos compostos foram submetidas a simulações de “docking” flexível com o “software” **GOLD** para o procedimento

denominado de “rescore”. Os parâmetros utilizados no “docking” de cada uma dessas moléculas com o sítio receptor do domínio KH3 e relativos ao algoritmo genético foram diferentes dos utilizados nas simulações de “screening” virtual: população equivalente a 100 confôrmers, 100.000 operações, 95 mutações e 95 “crossovers”. Também foi diferente o número de orientações geradas com o algoritmo genético empregado. Foram selecionadas 10 orientações de maior escore para cada composto investigado, onde cada uma delas foi analisada minuciosamente no sítio ligante do domínio KH3.

3.9. Determinação dos potenciais de interação molecular fármaco-receptor

A formação de um complexo fármaco-receptor se inicia através do processo de reconhecimento molecular, em que o receptor precisa reconhecer as propriedades moleculares do fármaco que se aproxima para realizar uma interação forte e específica. A etapa de reconhecimento molecular ocorre a distâncias consideravelmente grandes e precede a formação das interações que sacramentam a formação do complexo. O campo eletrostático que envolve cada molécula apresenta um papel crítico no processo de reconhecimento. Quando a distância entre a superfície do fármaco e do receptor diminui outras propriedades moleculares, como polarizabilidade e hidrofobicidade, se tornam preponderantes (HÖLTJE et al., 2003b).

Seguindo este contexto, potenciais de interação molecular podem ser determinados na estrutura do receptor através de cálculos sistemáticos que envolvem energias de interação entre o receptor e grupos químicos de prova de interesse, em que dados representativos para o entendimento dos potenciais de interação naquele receptor, sem informação prévia de ligantes, podem ser obtidos. Essa abordagem se torna ainda mais interessante quando se deseja identificar promissores protótipos para novos e atrativos alvos terapêuticos, como a proteína hnRNP K.

3.9.1. Potenciais eletrostáticos moleculares

O conhecimento de potenciais eletrostáticos se torna de vital importância quando interações moleculares são estudadas. As forças eletrostáticas de longo alcance governam o contato inicial de moléculas que se aproximam. Existem diversos tipos de interações intermoleculares que podem manter um complexo fármaco-receptor, entre

elas: interações iônicas, ligações de hidrogênio, interações de van der Waals, dipolo-dipolo, íon-dipolo e interações hidrofóbicas (HÖLTJE et al., 2003b; PATRICK, 2005).

Em princípio, as forças de interação molecular podem ser agrupadas em três componentes: eletrostática, indutiva e dispersiva. As interações de ordem eletrostática ocorrem entre moléculas polares que possuem carga ou um momento de dipolo permanente. As forças indutivas são formadas por moléculas polares que interagem com moléculas não-polares. As cargas ou dipolos das moléculas polares produzem um campo elétrico que é capaz de mudar a distribuição dos elétrons nas moléculas não-polares e, dessa forma, induzir um momento de dipolo nas mesmas. Quando as moléculas que interagem entre si apresentam características predominantemente hidrofóbicas, as forças dispersivas são majoritárias. Em moléculas hidrofóbicas a flutuação dos elétrons pode induzir a formação de um momento de dipolo na molécula vizinha. As forças dispersivas são consideradas fracas e se desfazem facilmente com o aumento da distância entre as moléculas. Entretanto, formam o principal componente de atração entre moléculas neutras apolares (HÖLTJE et al., 2003b).

As interações intermoleculares aparecem amplamente nas regiões moleculares que apresentam carga. Devido às cargas, sobretudo aos momentos de dipolo, um campo eletrostático tridimensional é gerado no ambiente que envolve as moléculas. Mesmo entre moléculas neutras, a distâncias consideradas moderadas, existe um potencial eletrostático significativo. Esse potencial pode ser representado como a energia de interação entre a distribuição eletrônica molecular e uma carga pontual positiva que está localizada em um “grid” tridimensional em qualquer região do espaço ao redor da molécula. Dessa forma, para a determinação de potenciais eletrostáticos as propriedades eletrônicas das moléculas precisam de tratamento minucioso (WADE, 2006).

3.9.2. Campos de interação molecular

As forças de interação não-covalentes determinam a geometria e a simetria do arranjo molecular entre um fármaco e seu sítio ligante. Como regra geral, a ligação entre fármaco e receptor só ocorre, efetivamente, se a energia de interação gerada supera as forças repulsivas de van der Waals. Os campos de interação molecular (MIF, do inglês, “molecular interaction fields”) podem ser utilizados na investigação das condições

energéticas entre um receptor e seu ligante (HÖLTJE et al., 2003b; GOODFORD, 1985).

Os campos de interação molecular podem ser calculados para qualquer molécula com estrutura tridimensional conhecida. Os MIFs descrevem a variação espacial da energia de interação entre um alvo molecular e um grupo químico de prova. O alvo molecular pode ser uma macromolécula, um complexo molecular ou até um composto de baixo peso molecular (WADE, 2006; HÖLTJE et al., 2003b).

Existem vários “softwares” capazes de computar os MIFs ao redor de uma molécula. Para que se proceda a essa análise, é necessária a obtenção das coordenadas atômicas x, y e z do alvo molecular. O alvo molecular é então envolvido por um “grid” ortogonal imaginário, onde os MIFs são calculados para os grupos químicos de prova em cada ponto do “grid”. Os grupos químicos de prova representam átomos ou pequenos grupos de átomos, como por exemplo, oxigênio de carbonila, que é um átomo de oxigênio com dois pares de elétrons sp². Os grupos químicos de prova refletem as características químicas de um componente que pode interagir com o alvo molecular (WADE, 2006). Através da utilização de gráficos computacionais, os campos de interação molecular podem ser representados como contornos tridimensionais isoenergéticos. Os contornos com energias altamente positivas indicam regiões pelas quais o grupo de prova seria repelido, enquanto que as regiões amplamente negativas correspondem a regiões que favorecem energeticamente interações com o grupo químico de prova (HÖLTJE et al., 2003b).

No decorrer do cálculo, o grupo de prova é movido sistematicamente através dos pontos regulares do “grid”. A cada ponto alcançado a energia de interação entre o grupo de prova e o alvo molecular é calculada. (GOODFORD, 1985). A energia de interação não-covalente é calculada a cada coordenada x, y e z através da soma de vários componentes:

$$E_{xyz} = \sum E_{lj} + \sum E_{el} + \sum E_{hb}$$

O termo E_{lj} é descrito pela função de Lennard-Jones e representa a energia de interações de van der Waals. O termo E_{el} representa a energia de interação

eletrostática e E_{nd} representa a energia de interação através da formação de ligações de hidrogênio.

3.9.2.1. Almond

Os campos de interação molecular foram gerados com o módulo **Almond** (PASTOR et al., 2000) do pacote computacional **Sybyl v.7.3** (SYBYL USER GUIDE, 2005) para o sítio ligante do domínio KH3 (código PDB: 1J5K) da proteína hnRNP K. Os cálculos foram realizados com base nas interações moleculares do domínio KH3 com 3 grupos químicos de prova, sendo eles: hidrofóbico (DRY), oxigênio de carbonila e nitrogênio de amida. Depois de gerados os MIFs, as orientações dos compostos selecionados nas simulações de “screening” virtual e da sequência oligonucleotídica TCCC foram carregadas individualmente nos modelos para efeito de comparação. O espaço do “grid” foi estabelecido em 0.5 Å e os nós filtrados em 100, com 35% de pesos relativos.

3.10. Predições ADMET

Nos últimos anos, percebeu-se um avanço considerável no desenvolvimento de técnicas de modelagem molecular que simulam as interações de um ligante em seu sítio receptor. Pode-se destacar também a evolução de técnicas que elevaram o poder de predição do comportamento dos ligantes em sistemas biológicos, que são aplicadas no estudo de diversas propriedades, como: absorção, distribuição, metabolismo, excreção e toxicidade (EKINS; ROSE, 2002).

Na busca de inovação e desenvolvimento de novos fármacos, é evidente a pressão do mercado sobre a otimização dos recursos financeiros. Pode-se destacar também a restrição, ou crescente dificuldade, com relação à disponibilização de animais para utilização em testes de toxicidade. A maioria dos fármacos retirados do mercado ocorre pelo fato de exercerem efeitos tóxicos indesejáveis (O'BRIEN; GROOT, 2005), citando o recente exemplo do antiinflamatório Vioxx® (REVISTA ÉPOCA, 2005). Dessa forma, os métodos de predição de toxicidade *in silico* surgem como uma importante e alternativa ferramenta na seleção ou priorização de moléculas promissoras a serem avaliadas com maior cautela em testes de toxicidade, reduzindo, sobremaneira,

os custos financeiros inerentes ao processo, o uso indiscriminado de animais e satisfazendo as precauções em relação à toxicidade desde as fases iniciais do processo de desenvolvimento de fármacos.

A grande maioria das predições de toxicidade de compostos *in silico* é baseada na avaliação da relação entre estrutura química e atividade biológica e podem identificar riscos potenciais à saúde humana associados aos compostos desenvolvidos. Os métodos preditivos podem ser divididos em qualitativos e quantitativos. Os sistemas qualitativos realizam previsões que podem confirmar ou descartar o tipo de risco avaliado. De maneira mais complexa, os métodos quantitativos se utilizam da aplicação de modelos matemáticos e tentam encontrar uma correlação entre estrutura química (gerada por descritores derivados de certas propriedades moleculares) e o nível de efeito biológico exercido (RIDINGS et al., 1996; SIMIN-HETTICH; ROTHFUSS; STEGER-HARTMANN, 2006).

A idéia do emprego de métodos computacionais na predição de mutagenicidade de novos compostos, por exemplo, representa uma abordagem atrativa. O sucesso dessas abordagens é muito grande e os custos financeiros, bem como o tempo gasto, são relativamente baixos. As predições *in silico* de mutagenicidade são baseadas no entendimento de que esse processo está intimamente relacionado à formação de ligações covalentes entre um composto químico e o DNA celular. A capacidade de formação de ligações covalentes é um fator que depende das propriedades eletroquímicas das moléculas, e a maioria das espécies químicas capazes de interagir com DNA podem ser identificadas em bases de dados (SNYDER et al., 2004). Os “softwares” de análise qualitativa disponíveis comercialmente, na atualidade, promovem grande confiabilidade em relação à predição de mutagenicidade (SIMIN-HETTICH; ROTHFUSS; STEGER-HARTMANN, 2006).

Existem outros efeitos tóxicos que podem ser diretamente atribuídos a mecanismos simples relacionados diretamente a certas propriedades físico-químicas de compostos químicos, e que se correlacionam com alguns tipos de efeitos no organismo humano. Efeitos tóxicos dessa natureza podem ser preditos com alto grau de confiabilidade *in silico*, como por exemplo, proliferação de peroxissomo (relacionada à hepatotoxicidade), irritação e hipersensibilidade cutânea (SIMIN-HETTICH; ROTHFUSS; STEGER-HARTMANN, 2006).

Em outro contexto se enquadram os efeitos tóxicos causados por mecanismos múltiplos que envolvem diversos fatores e variáveis distintas, em que o poder preditivo, ainda, é limitado a poucas dessas variáveis. Isso ocorre pelo fato do número de dados disponíveis para estes efeitos serem escassos e às vezes, pouco compreendidos. Nesse âmbito se destacam os tipos de toxicidade aguda e crônica, carcinogenicidade e letalidade (SIMIN-HETTICH; ROTHFUSS; STEGER-HARTMANN, 2006).

3.10.1. DEREK

O “software” **DEREK** (SANDERSON; EARNSHAW, 1991) foi utilizado na predição de toxicidade dos compostos sugeridos como potenciais ligantes do domínio KH3 da proteína hnRNP K. O “software” dispõe de um sistema que realiza predições do ponto de vista qualitativo e, dessa forma, alertas são gerados acerca da possível ação tóxica dos compostos químicos analisados. O sistema é capaz de interpretar subestruturas toxicofóricas presentes nos compostos como possíveis indutoras de certos tipos de toxicidade através das regras de correlação implementadas no “software”. As regras “knowledge-based” presentes no “software” **DEREK** operam em duas linguagens diferentes. A primeira é mais simples e faz uso do número de átomos e ligações para definir o grupo toxicofórico. A segunda linguagem é mais complexa e consegue responder questões a respeito da estrutura do grupo químico analisado (RIDINGS et al., 1996).

3.11. Dinâmica molecular

O estudo da dinâmica do movimento das moléculas é um atrativo para a química medicinal computacional. Pelo fato das técnicas modernas de cristalografia promoverem um excelente suporte na análise de estruturas moleculares estáticas, sejam elas de pequeno ou grande porte, a idéia de variações conformacionais está sempre presente. O reconhecimento do substrato pelas proteínas, o enovelamento de proteínas em suas conformações nativas e as reações químicas em geral, são processos inconcebíveis sem o conceito de flexibilidade molecular (DISCOVER USER GUIDE, 1993).

As simulações de dinâmica molecular constituem-se em uma importante estratégia para a exploração do espaço conformacional. O objetivo é reproduzir os

movimentos de uma molécula em função do tempo. As simulações de dinâmica molecular são baseadas nos conceitos físicos de mecânica molecular. Neste contexto, os átomos de uma molécula interagem com outros de acordo com as regras do campo de força empregado (HÖLTJE et al., 2003b). Em intervalos regulares de tempo, a equação de movimento representada pela segunda lei de Newton é resolvida:

$$-\frac{\partial U(x_i)}{\partial x_i} = F_i(t) = m_i a_i(t) = m_i \frac{\partial^2 x_i}{\partial t^2}$$

As simulações de dinâmica molecular resolvem a equação de movimento de Newton, em que F_i é a força sobre o átomo i no tempo t , m_i é a massa do átomo i , e $a_i \left(\partial^2 x_i / \partial t^2 \right)$ é a aceleração do átomo i no tempo t . O gradiente da função de energia potencial é usado para calcular as forças sobre os átomos, ao passo que a velocidade inicial dos átomos é gerada randomicamente no início da simulação. A força sobre o átomo i pode ser calculada diretamente pela derivada da energia potencial U com respeito às coordenadas x_i . Com uma expressão adequada para a energia potencial e massas conhecidas, é possível resolver a equação diferencial para futuras posições, que revelam uma trajetória ao longo do tempo. Baseadas nas coordenadas atômicas iniciais, novas posições e a velocidade dos átomos podem ser calculadas em um tempo t , logo, os átomos serão movidos para estas novas posições no espaço e uma nova conformação é criada. O ciclo, então, é repetido em um número pré-definido de etapas (DISCOVER USER GUIDE, 1993; HÖLTJE et al., 2003b). A energia total do sistema E é a somatório das contribuições das energias cinética e potencial (LANIG, 2003).

$$E = \frac{1}{2} m \left(\frac{\partial x}{\partial t} \right)^2 + U(x)$$

A temperatura é um conceito fundamental em uma simulação de dinâmica molecular. A temperatura é proporcional à energia cinética do sistema, que pode ser expressa em termos de velocidades atômicas. A justificativa para a relação entre temperatura e velocidade é promovida pela teoria cinética dos gases (DISCOVER USER GUIDE, 1993). Geralmente, as simulações são realizadas entre 300 K a 400 K. Se por um lado a temperatura deve ser suficientemente alta para prevenir o colapso do

sistema em determinada região do espaço conformacional, por outro lado não deve ser tão alta para resultar em conformações distorcidas de alta energia (LANIG, 2003). A temperatura T do sistema é relacionada ao meio da energia cinética do sistema de todos os átomos N , em que k_b é a constante de Boltzmann e $\langle v_i^2 \rangle$ a média das velocidades ao quadrado de todos os átomos i .

$$\frac{1}{2} \sum m_i \langle v_i^2 \rangle = \frac{3}{2} N k_b T$$

Diferentemente dos procedimentos de otimização de energia, as simulações de dinâmica molecular são capazes de transpor as barreiras de energia entre conformações diferentes. Para aumentar a amostragem conformacional, freqüentemente, altas temperaturas são aplicadas à simulação. A elevadas temperaturas, as moléculas são capazes de transpor até mesmo grandes barreiras de energia que podem existir entre algumas conformações. Logo, as chances para uma busca conformacional completa aumentam (HÖLTJE et al., 2003b).

Embora os recursos computacionais tenham se tornados cada vez mais robustos para lidar com sistemas moleculares grandes (até 50000 átomos, por exemplo), ainda é necessária a introdução de algumas simplificações para reduzir o tempo exigido para a realização dos cálculos. Uma grande vantagem do emprego de simplificações no sistema é o fato de que elas abrem a possibilidade da escolha de períodos de tempo mais longos para a realização da simulação, o que oferece uma observação mais completa do comportamento de sistemas macromoleculares. A realização destas modificações e a redução do número de graus de liberdade precisam ser checadas cuidadosamente, pois tais modificações podem levar o modelo a uma carência de exatidão (HÖLTJE et al., 2003a).

Um procedimento simplificado muito comum é o uso de funções de energia potencial de átomos unidos (do inglês, “united atom potencial energy functions”). A maioria dos campos de força em modelagem molecular de proteínas, como AMBER e GROMOS, são baseados nestes algoritmos. A omissão de hidrogênios não polares em um campo de força dessa categoria reduz significativamente o número de partículas em uma biomacromolécula. Uma outra possibilidade, para reduzir o tempo exigido para a realização dos cálculos, é promovida pelo algoritmo SHAKE. Nesse procedimento, forças adicionais são determinadas para os átomos, com o objetivo de manter o

comprimento das ligações em valores fixos de equilíbrio. Logo, os termos de energia de estiramento das ligações não seriam calculados para ligações rígidas (HÖLTJE et al., 2003a).

3.11.1. Simulações de dinâmica molecular

As simulações de dinâmica molecular foram realizadas com o módulo **Discover_3** do pacote computacional **Insight II** (INSIGHT II USER GUIDE, 2005) para os ligantes selecionados nas simulações de screening virtual em complexo com o domínio KH3. As geometrias iniciais dos ligantes e do domínio KH3 foram as mesmas obtidas no procedimento de “rescore” das simulações de “docking”, bem como para a estrutura de KH3 com a sequência oligonucleotídica depositada no PDB (código 1J5K), incluindo a adição de aproximadamente 400 moléculas de água a partir de um raio de 20 Å, tendo como centro o átomo de carbono delta 1 da cadeia lateral do resíduo de I49, criando um ambiente solvatado. Para a realização das simulações, os potenciais eletrostáticos foram estabelecidos para o sistema, seguido pelo processo de minimização de energia dos complexos solvatados ligante-KH3. A energia dos complexos foi minimizada utilizando 1500 passos de um protocolo combinado de algoritmos steepest descent/ gradiente conjugado. Para cada complexo analisado foi gerada uma trajetória, com um tempo simulado de 1500 ps a uma temperatura de 298 K. As coordenadas do sistema foram salvas a cada 1 ps do tempo simulado, gerando, assim, 1500 coordenadas de conformações para cada complexo. Para cada trajetória foram analisados os valores de energia total do sistema, o RMSD da conformação dos ligantes e o RMSD do contato dos ligantes com os resíduos de R59 e R40 do domínio KH3.

O RMSD (desvio de mínimos quadrados) é uma medida frequentemente usada para discriminar as diferenças de valores entre um modelo e um sistema estimado. Nas simulações aqui realizadas o RMSD representa a medida das distâncias dos átomos das estruturas iniciais em comparação com as coordenadas das 1500 conformações geradas nas simulações de dinâmica molecular para cada complexo KH3-ligante analisado. Dados dois grupos ($\mathbf{v}$ e $\mathbf{w}$) de n pontos, o RMSD pode ser definido como:

$$\text{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2}$$

As simulações foram realizadas em dois campos de força distintos, em virtude da natureza estrutural diferenciada dos ligantes (selecionados nas simulações de “screening” virtual) e da sequência oligonucleotídica. A simulação do complexo contendo a estrutura oligonucleotídica foi realizado com o campo de força AMBER, um dos campos de força mais populares e bastante apropriado para estruturas de ácidos nucléicos. Em relação às simulações dos complexos dos ligantes com KH3, o campo de força CVFF foi utilizado. O campo de força CVFF, que se utiliza de procedimentos de mecânica quântica, foi desenvolvido para calculo de energias e frequencias vibracionais de estruturas proteicas e pequenas moleculas orgânicas (LANIG, 2003).

4. RESULTADOS E DISCUSSÃO

4.1. Busca de seqüências homólogas

A busca de seqüências homólogas foi realizada com o “software” **BLAST** (www.ncbi.nlm.nih.gov/BLAST/), de maneira individual para a seqüência de cada um dos três domínios KH das duas isoformas da proteína hnRNP K. O PDB foi o banco de dados selecionado para a realização dessa busca. Para maior confiabilidade e segurança das análises posteriores, foram selecionadas, apenas, seqüências com identidade seqüencial igual ou superior a 30%, e que possuem resolução acima de 2,5 Å quando resolvidas por cristalografia de raios-X. Na tabela 4 estão descritas as proteínas selecionadas pelo **BLAST** para cada um dos domínios KH, com seus respectivos códigos PDB e a identidade seqüencial calculada pelo “software”.

Tabela 4. Descrição das seqüências selecionadas na busca com o **BLAST**, e seus respectivos códigos PDB, com os valores de identidade seqüencial obtidos. Para as estruturas resolvidas por cristalografia de raios-X é indicada a resolução, e para as estruturas resolvidas por ressonância magnética nuclear é indicado RMN.

Domínios KH	Código PDB e descrição da seqüência	Identidade seqüencial	Resolução
KH1	2CXC: fator de transcrição NusA de arquobactérias	40%	2,00 Å
	1J5K: domínio KH3 da hnRNP K	36%	RMN
	1ZZI: domínio KH3 da hnRNP K.	36%	1,80 Å
	1KHM: domínio KH3 da hnRNP K	34%	RMN
	1X4M: domínio KH da proteína <i>far upstream element binding 1</i>	33%	RMN
	2AXY: domínio KH1 da proteína <i>poly(C) binding protein 2</i>	31%	1,70 Å
KH2	1J5K: domínio KH3 da hnRNP .	36%	RMN
	1ZZI: domínio KH3 da hnRNP K	36%	1,80 Å
	1WVN: domínio KH da proteína <i>poly(C) binding protein 1</i>	35%	2,10 Å
	1KHM: domínio KH3 da hnRNP K	34%	RMN
	1X4M: domínio KH da proteína <i>far upstream element binding 1</i>	31%	RMN
	2AXY: domínio KH1 da proteína <i>poly(C) binding protein 2</i>	30%	1,70 Å
KH3 (Isoforma a)	1J5K: domínio KH3 da hnRNP K	98%	RMN
	1ZZI: domínio KH3 da hnRNP K	98%	1,80 Å
	1KHM: domínio KH3 da hnRNP K	97%	RMN
	1WVN: domínio KH da proteína <i>poly(C) binding protein 1</i>	50%	2,10 Å
	2AXY: domínio KH1 da proteína <i>poly(C) binding protein 2</i>	47%	1,70 Å
KH3 (Isoforma b)	1J5K: domínio KH3 da hnRNP K	100%	RMN
	1ZZI: domínio KH3 da hnRNP K	100%	1,80 Å
	1KHM: domínio KH3 da hnRNP K	98%	RMN
	2AXY: domínio KH1 da proteína <i>poly(C) binding protein 2</i>	47%	1,70 Å
	1WVN: domínio KH da proteína <i>poly(C) binding protein 1</i>	46%	2,10 Å

Para uma análise mais refinada e com maior rigor, foi feito o “download”, no PDB, das estruturas protéicas obtidas através da busca com o **BLAST** (ALTSCHUL et al., 1990), e as pertencentes a cada um dos quatro grupos (KH1, KH2, KH3 isoforma a e KH3 isoforma b) foram sobrepostas uma em relação às outras com auxílio do “software” **SPDB Viewer**, e posteriormente visualizadas no software **DS ViewerPro5.0**. O objetivo dessa sobreposição foi, exatamente, o de verificar se o enovelamento das proteínas selecionadas era, realmente, característico de um domínio KH. Essa análise foi facilitada pelo fato da maioria das proteínas selecionadas pertencerem, efetivamente, ao grupo de proteínas com domínios KH. Nessa sobreposição, apenas as estruturas com códigos PDB 2CXC e 1X4M apresentaram um enovelamento distante do padrão esperado para a realização da modelagem molecular por homologia estrutural, apesar da identidade sequencial de 2CXC ter sido a maior em relação ao domínio KH1. Dessa maneira, as estruturas 2CXC e 1X4M foram descartadas.

4.2. Alinhamento múltiplo

O alinhamento múltiplo das seqüências selecionadas foi realizado para cada uma das seqüências homólogas dos domínios KH com o “software” **Multalign**, pertencente ao pacote computacional **AMPS** (BARTON; STERNBERG, 1987), com exceção da isoforma b do domínio KH3, que apresentou na busca com o **BLAST** (ALTSCHUL et al., 1990) seqüências com identidade de 100%. O “software” realiza, primeiramente, um alinhamento global entre todas as seqüências, e nesse alinhamento foi obtido um valor de identidade sequencial para cada um dos pares de seqüências, que foi comparado com o valor obtido pelo **BLAST**, como mostrado na tabela 5. Dessa forma, o alinhamento entre as seqüências foi realizado, em um primeiro momento (**BLAST**), de maneira local, funcionando como um processo de triagem, e em um segundo momento (**Multalign**), de forma mais robusta e específica compreendendo um alinhamento global, para confirmar, efetivamente, as melhores seqüências a serem empregadas na construção dos modelos por homologia.

Tabela 5. Comparação entre os valores de identidade seqüencial obtidos pelos softwares **BLAST** e **Multalign**.

Domínios	Códigos PDB	Identidade seqüencial BLAST	Identidade seqüencial MULTALIGN
KH1	1J5K	36%	23.94 %
	1ZZI	36%	23.94%
	1KHM	34%	22.54%
	2AXY	31%	29.58%
KH2	1J5K	36%	27.40%
	1ZZI	36%	27.40%
	1WVN	35%	28.38%
	1KHM	34%	27.40%
	2AXY	30%	31.43%
KH3 Isoforma a	1J5K	98%	95.06%
	1ZZI	98%	92.59%
	1KHM	97%	93.83%
	1WVN	50%	41.46%
	2AXY	47%	32.88%
KH3 Isoforma b	1J5K	100%	100%
	1ZZI	100%	97.53%
	1KHM	98%	98.77%
	2AXY	47%	32.88%
	1WVN	46%	43.21%

Após o alinhamento aos pares, realizado pelo “software” **Multalign**, e a seleção das melhores seqüências, foi realizado o alinhamento múltiplo para as seqüências de cada um dos domínios KH e suas homólogas, com o objetivo de se obter o melhor alinhamento para a construção dos modelos estruturais. O “software”, então, cria uma ordem para o alinhamento das seqüências, dos pares mais semelhantes aos menos semelhantes, como mostrado nas Tabelas 6, 7 e 8.

Tabela 6. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH1, na ordem requisitada pelo “software” AMPS. Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI, 4 – 2AXY e 5 – seqüência alvo KH1.

	Alinhamento múltiplo – domínio KH1
1	GSHMSYGD LGGPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
2	GSPNSYGD LGGPIITTQVTIPKDLARSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
3	GAM GPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
4	KNVTLTIRLLMHGKEVGSIIIGKKGESVKMREESGARINISEGN CPERIITLAGPTNAIFKAFAMIIDKLEEDI
5	MVELRILLQSKNAGAVIGKGGKNIKALRTDYNASVSVPD SSGPERILSISADIETIGEILKKIIPTELEGL

Tabela 7. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH2, na ordem requisitada pelo “software” AMPS. Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI, 4 – 1WVN, 5 – 2AXY e 6 – seqüência alvo KH2.

	Alinhamento múltiplo – domínio KH2
1	GSHMSYGD LGGPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKI DEPLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
2	GSPNSYGD LGGPIITTQVTIPKDLARSIIGKGGQRIKQIRHESGASIKI DEPLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
3	GAM GPIITTQVTIPKDLAGSIIGKGGQRIKQIRHESGASIKI DEPLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSGKFF
4	PL GSQTTHELTIPNNLIGCIIGRQGANINEIRQMGAQIKI ANPVEGSSGRQVTITGSAASISLAQYLINARLSSEKMGCS
5	KNVTLTIRLLMHGKEVGSIIIGKKGESVKMREESGARINI SEGN CPERIITLAGPTNAIFKAFAMIIDKLEED
6	DCELRLLIHQSLAGGIIGVKGAKIKELRENTQTTIKLFQECPPHSTDRVVLIGGKPD RVVECIKIILDLISESP

Tabela 8. Alinhamento múltiplo entre as seqüências-molde extraídas do PDB e a seqüência-alvo do domínio KH3 (isoforma a), na ordem requisitada pelo “software” AMPS. Em que, 1 – 1J5K, 2 – 1KHM, 3 – 1ZZI e 4 – seqüência alvo da isoforma a do domínio KH3.

	Alinhamento múltiplo – domínio KH3 (isoforma a)
1	GSHMSYGD LGGPIITTQVTI PKDLAGSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSG KFF
2	GSPNSYGD LGGPIITTQVTI PKDLARSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSG KFF
3	GAM GPIITTQVTI PKDLAGSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYSG KFF
4	LG GPIITTQVTI PKDLAGSIIGKGGQRIKQIRHESGASIKIDE PLEGSSEDRIITITGTQDQIQNAQYLLQNSVKQYAD VEGF

4.3. Construção dos modelos

Uma vez obtidos os alinhamentos entre as seqüências-alvo e as seqüências-molde, para os domínios KH1, KH2 e KH3 (isoforma a), a etapa seguinte compreendeu a construção dos modelos estruturais através da modelagem molecular por homologia estrutural, dispondo-se do “software” **Modeller** (SALI; BLUNDELL, 1993).

O “software” **Modeller** gerou 3 modelos para cada um dos três domínios estudados, e um processo posterior de validação foi realizado, para que os modelos de melhor qualidade para cada domínio fossem filtrados. Apesar de todas as restrições impostas pelo software, em alguns casos os modelos podem apresentar maus contatos atômicos e enovelamentos incorretos (SALI; BLUNDELL, 1993).

4.4. Validação dos modelos

Os modelos gerados pelo “software” **Modeller** foram analisados por três “softwares”: **Procheck** (LAKOWSKI; MACATHUR; THORNTON, 1993), **Whatif** (VRIEND; SANDER, 1993) e **Verify 3D** (LUTHY; BOWIE; EISENBERG, 1992). O “software” **Procheck** avalia a qualidade estereoquímica dos modelos. Já **Whatif** avalia a qualidade dos modelos finais por análise dos contatos atômicos dos resíduos, e o “software” **Verify 3D** os ambientes químicos dos resíduos.

4.4.1. Domínio KH1

O “software” **Procheck** (LAKOWSKI; MACATHUR; THORNTON, 1993) promove uma extensa verificação com relação aos parâmetros estereoquímicos dos modelos protéicos. O “output” compreende vários gráficos, que concedem uma avaliação completa da qualidade estereoquímica do modelo em comparação com estruturas bem definidas no mesmo nível de resolução. As Figuras 6 e 7 mostram o nível de qualidade estereoquímica do modelo construído para o domínio KH1.

Idealmente, em relação ao gráfico de Ramachandran (Figura 6), a estrutura deve apresentar acima de 90% dos seus resíduos na região vermelha (A, B e L), desconsiderando os resíduos de glicina (não possuem cadeia lateral), prolina (o C α está ligado à cadeia lateral) e os resíduos das extremidades (carboxi-terminal e amino-terminal) que apresentam padrões estereoquímicos diferentes dos outros resíduos

(LASKOWSKI, MACARTHUR; THORNTON, 1993). Em relação a esse critério, o melhor modelo gerado para o domínio KH1 apresentou um valor de 94,9%. O referido modelo do domínio KH1 apresentou suas respectivas propriedades estereoquímicas da cadeia principal (Figura 7), sempre dentro da margem ou em melhores condições que os parâmetros de estruturas protéicas do PDB com nível de resolução estrutural semelhante, ressaltando a qualidade estereoquímica total do modelo, representada pelo fator-G, que foi acima da média.

N° de resíduos nas regiões mais favoráveis [A,B,L]	56	94,9%
N° de resíduos em regiões adicionalmente permitidas [a,b,l,p]	2	3,4%
N° de resíduos em regiões generosamente permitidas [~a,~b,~l,~p]	1	1,7%
N° de resíduos em regiões desfavoráveis [branco]	0	0%
N° de resíduos não-glicina e não-prolina	59	100%
N° de resíduos em C e N-terminal (exceto glicina e prolina)	2	
N° de resíduos de glicina (triângulos)	7	
N° de resíduos de prolina	3	
N° total de resíduos	71	

Figura 6. Gráfico de Ramachandran do modelo do domínio KH1, gerado pelo software **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). Os resíduos de glicina (7 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu C α não apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados.

Figura 7. Representação dos gráficos de cinco propriedades estruturais da cadeia principal. Os valores do modelo do domínio KH1 são marcados por quadrados e comparados com estruturas bem definidas com resolução estrutural similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média.

O módulo **Coarse Packing Quality Control** do “software” **Whatif** (VRIEND; SANDER, 1993) apresenta a possibilidade de analisar os contatos atômicos de qualquer tipo utilizando como referência comparativa estruturas depositadas no PDB. O software realiza o cálculo do chamado “índice da qualidade de contato”. O valor médio do índice da qualidade de contato para o modelo protéico pode ser interpretado da seguinte maneira:

- maior que -0,5: ótimo modelo.
- em -0,5: valor médio para um bom modelo.
- entre -1,0 e -0,5: ainda um bom modelo.
- em -1,5: ainda bom, mas com pequenos erros.
- em -2,0: modelo considerado pobre.
- em -3,0: modelo ruim.

Os índices da qualidade de contato calculados para cada um dos resíduos do modelo do domínio KH1, bem como o índice global do modelo, estão descritos na

Tabela 9. Normalmente, a escala do índice da qualidade de contato para cada resíduo abrange valores entre -5 e 5. Um valor menor que -5 pode significar algum tipo de erro, como: empacotamento improvável ou coordenadas atômicas incorretas (VRIEND; SANDER, 1993).

Tabela 9. Valores dos índices da qualidade de contato para todos os resíduos do modelo estrutural gerado para o domínio KH1 e o índice total do modelo.

1 MET (1): -4.644	25 LYS (25): 2.072	49 ILE (49): 6.158
2 VAL (2): 0.985	26 ALA (26): 1.268	50 SER (50): 0.246
3 GLU (3): -0.750	27 LEU (27): 0.912	51 ALA (51): 0.312
4 LEU (4): -1.248	28 ARG (28): 0.274	52 ASP (52): 3.925
5 ARG (5): -2.976	29 THR (29): -1.069	53 ILE (53): -0.075
6 ILE (6): 1.471	30 ASP (30): -2.766	54 GLU (54): 0.177
7 LEU (7): -2.133	31 TYR (31): -4.245	55 THR (55): -1.343
8 LEU (8): 1.916	32 ASN (32): -1.442	56 ILE (56): 1.297
9 GLN (9): 0.486	33 ALA (33): 2.168	57 GLY (57): 1.519
10 SER (10): -0.266	34 SER (34): 0.367	58 GLU (58): 2.464
11 LYS (11): -1.178	35 VAL (35): -0.126	59 ILE (59): 0.995
12 ASN (12): -2.792	36 SER (36): -0.365	60 LEU (60): 1.789
13 ALA (13): 0.981	37 VAL (37): -1.524	61 LYS (61): 0.018
14 GLY (14): -0.024	38 PRO (38): -1.465	62 LYS (62): -0.540
15 ALA (15): -1.359	39 ASP (39): -4.669	63 ILE (63): -1.315
16 VAL (16): -2.778	40 SER (40): -4.387	64 ILE (64): 0.452
17 ILE (17): -1.936	41 SER (41): -4.876	65 PRO (65): -1.282
18 GLY (18): -3.224	42 GLY (42): -0.632	66 THR (66): -1.410
19 LYS (19): -4.075	43 PRO (43): -1.923	67 LEU (67): -1.400
20 GLY (20): -4.298	44 GLU (44): -2.262	68 GLU (68): -4.950
21 GLY (21): -1.183	45 ARG (45): 0.537	69 GLU (69): -4.979
22 LYS (22): 1.881	46 ILE (46): 0.885	70 GLY (70): -3.684
23 ASN (23): -0.011	47 LEU (47): 5.212	71 LEU (71): -7.414
24 ILE (24): 1.574	48 SER (48): 2.542	

Índice do modelo: -0,711

O resíduo de LEU71 do modelo (Tabela 9) apresentou um valor abaixo de -5, mas isso não significa, necessariamente, que este resíduo esteja incorreto. Pois, resíduos pequenos realizam menos contatos que resíduos grandes, logo, seus índices da qualidade de contato são pequenos, mesmo quando empacotados corretamente. Da mesma forma,

resíduos que se encontram na superfície das proteínas fazem poucos contatos quando comparados com resíduos que se encontram no interior das proteínas, logo, é esperado que tais resíduos apresentem, também, menor valor para o índice de qualidade. No caso do resíduo LEU71, o mesmo se encontra na extremidade da proteína, fato que pode proporcionar poucos contatos atômicos com outros resíduos.

O método utilizado pelo “software” **Verify 3D** (LUTHY; BOWIE; EISENBERG, 1992) avalia a compatibilidade da estrutura do modelo protéico com a sua sequência, através do perfil 3D. Ou seja, a posição de cada resíduo no modelo 3D é caracterizada pelo seu ambiente químico, e as preferências estatísticas para cada um dos aminoácidos são determinadas para cada um dos ambientes. A avaliação do modelo do domínio KH1 (Figura 8) encontra-se dentro dos níveis aceitáveis, com nenhum resíduo apresentando escore abaixo de zero, o que está em conformidade com os parâmetros estereoquímicos e de contato atômico descritos anteriormente. Dessa forma o referido modelo foi validado para ser utilizado em simulações posteriores. Vale lembrar que, antes disso, um protocolo de minimização de energia deve ser empregado.

Figura 8. Representação do perfil 3D do modelo do domínio KH1. Os valores dos 10 primeiros resíduos de cada extremidade são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore.

4.4.2. Domínio KH2

O gráfico de Ramachandran para o modelo do domínio KH2 (Figura 9) revelou que 95,3% de seus resíduos, desconsiderando glicina e prolina, se encontraram nas regiões mais favorecidas, o que garante confiabilidade em relação à qualidade dos ângulos torcionais da cadeia principal.

N° de resíduos nas regiões mais favoráveis [A,B,L]	61	95,3%
N° de resíduos em regiões adicionalmente permitidas [a,b,l,p]	3	4,7%
N° de resíduos em regiões generosamente permitidas [~a,~b,~l,~p]	0	0%
N° de resíduos em regiões desfavoráveis [branco]	0	0%
N° de resíduos não-glicina e não-prolina	64	100%
N° de resíduos em C e N-terminal (exceto glicina e prolina)	1	
N° de resíduos de glicina (triângulos)	6	
N° de resíduos de prolina	3	
N° total de resíduos	74	

Figura 9. Gráfico de Ramachandran do modelo do domínio KH2, gerado pelo “software” **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). Os resíduos de glicina (6 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu *C α* não apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados.

As propriedades estereoquímicas da cadeia principal do modelo molecular construído para o domínio KH2, em comparação com estruturas do PDB, são mostradas na Figura 10. As propriedades de planaridade de ligação peptídica, maus contatos atômicos e energia das ligações de hidrogênio se encontraram dentro da margem considerada ideal para estruturas com o mesmo nível de resolução estrutural, cerca de 1,8 Å. Já as propriedades de distorção dos carbonos α e o fator-G, que é uma média da

qualidade estereoquímica total do modelo, se mostram acima da média para um bom modelo.

Figura 10. Representação das propriedades estruturais da cadeia principal. Os valores do modelo do domínio KH2 são marcados por quadrados e comparados com estruturas bem definidas com resolução similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média.

A análise dos índices da qualidade de contato para o modelo construído do domínio KH2, mostra um bom nível de qualidade para contatos atômicos, em que o índice encontrado para o modelo (-0,577) está muito próximo do valor médio para um modelo de boa qualidade (-0,5). A tabela 10 descreve os valores do índice de qualidade para cada um dos resíduos do modelo. Três resíduos de aminoácidos apresentaram o índice da qualidade abaixo de -5.0, sendo eles GLU40, CYS41 e HIS44.

Tabela 10. Valores dos índices da qualidade de contato para todos os resíduos do modelo do domínio KH2 e o índice do modelo.

1 ASP (1): -4.107	26 GLU (26): 1.434	51 LEU (51): 3.544
2 CYS (2): -2.653	27 LEU (27): 1.043	52 ILE (52): 5.062
3 GLU (3): -1.364	28 ARG (28): -0.152	53 GLY (53): -1.259
4 LEU (4): 1.096	29 GLU (29): -1.252	54 GLY (54): -2.208
5 ARG (5): -4.198	30 ASN (30): -4.146	55 LYS (55): 1.505
6 LEU (6): 4.122	31 THR (31): -3.729	56 PRO (56): 0.464
7 LEU (7): -0.657	32 GLN (32): -2.778	57 ASP (57): 2.073
8 ILE (8): 1.271	33 THR (33): 0.061	58 ARG (58): 2.793
9 HIS (9): 2.869	34 THR (34): -1.418	59 VAL (59): 0.581
10 GLN (10): -0.345	35 ILE (35): 0.826	60 VAL (60): 0.546
11 SER (11): -2.005	36 LYS (36): -1.064	61 GLU (61): 1.935
12 LEU (12): -3.307	37 LEU (37): -0.678	62 CYS (62): 1.662
13 ALA (13): -0.624	38 PHE (38): -2.216	63 ILE (63): 0.852
14 GLY (14): 0.134	39 GLN (39): -3.091	64 LYS (64): 1.456
15 GLY (15): -1.709	40 GLU (40): -5.363	65 ILE (65): 1.467
16 ILE (16): -2.963	41 CYS (41): -5.332	66 ILE (66): 1.809
17 ILE (17): -2.208	42 CYS (42): -4.061	67 LEU (67): 0.890
18 GLY (18): -0.354	43 PRO (43): -2.991	68 ASP (68): 0.568
19 VAL (19): -3.256	44 HIS (44): -5.372	69 LEU (69): -0.028
20 LYS (20): -4.310	45 SER (45): -1.138	70 ILE (70): -0.158
21 GLY (21): -0.809	46 THR (46): -2.545	71 SER (71): -1.809
22 ALA (22): 1.367	47 ASP (47): -2.399	72 GLU (72): -5.061
23 LYS (23): 1.313	48 ARG (48): -0.358	73 SER (73): -2.859
24 ILE (24): 0.981	49 VAL (49): 4.420	74 PRO (74): -3.642
25 LYS (25): 1.751	50 VAL (50): 7.297	

Índice do modelo: -0577

A Figura 11 destaca a localização dos resíduos (em amarelo) que apresentam o índice da qualidade abaixo de -5.0 para o modelo do domínio KH2. Nessa ilustração, fica evidente que esses resíduos se localizam na superfície do domínio, sendo, então, perfeitamente aceitável o índice da qualidade de contato dos mesmos.

Figura 11. Localização dos resíduos (em amarelo) com baixo índice da qualidade de contato no modelo do domínio KH2.

Como nas avaliações anteriores, em relação a parâmetros estereoquímicos e contatos atômicos, a análise dos ambientes químicos, também apontou um bom nível de qualidade para o modelo do domínio KH2, o qual será utilizado em simulações posteriores. A avaliação do perfil 3D do modelo é apresentada na Figura 12.

Figura 12. Representação do perfil 3D do modelo do domínio KH2. Os valores dos 10 resíduos mais próximos de ambas as extremidades são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore.

4.4.3. Domínio KH3 (isoforma a)

O gráfico de Ramachandran gerado para o melhor modelo da isoforma a do domínio KH3 (Figura 13) revela que 98,5% de seus resíduos, desconsiderando glicina, prolina e os resíduos das extremidades, se encontram em regiões favorecidas. Os parâmetros estereoquímicos da cadeia principal para o modelo do domínio KH3 (Figura 14) também estão em consonância com os resultados obtidos no gráfico de Ramachandran e mostram boa qualidade para o modelo. Dessa forma, o fator-G obtido para o modelo se mostrou acima da média de qualidade para bons modelos.

N° de resíduos nas regiões mais favoráveis [A,B,L]	66	98,5%
N° de resíduos em regiões adicionalmente permitidas [a,b,l,p]	1	1,5%
N° de resíduos em regiões generosamente permitidas [~a,~b,~l,~p]	0	0%
N° de resíduos em regiões desfavoráveis [branco]	0	0%
N° de resíduos não-glicina e não-prolina	67	100%
N° de resíduos em C e N-terminal (exceto glicina e prolina)	2	
N° de resíduos de glicina (triângulos)	10	
N° de resíduos de prolina	3	
N° total de resíduos	82	

Figura 13. Gráfico de Ramachandran do modelo do domínio KH3, gerado pelo “software” **Procheck**, onde é feita uma correlação entre os ângulos torcionais da cadeia principal *Phi* e *Psi* para cada resíduo. As diferentes regiões são mostradas por cores e/ou tonalidades distintas (vermelho, amarelo e branco). Os resíduos de glicina (10 ao todo) possuem como cadeia lateral um átomo de hidrogênio, logo, seu *Cα* não

apresenta quiralidade e os resíduos são representados por triângulos, diferentemente dos resíduos convencionais, representados por quadrados.

Figura 14. Representação das propriedades estruturais da cadeia principal. Os valores obtidos do modelo do domínio KH3 (isoforma a) são marcados por quadrados e comparados com estruturas bem definidas com resolução similar. As bandas escuras em cada gráfico representam os resultados dessas estruturas bem definidas, em que a linha central representa uma média dos valores em função da resolução, e as linhas das extremidades o desvio em relação à média.

A análise dos índices da qualidade de contato para o modelo do domínio KH3 (isoforma a), revelou que o mesmo apresenta um padrão de empacotamento próximo de um bom modelo, com um índice global de -0,804. A Tabela 11 descreve os valores do índice de qualidade para cada resíduo do referido modelo.

Tabela 11. Valores dos índices da qualidade de contato para todos os resíduos do modelo do domínio KH3 (isoforma a) e o índice do modelo.

1 LEU (1): -7.372	29 LYS (29): 0.687	56 THR (56): 0.083
2 GLY (2): -4.615	30 GLN (30): 0.678	57 GLY (57): 0.101
3 GLY (3): -1.614	31 ILE (31): 1.551	58 THR (58): 4.153
4 PRO (4): -0.831	32 ARG (32): -0.850	59 GLN (59): 1.293
5 ILE (5): -4.246	33 HIS (33): -2.380	60 ASP (60): 1.694
6 ILE (6): -2.757	34 GLU (34): -4.907	61 GLN (61): 2.259
7 THR (7): -1.125	35 SER (35): -4.128	62 ILE (62): 1.920
8 THR (8): -0.023	36 GLY (36): 0.173	63 GLN (63): 1.319
9 GLN (9): -1.348	37 ALA (37): 0.803	64 ASN (64): 0.327
10 VAL (10): 2.319	38 SER (38): 0.065	65 ALA (65): 1.253
11 THR (11): -1.325	39 ILE (39): 2.748	66 GLN (66): 2.331
12 ILE (12): 1.263	40 LYS (40): -0.341	67 TYR (67): 0.806
13 PRO (13): 4.999	41 ILE (41): -0.507	68 LEU (68): 1.946
14 LYS (14): -0.467	42 ASP (42): -3.660	69 LEU (69): 1.049
15 ASP (15): -1.574	43 GLU (43): -3.630	70 GLN (70): -0.822
16 LEU (16): -1.648	44 PRO (44): -2.259	71 ASN (71): -1.822
17 ALA (17): 0.941	45 LEU (45): -5.524	72 SER (72): -1.127
18 GLY (18): -0.312	46 GLU (46): -2.918	73 VAL (73): -0.590
19 SER (19): -2.250	47 GLY (47): -2.465	74 LYS (74): -0.572
20 ILE (20): -3.324	48 SER (48): -0.651	75 GLN (75): -2.763
21 ILE (21): -2.886	49 GLU (49): -3.131	76 TYR (76): -3.042
22 GLY (22): -3.171	50 ASP (50): -2.044	77 ALA (77): -2.199
23 LYS (23): -3.863	51 ARG (51): -0.752	78 ASP (78): -1.945
24 GLY (24): -4.369	52 ILE (52): 1.005	79 VAL (79): -3.470
25 GLY (25): -1.159	53 ILE (53): 8.160	80 GLU (80): -3.119
26 GLN (26): 2.403	54 THR (54): 2.459	81 GLY (81): -3.994
27 ARG (27): 1.119	55 ILE (55): 3.366	82 PHE (82): -7.481
28 ILE (28): 0.442		
Índice do modelo: -0,804		

Os resíduos LEU1, LEU45 e PHE82 apresentaram seus índices de qualidade abaixo de -5.0. Isso é perfeitamente justificável, pelo fato de os resíduos LEU1 e PHE82 se encontrarem nas extremidades e o resíduo LEU45 na superfície do domínio (Figura 15).

Figura 15. Localização do resíduo LEU 45 (em amarelo) no modelo 3 do domínio KH3 (isoforma a).

O modelo construído para o domínio KH3 (isoforma a) mostrou-se dentro de um nível aceitável em relação à concordância entre cada resíduo e o seu respectivo ambiente químico, onde nenhum resíduo apresentou escore negativo, corroborando as análises estereoquímica e de contatos atômicos descritas previamente. O perfil 3D do modelo é mostrado na Figura 16.

Figura 16. Representação do perfil 3D do modelo da isoforma a do domínio KH3. Os valores dos 10 resíduos mais próximos de ambas as extremidades são desconsiderados e, por este motivo, se encontram no mesmo patamar de escore.

4.5. Análise do complexo KH3-ssDNA

A busca de seqüências homólogas com o software **BLAST** (ALTSCHUL et al., 1990) para o domínio KH3 (isoforma b) revelou identidade seqüencial de 100% para seqüências de estruturas depositadas no PDB. A resolução estrutural de complexos do domínio KH3 com ssDNA está disponível no PDB, sendo resultado de estudos que buscaram revelar a importância dos resíduos de aminoácidos na promoção de interações do domínio KH3 com os oligonucleotídeos de DNA e/ou RNA. Os códigos desses complexos depositados no PDB são: 1J5K e 1ZZI.

A estrutura com código PDB 1J5K, um complexo entre uma seqüência oligonucleotídica (TCCCT) e o domínio KH3, foi utilizada para a identificação espacial dos resíduos de aminoácidos que compõem o sítio ligante do domínio (Figura 17), servindo como um preâmbulo para simulações posteriores com os outros domínios, bem como para a realização de simulações de “screening” virtual em bases de dados para seleção de moléculas com potencial de se ligarem ao sítio ligante do domínio KH3.

Na análise visual do complexo, vários resíduos de aminoácidos do domínio podem ser identificados como importantes para a realização e manutenção de interações, e conseqüente reconhecimento da seqüência oligonucleotídica. Os aminoácidos que mais se destacam são: I29, K31, I36, K37, K48, R40, S46, I48, I49, R59 (Figura 17). Como o padrão de cores adotado neste trabalho, todas as figuras tridimensionais de compostos apresentados, incluindo oligonucleotídeos, proteína e ligantes serão mostrados da seguinte forma:

- A cor vermelha representa átomos de oxigênio
- A cor azul representa átomos de nitrogênio
- A cor rosa representa átomos de bromo
- A cor amarela representa átomos de enxofre
- A cor magenta representa átomos de fósforo
- A cor verde fluorescente representa átomos de cloro
- Os átomos com cores diferentes das supracitadas ou com indicação nas Figuras correspondem a átomos de carbono, que em uma mesma Figura podem adotar cores distintas de acordo com a representação de cada composto apresentado, para efeito de diferenciação.

Figura 17. Estrutura do domínio KH3 (código PDB 1J5K) em complexo com a sequência oligonucleotídica TCCCT. Os átomos de carbono dos resíduos de aminoácidos do sítio ligante da proteína estão indicados em verde e os da sequência nucleotídica em amarelo.

4.6. “Screening” virtual

As simulações de “screening” virtual foram realizadas com o “software” **GOLD** (VERDONK et al., 2003) para o domínio KH3 disponível no PDB (código 1J5K, complexo da isoforma b de KH3 com a sequência de ssDNA TCCCT). As bases de dados de compostos utilizadas foram: Ilibdiverse, IResearch Library e Chembridge. As bases de dados Chembridge e IResearch Library não foram utilizadas em sua plenitude ainda, mas sim para algumas de suas sub-coleções de compostos (Diverset, MolecularWeightset, CNSset e MicroFormats para Chembridge; Archive01 para IResearch Library) contendo estruturas com propriedades “drug-like” (Figura 18).

Em um primeiro momento, essas bases de dados foram utilizadas com o objetivo de selecionar compostos com potencial de se ligarem ao domínio KH3 através da abordagem de “docking” flexível, presente no “software” **GOLD** (VERDONK et al., 2003). Na primeira simulação, foi selecionada apenas a orientação de melhor escore para as 30 melhores estruturas filtradas pelo “software” para cada uma das coleções e/ou subcoleções de compostos. Dessa forma, os 30 melhores compostos de cada subcoleção foram submetidos a novas simulações, de maneira individual e de caráter mais refinado e criterioso, em que foram obtidas as 10 orientações de melhor escore para cada estrutura em relação ao sítio ligante do domínio KH3. Assim, houve a realização de um “rescore”, onde algumas estruturas foram descartadas e outras apresentaram orientações com interações favoráveis no sítio ligante do domínio KH3. O objetivo do “rescore” foi o de propor um modo de ligação para os compostos selecionados.

Figura 18. Bases de dados e suas respectivas subcoleções de compostos utilizadas nas simulações de “screening” virtual.

Para as moléculas selecionadas no “screening” virtual que mostraram maior escore de interação com o sítio ligante do domínio KH3 e que foram refinadas no procedimento de “rescore”, observou-se um *consensus* estrutural com a fita simples de DNA, em relação à presença de estruturas cíclicas e, em sua grande maioria, com substituintes e espaçadores polares. Esse padrão estrutural geral apresentado pela maioria dos compostos mimetiza os anéis nucleotídicos de pirimidina com espaçadores que contêm grupos fosfatos, como, por exemplo, na sequência oligonucleotídica TCCCT. As estruturas lineares dos compostos com melhor escore estão presentes nas Figuras 19, 20 e 21.

Figura 19. Fórmula estrutural dos compostos da base de dados Ilibdiverse que apresentaram maior escore nas simulações de “screening” virtual.

Figura 20. Fórmula estrutural dos compostos da base de dados IResearch Library que apresentaram maior “escore” nas simulações de “screening” virtual.

Figura 21. Fórmula estrutural dos compostos da base de dados Chembridge que apresentaram maior escore nas simulações de “screening” virtual e suas respectivas subcoleções de compostos.

Na Tabela 12 estão dispostos os valores de escore obtidos nas simulações de “docking” flexível (procedimento de “rescore”) e o nome IUPAC dos compostos selecionados. A comparação entre os valores de escore obtidos para os quinze compostos selecionados não se torna válida ainda em relação à predição de atividade, já que não existe até o presente momento uma série de compostos ativos descritos e validados que possam competir com a fita oligonucleotídica pela ligação ao domínio KH3 da proteína hnRNP K. Dessa forma, não é possível dizer ainda que dos quinze compostos selecionados os que apresentam maiores valores de escore sejam os mais potentes. Pode-se observar, somente, que dentre os compostos selecionados os maiores valores de escore significam um maior potencial de interação com o sítio ligante do domínio, do ponto de vista de energia das interações sugeridas pelas simulações de “docking”. As simulações também foram realizadas com o oligonucleotídeo de DNA TCCCT, em que o valor de escore obtido foi de 40,01.

Tabela 12. Nome IUPAC e valores obtidos pela função Goldscore nas simulações de “docking” flexível dos quinze compostos selecionados nas simulações de “screening” virtual.

Compostos	Nome IUPAC	Goldscore
1	(<i>E</i>)-3-(2-clorobenzamino)- <i>N'</i> -(4-metoxibenzilideno)benzohidrazida	33,98
2	2-(2-iltio-4,5-dihidrotiazol)-4,6-di(1-il-piperidina)-1,3,5-triazina	46,08
3	<i>N</i> -(4-6-metil-4-oxo-4 <i>H</i> -benzo[<i>d</i>] [1,3]2-il-oxazina)fenil)acetamida	31,47
4	2,6-bis((piridina-2-il)metilamino)pirimidina-4-ol	42,14
5	Cianometil-3-(5-((<i>E</i>)-(tetrahydro-3-(2-metoxifenil)-2,4,6-trioilideno)metil)furano-2-il)-4-metilbenzoato	39,28
6	(<i>E</i>)-2-(2-hidroxi-5-metilbenzilideno)benzo[<i>b</i>]tiofeno-3(2 <i>H</i>)-ona	31,20
7	(5 <i>E</i>)-7-((1 <i>R</i> ,2 <i>R</i> ,3 <i>R</i> ,5 <i>R</i>)-3,5-dihidroxi-2-((<i>E</i> ,3 <i>S</i> ,7 <i>R</i>)-3,7-dihidroxioceta-1-enil)ciclopentil) ácido 5-heptenoico	55,88
8	(<i>S</i>)-2-(benzamido(4-3(2-amino-6-il-3,4-dihidro-oxoquinazolina))ácido propanoico) ácido pentanodioico	39,01
9	4-(2-il-1,3-dioxoisindolina-2)ácido benzóico	27,03
10	1-(4-(5,6-dimetil-4-oxo-4 <i>H</i> -tieno[2,3- <i>d</i>] [1,3]2-il-oxazina)fenil)pirrolidina-2,5-diona	36,55
11	1-(4-(6-(1-il-2,5-dioxopirrolidina)-1 <i>H</i> -benzo[<i>d</i>]imidazol-2-il)fenil)2,5-pirrolidinodiona	32,65
12	2-(3-(6-bromo-4-oxo-4 <i>H</i> -benzo[<i>d</i>] [1,3]oxazina-2-il)fenil)-5-(3-nitrofenoxi)-isoindolina-1,3-diona	48,93
13	2-(1,3-dioxo-5-(4-oxo-4 <i>H</i> -benzo[<i>d</i>] [1,3]oxazina-2-il)-2-il-isoindolina)ácido benzóico	40,55
14	<i>N</i> -(4-(4-benzoilpiperazina-1-il)fenil)-2,3-dihidrobenzo[<i>b</i>] [1,4]dioxano-6-carboxamida	37,92
15	2-(5-benzoil-1,3-dioxoisindolina-2-il)ácido benzóico	35,56

Os resultados das simulações de “docking” (Figura 22) corroboram a relevância do papel do resíduo de R59 no reconhecimento de fitas de DNA e possíveis ligantes do domínio KH3 da proteína hnRNP K (Backe et al., 2005; Braddock et al., 2002), assim

também sugerido pelos cálculos dos campos de interação molecular, apresentados posteriormente na seção 4.8. Os quinze compostos selecionados até o presente momento mostraram orientações, nas simulações de “docking” flexível, dentro da fenda estreita considerada primordial no reconhecimento dos oligonucleotídeos. Os resíduos mais internos dessa fenda apresentam características hidrofóbicas, região composta basicamente por resíduos de isoleucina. A região mais externa, incluindo as extremidades, é formada por resíduos com características hidrofílicas, onde se destacam resíduos de arginina, lisina e serina. As moléculas selecionadas apresentam certo ajuste no sítio ligante do domínio KH3, onde os sistemas de anéis dos compostos podem ser acomodados na superfície hidrofóbica do centro do sítio ligante e os grupos polares dos mesmos interagir com os resíduos de aminoácidos da superfície externa do domínio, especialmente com o resíduo de R59. Esses resultados sugerem que esse sítio ligante poderia ser explorado no desenvolvimento de protótipos que pudessem bloquear a atividade da proteína hnRNP K por competição com a fita de DNA.

Figura 22. Orientações de melhor escore dos compostos selecionados nas simulações de “screening” virtual com o sítio ligante do domínio KH3.

Os compostos 1 e 14 apresentaram um padrão semelhante de interação com o sítio ligante do domínio KH3, além de uma certa semelhança do ponto de vista estrutural. Os dois compostos são formados por um sistema de anéis, destacando-se basicamente anéis aromáticos para o composto 1 e, além dos aromáticos, um anel piperazina e um dioxano para o composto 14. Nesses dois compostos há a presença de espaçadores entre os anéis, formados por grupamentos amida, que se mostram importantes, através de seus átomos de oxigênio carbonílico, na formação de interações íon-dipolo com a porção de guanidina do resíduo de R59. As orientações dos compostos 1 e 14 são mostradas na Figura 23.

Figura 23. Orientação dos compostos 1 e 14 no sítio ligante do domínio KH3, representados por A e B, respectivamente. A orientação do composto 1 (átomos de carbono em azul) é mostrada em comparação com a orientação do oligonucleotídeo TCCCT (carbonos em magenta) no complexo depositado no PDB (código 1J5K). As regiões circuladas destacam os átomos de oxigênio carbonílico dos grupamentos amida dos compostos em torno do resíduo de R59 da proteína.

Os compostos 3, 12 e 13 apresentam em comum um anel benzoxazina em sua estrutura, o qual é responsável, segundo as orientações sugeridas nas simulações de “docking”, por manter interações com o grupo guanidina do resíduo de R59 da proteína. Esses resultados revelam que derivados de compostos com grupos benzoxazina podem ser viáveis, do ponto de vista estrutural, na manutenção de interações com o domínio KH3. Dessa forma, o anel oxazina poderia ser uma subestrutura importante na realização de interações com o sítio ligante do domínio KH3, se tornando em um componente estrutural inicial na busca e planejamento de substâncias ativas. As orientações dos

compostos 3, 12 e 13, obtidas nas simulações de “docking”, são apresentadas na Figura 24. Já o composto 10 apresenta uma variante do anel benzoxazina, onde existe um anel tienoxazina, que também se mostrou capaz de interagir com o resíduo de R59 através das simulações de “docking”. Nessas orientações, os átomos de oxigênio carbonílico e os heteroátomos de oxigênio presentes nos anéis oxazina supracitados, mostraram grande potencial de realizar interações de caráter polar com o resíduo R59. Além disso, o composto 10 apresentou uma interação adicional entre a carbonila de seu anel pirrolidinodiona e o resíduo de K22.

Figura 24. Orientação dos compostos 3, 12, 13 e 10 no sítio ligante do domínio KH3, representados por A, B, C e D, respectivamente. A orientação do composto 3 (átomos de carbono em amarelo) é mostrada em comparação com a orientação do oligonucleotídeo TCCCT (carbonos em magenta) do complexo depositado no PDB (código 1J5K). As regiões circuladas destacam os átomos de oxigênio dos compostos que interagem com o resíduo de R59 do domínio KH3. Em D a linha tracejada representa uma interação entre o composto 10 e o resíduo de K31.

Os compostos 9 e 15, apesar de serem extraídos de bases de dados diferentes são compostos bastante semelhantes e apresentam um núcleo estrutural comum formado por

uma dioxoisindolina ligada a um grupamento de ácido benzóico. A diferença entre os dois compostos se deve à presença de um grupo benzoil no composto 15 e o posicionamento do grupamento carboxílico no anel aromático de ácido benzóico. A semelhança dos dois compostos também se mostrou evidente no modo de ligação sugerido pelas simulações de “docking”, onde em ambos os compostos o grupamento carboxílico se mostrou importante na manutenção de interações iônicas com o resíduo de R59. Do ponto de vista das interações e dos valores de escore obtidos, o composto 15 parece ser um composto com maior potencial de interação, pois o posicionamento de seu grupamento carboxilato na posição *orto* favorece a interação de uma das carbonilas de seu anel dioxoisindolina com o resíduo de R59. As orientações de melhor escore dos compostos 9 e 15 estão presentes na Figura 25.

Figura 25. Orientações de melhor escore dos compostos 9 e 15 no sítio ligante do domínio KH3, representados por A e B, respectivamente. As regiões circuladas destacam os átomos de oxigênio carboxílico e/ou carbonílico dos compostos em torno do resíduo de R59 da proteína.

Os compostos 2 e 4 também apresentam similaridade estrutural entre si, mas suas respectivas orientações se mostraram distintas (Figura 26). As simulações sugerem que o composto 4 busca uma orientação em torno do resíduo de R59, onde dois átomos de nitrogênio se localizam muito próximos de regiões que contém átomos de nitrogênio na sequência oligonucleotídica. A orientação do composto 2 não se apresenta tão próxima do resíduo de R59 como nos outros compostos, e o anel central dessa molécula está localizado acima do resíduo de I49. O interessante dessa orientação diferenciada é o posicionamento do átomo de nitrogênio do anel tiazol próximo ao resíduo de R40, que

também se mostra importante no reconhecimento de fitas de DNA (Backe et al., 2005; Braddock et al., 2002).

Figura 26. Orientações de melhor escore dos compostos 2 e 4 em comparação com a orientação do oligonucleotídeo TCCCT. (A) composto 2, com destaque para o posicionamento do anel tiazol próximo ao resíduo de R40. (B) composto 4 (carbonos em rosa) ao redor do resíduo de R59, com destaque para a posição dos átomos de nitrogênio que quase se sobrepõem aos átomos de nitrogênio da citosina 2 da sequência oligonucleotídica TCCCT (carbonos em magenta).

As orientações de maior escore dos compostos 5, 8 e 11 (Figura 27) revelaram que a principal interação dos mesmos com o resíduo de R59 é realizada com grupamentos de carbonila presentes em anéis. Esses grupamentos carbonila tem um potencial de interação muito grande com o resíduo de arginina, onde interações do tipo íon-dipolo podem ser formar com a região guanidina do resíduo.

Figura 27. Orientações de melhor escore dos compostos 5, 8 e 11, representados por A, B e C, respectivamente. As regiões circuladas destacam os átomos de oxigênio carbonílico dos compostos que interagem com o resíduo R59.

A orientação sugerida para o composto 6 é mantida basicamente por interações hidrofóbicas. Um de seus anéis, inclusive, se encontra posicionado entre os resíduos de I49 e P52. A orientação de maior escore do compostos 6 pode ser visualizada Figura 28. O composto 7, que pertence à classe das prostaglandinas, apresentou a orientação de maior escore dentre os 15 compostos selecionados nas simulações de “screening” virtual. A orientação de maior escore do composto 7 é apresentada na Figura 29. O composto 7 foi o único dentre os compostos selecionados que conseguiu alcançar os resíduos de R40 e R59, com seus grupamentos carboxila e hidroxila, respectivamente. Além disso, seu anel central encontra-se posicionado em uma região favorável a interações hidrofóbicas, entre os resíduos de I49 e P52.

Vale ressaltar que vários compostos apresentaram um anel dioxoisindolina em sua estrutura (compostos 9, 12, 13 e 15), constituindo em uma subestrutura que poderia ser investigada posteriormente quanto à sua importância na ligação ao domínio KH3.

Figura 28. Orientação de maior escore do composto 6 no sítio ligante do domínio KH3.

Figura 29. Orientação de melhor escore do composto 7 (19(R)-hidroxiprostaglandinaF2a) no sítio ligante do domínio KH3. Em círculos estão destacados um grupamento hidroxila próximo a R59 e um grupamento carboxilato próximo a R40.

4.7. Propriedades físico-químicas

As propriedades físico-químicas relacionadas aos parâmetros da Regra dos Cinco, de Lipinski et al. (1997), foram calculadas e são mostradas na Tabela 13. Segundo a Regra dos Cinco, a maioria dos fármacos que apresentam biodisponibilidade por via oral obedece pelo menos três dos seguintes parâmetros: peso molecular menor que 500, LogP menor que 5, número de receptores de ligação de hidrogênio menor ou igual a 10 e número de doadores de ligação de hidrogênio menor ou igual a 5. Todos os 15 compostos selecionados nas simulações de “screening” virtual se enquadram nos parâmetros da Regra dos Cinco.

Tabela 13. Propriedades físico-químicas relacionadas à Regra dos Cinco dos 15 compostos selecionados nas simulações de “screening” virtual.

Compostos	Peso molecular	Nº receptores de lig. de H	Nº doadores de lig. de H	LogP
1	407.86	5	2	4.11
2	364.54	6	0	5.27
3	294.31	4	1	2.75
4	304.31	7	1	4.26
5	485.46	9	1	4.94
6	268.34	3	1	3.91
7	369.48	6	4	1.38
8	479.43	9	4	0.74
9	266.23	5	0	1.98
10	354.39	7	0	3.32
11	388.38	7	1	0.53
12	584.34	10	0	4.23
13	411.35	7	0	3.73
14	443.51	5	1	3.55
15	370.34	6	0	3.42

4.8. Campos de Interação Molecular

Os campos de interação molecular foram gerados a partir da estrutura do domínio KH3 (código PDB: 1J5K) para três grupos químicos de prova diferentes: hidrofóbico (DRY), oxigênio de carbonila e nitrogênio de amida. A utilização desses grupos de provas distintos tem por objetivo a definição de sítios receptores virtuais no sítio ligante do domínio KH3 para grupos com características químicas consideradas relevantes na realização de interações de ligantes com proteínas. O grupo de prova hidrofóbico identifica sítios na proteína que favorecem a acomodação, do ponto de vista energético, de porções hidrofóbicas de ligantes. O grupo de prova oxigênio de carbonila representa fragmentos de ligantes que podem agir como receptores de ligação de hidrogênio. O grupo de prova nitrogênio de amida identifica regiões na proteína que favorecem interações com regiões doadoras de ligação de hidrogênio em ligantes (PASTOR et al., 2000).

Considerando o grupo químico de prova hidrofóbico, as orientações dos compostos 3 e 6 mostraram ser capazes de posicionar regiões dos ligantes com características hidrofóbicas (anéis aromáticos) em pelo menos um dos sítios virtuais identificados pelos cálculos, como mostrado na Figura 30.

Figura 30. Orientações do oligonucleotídeo TCCC (A) e dos compostos 3 (B) e 6 (C) no sítio ligante do domínio KH3. As superfícies representam os sítios virtuais de interação hidrofóbica. As regiões dos ligantes mais próximas dos sítios hidrofóbicos são evidenciadas por círculos.

Os resultados obtidos com o grupo químico de prova oxigênio de carbonila mostram que os resíduos de R59 e R40 são os principais sítios responsáveis pela interação com grupos receptores de ligação de hidrogênio. As orientações dos compostos 1, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14 e 15 apresentam átomos de oxigênio capazes de receber ligações de hidrogênio do resíduo R59 ou até grupamentos carboxila que podem realizar interações iônicas. As figuras 31 e 32 mostram as orientações desses compostos e do oligonucleotídeo TCCC com o sítio ligante do domínio KH3, de acordo com as bases de dados de onde foram extraídos (Ilibdiverse, IResearch Library e Chembridge). Das orientações analisadas apenas a sequência oligonucleotídica é capaz de sobrepor átomos de oxigênio exatamente na região do sítio receptor virtual gerado pelo resíduo de R40.

Figura 31. Orientações do oligonucleotídeo TCCC (A) e dos compostos 1 (B), 3 (C), 5 (D), 7 (E) e 8 (F), pertencentes às bases de dados IResearch Library e Ilibdiverse, no sítio ligante do domínio KH3. As superfícies representam os sítios virtuais de interação. As regiões dos ligantes mais próximas dos sítios são evidenciadas por círculos.

Figura 32. Orientações dos compostos 9 (A), 10 (B), 11 (C), 12 (D), 13 (E), 14 (F) e 15 (G), pertencentes à base de dados Chembridge, no sítio ligante do domínio KH3. As superfícies representam os sítios

virtuais que favorecem interações polares. As regiões favoráveis dos ligantes mais próximas dos sítios são evidenciadas por círculos.

Os campos de interação molecular também foram computados para o grupo químico de prova nitrogênio de amida. Os resultados obtidos para este grupo de prova não se mostraram significantes, pois a fenda estreita do sítio ligante que comporta o oligonucleotídeo não apresentou nenhum sítio virtual capaz de receber ligações de hidrogênio. Apenas os resíduos de aminoácidos que estão fora dessa fenda apresentaram sítios virtuais dessa natureza. Dessa forma, nenhum dos quinze compostos apresentados, inclusive a sequência olinucleotídica, mostrou grupamentos químicos capazes de sobrepor as superfícies geradas para o grupo de prova em questão. Esse resultado corrobora o baixo número de grupos doadores de ligação de hidrogênio observados para os compostos selecionados nas simulações de “screening” virtual (Tabela 13).

4.9. Dinâmica molecular

Simulações de dinâmica molecular foram realizadas com os quinze compostos selecionados nas simulações de “screening” virtual no sítio ligante do domínio KH3 (código PDB: 1J5K). As conformações iniciais de partida para as simulações foram os próprios modos de ligação sugeridos nas simulações de “docking” molecular apresentadas na seção 4.6. O objetivo da realização dessas simulações foi o de avaliar a estabilidade energética e conformacional dos ligantes no domínio KH3, bem como a estabilidade das interações sugeridas com os resíduos de R40 e R59. Para a realização das simulações de dinâmica molecular, moléculas de água foram adicionadas ao perímetro espacial da estrutura do domínio KH3, simulando um ambiente solvatado, como mostrado na Figura 33 para o composto 7, que apresentou o maior escore nas simulações de “docking”.

Figura 33. Orientação do composto 7 no sítio ligante do domínio KH3. Os elementos coloridos em roxo representam os átomos de oxigênio das moléculas de água que foram adicionadas ao sistema.

4.9.1. Estabilidade energética

O comportamento energético dos complexos ligante-proteína foi avaliado de acordo com o cálculo da energia total do sistema na trajetória gerada simulando o tempo de 1500 ps. Os cálculos de energia foram realizados a cada 1,0 ps em relação ao tempo decorrido de simulação, gerando dessa forma 1500 valores de energia calculados para cada complexo (valores referentes à cada uma das 1500 conformações avaliadas). Os gráficos da energia total para cada um dos 15 compostos no sítio ligante do domínio KH3 são mostrados na Figura 34.

Figura 34. Gráficos da energia total em função do tempo de simulação dos 15 compostos selecionados nas simulações de “screening” virtual complexados ao domínio KH3.

O comportamento energético dos 15 sistemas analisados, formados por complexos entre os 15 compostos apresentados e o domínio KH3, mostrou um nível semelhante de variações energéticas, com um favorecimento de queda de energia com o decorrer da simulação. Não houve grandes variações de energia em nenhum dos quinze complexos analisados. Dessa forma, não há indícios de grandes variações conformacionais nos complexos, o que, de maneira geral, corrobora as orientações sugeridas pelas simulações de “docking”, que se encontram praticamente estabilizadas do ponto de vista energético. A queda de energia durante a realização das simulações sugere a busca de um estado energético e de um ajuste mais favorável por parte dos complexos. A trajetória do complexo formado pelo composto 11 com KH3 apresentou a menor variação energética dentre as conformações obtidas, de 1450 Kcal. Já o complexo entre o composto 8 e KH3 apresentou a maior variação energética, de 5000 Kcal. Os compostos 4, 6, 13 e 15 se destacam por terem mantido um nível de variabilidade energética mais próximo de uma constante na maior parte do tempo de simulação.

As simulações de dinâmica molecular também foram realizadas com a tétrade oligonucleotídica TCCC em complexo com o domínio KH3, estrutura depositada no PDB, com código 1J5K. Embora realizada em um campo de força diferente, o mesmo comportamento energético dos quinze compostos selecionados (Figura 34) foi observado. O Gráfico 2 mostra a variação energética em função do tempo para a tétrade oligonucleotídica complexada ao domínio KH3.

Gráfico 2. Variação da energia total do complexo DNA - domínio KH3 em função do tempo de simulação.

4.9.2. Estabilidade conformacional

A avaliação da estabilidade conformacional das orientações dos compostos no sítio ligante do domínio KH3 foi realizada através do cálculo do RMSD da trajetória das 1500 conformações de cada ligante geradas nas simulações de dinâmica molecular. A Figura 35 revela os gráficos do RMSD em função do tempo para cada um dos quinze compostos analisados. Os compostos 4, 5, 6, 7, 12, 13 e 15 mostraram um padrão de variação de conformações bastante estável em praticamente todo o tempo de simulação, mostrando-se, dessa forma, como orientações estáveis no sítio ligante do domínio KH3. As orientações dos compostos 2, 9, 10 e 11 possuem um padrão de trajetórias conformacional instável no início das simulações, mas que logo se estabilizam com o decorrer do tempo em baixos níveis de RMSD. Isso mostra que esses compostos buscam certo ajuste no sítio ligante e logo se estabilizam com valores de energia e variação de RMSD mais baixos. Em relação à estabilidade conformacional no sítio ligante do domínio KH3, as orientações dos compostos 1, 3, 8 e 14 se mostraram instáveis, sem um padrão definido ao longo das simulações. Dentre os quinze compostos apresentados, o composto 7 foi o que apresentou os maiores valores de RMSD, significando um maior nível de variação e movimento de seus átomos, que é compensado pela estabilidade alcançada ao longo da simulação. Embora o composto 7 pareça ser, do ponto de vista conformacional, o menos rígido dos quinze compostos, as interações são suficientemente fortes para manter sua estabilidade durante a trajetória. O composto 7 é formado por apenas um anel central com duas cadeias com pelo menos sete átomos de carbono. A capacidade de variação conformacional inerente a essas cadeias carbônicas pode induzir a formação de vários estados conformacionais para o composto 7, dependendo do meio ao qual o mesmo seja introduzido.

Figura 35. Gráficos referentes à variação dos valores de RMSD em função do tempo de simulação dos 15 compostos selecionados nas simulações de "screening" virtual em complexo com o domínio KH3.

Em relação aos valores de RMSD calculados para o oligonucleotídeo TCCC (Gráfico 3), observa-se pouco grau de variação conformacional até cerca de 1200 ps do tempo de simulação, indicando que a conformação inicial da fita simples de DNA, obtida experimentalmente (código PDB: 1J5K), esteja estabilizada pelas interações realizadas com o sítio ligante do domínio KH3. A partir de 1200 ps ocorre um aumento no nível de variação de RMSD, que para ser verificado como estável deveria ser investigado em um tempo de simulação maior que 1500 ps.

Gráfico 3. Variação dos valores de RMSD do complexo DNA - domínio KH3 em função do tempo de simulação.

4.9.3. Estabilidade das interações com R40 e R59

A avaliação da estabilidade das interações dos átomos dos ligantes com os resíduos de R49 e R50, sugeridas pelas simulações de “docking”, foram realizadas através do cálculo do RMSD da trajetória das 1500 conformações de cada um dos compostos e da tétrade oligonucleotídica TCCC (presente na estrutura com código PDB 1J5K) obtidas durante as simulações de dinâmica molecular. Esses dois resíduos foram escolhidos por serem considerados os mais importantes no reconhecimento de seqüências nucleotídicas por parte do domínio KH3 (Backe et al., 2005; Braddock et al., 2002). Dessa forma, a variação da distância entre os átomos dos ligantes e dos resíduos de R40 e R59, que apresentaram potencial de interação, foi mensurada durante as simulações de dinâmica molecular. A Figura 36 mostra os valores de RMSD obtidos

para a distância de interação entre a tétrede oligonucleotídica e o resíduo de R59 do domínio KH3 (Gráfico 3).

Figura 36. Avaliação da estabilidade da interação (indicada por uma linha tracejada) do resíduo de R59 com o elemento C2 da tétrede oligonucleotídica, através do cálculo dos valores de RMSD em função do tempo de simulação.

Na Figura 37 são mostrados os gráficos obtidos para os compostos 1 e 14, que interagem com o resíduo de R59 através de carbonilas de seus grupamentos amida. De acordo com os valores de RMSD obtidos para os dois compostos, fica evidente a estabilidade da interação sugerida para o composto 1 até aproximadamente 1000 ps. Em relação à interação do composto 14 com o átomo de nitrogênio do grupamento guanidina do resíduo R59, observa-se que a mesma apresenta um padrão de variação muito maior que o do composto 1, traduzido em maiores valores de RMSD, mas manteve-se estável por mais tempo durante a simulação, iniciando em um nível mais baixo de RMSD seguido por um aumento que tende a estabilizá-lo. Essas variações de ambos os compostos podem ocorrer devido à busca, pelo sistema, de um melhor ajuste, do ponto de vista energético, dos átomos envolvidos na interação.

Figura 37. Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 1 e R59. (B) composto 14 e R59.

Os modos de ligação sugeridos para os compostos 3, 10, 12 e 13 supõem que os mesmos, do ponto de vista teórico, podem interagir com o resíduo de R59 através de átomos de oxigênio presentes em anéis oxazina. A estabilidade dessas interações, por cálculo de RMSD, está presente na Figura 38. Os gráficos mostrados para os compostos 12 e 13 parecem conter as interações mais estáveis dos anéis oxazina com o resíduo de R59. O composto 3, embora não tenha apresentado um padrão constante e estável de variação, apresentou os mais baixos índices de RMSD, entre 0,03 e 0,36, resultando em baixo grau de moção dos átomos ao longo da simulação. Já a interação do modo de ligação do composto 10 com o resíduo de R59 mostra-se instável ao longo do tempo de simulação. Isso pode ter ocorrido ao alinhamento de seu anel oxazina com R59, que foi diferente do observado para os outros compostos. Nos compostos 3, 12 e 13 alinhamento dos anéis oxazina com o resíduo de R59 propicia interações com dois átomos de oxigênio, uma carbonila e outro heteroátomo do anel, deixando a interação mais estável, o que parece não ocorrer no modo de ligação sugerido para o composto 10. Essas indicações sugerem que o anel oxazina pode ser uma subestrutura importante na

realização de interações com o domínio KH3. Corroborando os dados obtidos nas simulações de “dinâmica” molecular, os compostos 12 e 13 foram os que apresentaram maior escore dentre os quatro compostos analisados (Tabela 12), possivelmente, pela estabilidade de suas conformações e suas interações no sítio ligante do domínio KH3.

Figura 38. Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 3 e R59. (B) composto 10 e R59. (C) composto 12 e R59. (D) composto 13 e R59.

Em relação aos compostos 9 e 15, a avaliação das interações iônicas de seus grupamentos carboxilato com o resíduo de R59 (Figura 39) sugere que, do ponto de vista teórico, as duas propostas são estáveis, com destaque para a interação do composto 9, variando em grau muito baixo ao longo da simulação (RMSD menor que 0,2). Isso pode ocorrer devido à diferença no posicionamento espacial dos grupamentos carboxilato nos dois compostos. No composto 9, o mesmo se encontra na posição *para* no anel benzênico, e no composto 2 na posição *orto*. Esse aspecto poderia influenciar, do ponto de vista estérico, no posicionamento ideal do grupamento carboxilato para realizar uma interação mais estável com o resíduo de arginina. Por ser um composto maior e ter possibilidade de realizar mais interações, o composto 15 apresentou maior escore nas simulações de “docking” (Tabela 12), o que, de certa forma, não o torna mais promissor que o composto 9, do ponto de vista teórico, em relação à manutenção da interação com o resíduo de R59.

Figura 39. Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 9 e R59. (B) composto 15 e R59.

Embora os compostos 2 e 4 apresentem certo grau de similaridade estrutural, os modos de ligação sugeridos para ambos mantêm interações com resíduos diferentes. Em relação à essas interações, fica evidente a maior estabilidade da interação realizada do nitrogênio do anel piridina do composto 4 com o resíduo de R59. Em relação a essa interação, os valores de RMSD se mantiveram no mesmo patamar durante toda a simulação, indicando um bom ajuste do composto no sítio ligante do domínio KH3. Vale ressaltar, que o átomo de nitrogênio do anel piridina do composto 4 encontra-se posicionado espacialmente na mesma região de um átomo de nitrogênio do resíduo de citosina 2 da sequência oligonucleotídica TCCCT (Figura 26).

Figura 40. Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 2 e R40. (B) composto 4 e R59.

Os compostos 5, 8 e 11 interagem com o resíduo de R59 através de grupamentos carbonila, e além deste o composto 8 apresenta um grupamento carboxilato com potencial de interação com o mesmo resíduo de arginina. A avaliação da estabilidade das interações sugeridas encontra-se disposta na Figura 41. É possível observar que nenhum dos três compostos apresentou interações em um nível satisfatório de estabilidade durante a simulação. Nesse caso, há a necessidade de se realizar uma simulação em tempo maior, com o intuito de verificar se as interações propostas mantêm um patamar de estabilidade em algum período da trajetória, indicando qual o melhor ajuste da interação.

Figura 41. Avaliação da estabilidade das interações sugeridas nas simulações de “docking” (indicadas por linhas tracejadas) através do cálculo dos valores de RMSD em função do tempo. (A) composto 5 e R40. (B) composto 8 e R59. (C) composto 11 e R59.

Em consonância com o fato de ter apresentado o maior grau de variação conformacional dentre os quinze compostos apresentados nas simulações de dinâmica molecular (Figura 35), a interação polar proposta do composto 7 com o resíduo de R59 no sítio ligante do domínio KH3 apresentou os maiores valores de RMSD (Figura 42), embora a interação seja favorável o suficiente para retomar o estado inicial. Isso pode ser influenciado pela falta de rigidez do composto 7, que é formado por apenas um anel central substituído em duas posições por cadeias carbônicas. O que de fato ocorre é que durante a realização da simulação algumas moléculas de água conseguem deslocar o composto 7 e realizar interações com o resíduo de R59, e além disso, alguns dos grupamentos polares do composto às vezes substituíam suas interações com a proteína por interações com as próprias moléculas de água. Por várias vezes isso ocorreu e por várias vezes a conformação do composto retomava um estado conformacional próximo do proposto nas simulações de “docking”, situações às quais os valores de RMSD se aproximam de 0,99 (Figura 42).

Figura 42. Avaliação da estabilidade da interação (indicada por uma linha tracejada) do resíduo de R59 com uma hidroxila do composto 7, através do cálculo dos valores de RMSD em função do tempo.

4.10. Predição de toxicidade

Os quinze compostos selecionados foram analisados quanto à presença de grupamentos toxicofóricos com o “software” DEREK (SANDERSON; EARNSHAW, 1991) e os alertas de toxicidade foram gerados para as subestruturas correspondentes em cada um dos mesmos.

4.10.1. Amidas e aminas aromáticas

O “software” **DEREK** (SANDERSON; EARNSHAW, 1991) identificou a presença de amidas aromáticas nas estruturas dos compostos 1, 3, 9, 10, 11 e 12, assim como a presença de dois grupamentos amina aromáticos no composto 2 (Figura 43). É plausível a associação de amidas e aminas aromáticas ao processo de carcinogênese em humanos, pois esse grupamento pode ser convertido a hidroxilamina por redutases, oxidades ou hidrolases na maioria dos tecidos endógenos. A hidroxilamina é reconhecida por ser um agente indutor do processo de carcinogênese (RIDINGS et al., 1996).

Vale ressaltar que os alertas gerados apenas indicam a presença de uma subestrutura com histórico de causar um determinado efeito tóxico, e não necessariamente que os compostos supracitados sejam carcinogênicos. Em relação a tais compostos, esse alerta não aparenta ser motivo de grande preocupação, pois tanto os grupamentos de amidas como as aminas aromáticas se encontram estericamente protegidas por outros anéis, o que de fato pode dificultar o acesso de enzimas. A grande

exceção fica a cargo do composto 3, onde seu grupamento amida se encontra na extremidade da estrutura e pode, perfeitamente, ser acessível às enzimas responsáveis pela conversão à hidroxilamina.

Figura 43. Subestruturas de amidas e aminas aromáticas responsáveis pelos alertas tóxicos gerados para os compostos 1, 2, 3, 9, 10, 11 e 12.

4.10.2. Fenóis, precursores fenólicos, hidrazidas e análogos de anidrido ácido

Compostos formados por fenóis e seus precursores, hidrazidas e derivados de anidrido ácido apresentam grande potencial de causar hipersensibilidade cutânea. Normalmente, essas subestruturas são capazes de reagir com proteínas cutâneas, mas apenas a presença desses grupamentos em uma molécula não a credencia a induzir tal efeito. Outros aspectos, inerentes às propriedades físico-químicas da molécula em questão, como por exemplo, capacidade de absorção percutânea, devem ser levadas em consideração (CRONIN; BASKETTER, 1994; ITOH, 1982; RYCROFT; WILKINSON, 1991). Dos quinze compostos selecionados nas simulações de “screening” virtual, os compostos 1, 3, 5, 6, 10, 12 e 13 apresentaram subestruturas que se encaixam no perfil

de sensibilizadores cutâneos. A figura 44 mostra os compostos que apresentam fenóis, precursores fenólicos e o grupamento hidrazida e a Figura 45 os compostos formados por anidridos ácidos (anéis oxazina).

Figura 44. Subestruturas de fenóis e precursores e do grupamento hidrazida, responsáveis pelos alertas de hipersensibilidade cutânea gerados para os compostos 1, 5 e 6.

Figura 45. Subestrutura básica de um anidrido ácido presente nos anéis oxazina dos compostos 3, 10, 12 e

4.10.3. Pirimidina

O alerta de toxicidade gerado para o composto 4 diz respeito à presença de um anel de pirimidina substituído, que corresponde ao anel central do composto (Figura 46). Alguns derivados de pirimidina mostram potencial carcinogênico, incluindo uracil e timidina. Os derivados de pirimidina destacam-se por sua capacidade de formar cálculos urinários em ratos e camundongos, que se desenvolvem com a saturação dos compostos na urina. A formação dos cálculos causa danos aos epitélios dos órgãos urinários, o que se constitui em um estímulo para a síntese de DNA nas células, resultando em formação de tumor (ARCOS; ARGUS, 1974).

Figura 46. Anel de pirimidina do composto 4, responsável pelo alerta de toxicidade gerado para o composto 4.

4.10.4. Nitrila

Além da presença de um precursor fenólico (Figura 44), o “software” **DEREK** (SANDERSON; EARNSHAW, 1991) gerou outro alerta de toxicidade para o composto 5. O alerta em questão diz respeito à presença de um grupamento nitrila formado por um éster de cianohidrina (Figura 47). Compostos com nitrila podem liberar cianido no metabolismo e desencadear efeitos tóxicos. Os ésteres de cianohidrina, em quase todos os casos, são hidrolisados e liberam cianido. O cianido afeta, virtualmente, todos os tecidos humanos, pois é capaz de se ligar às metaloenzimas e inativá-las. Seu principal efeito tóxico resulta da inativação da enzima citocromo oxidase, inibindo o processo de respiração celular (LEAVESLEY et al., 2008).

Figura 47. Éster de cianohidrina presente no composto 5, responsável pelo alerta de toxicidade gerado.

4.10.5. Precursores de anilina

O alerta de toxicidade gerado para o composto 8 está relacionado à presença de um precursor de anilina em sua estrutura (Figura 48). Compostos capazes de serem metabolizados ou hidrolizados para formar anilina são potenciais causadores de metahemoglobinemia. A espécie humana é uma das mais susceptíveis a esse efeito. Esses compostos de nitrobenzenos estão associados ao desenvolvimento de toxicidade no baço, pois a anilina consegue se ligar a proteínas dos eritrócitos, que são danificados e se acumulam no baço, podendo ocasionar a formação de tumores (BUS; POPP, 1987).

Figura 48. Precursor de anilina presente na estrutura do composto 14, responsável pelo alerta de toxicidade gerado.

4.10.6. Diarilcetona

O “software” **DEREK** identificou a presença de um grupamento diarilcetona na estrutura do composto 5 (Figura 49). O grupamento diarilcetona, geralmente, está associado ao desenvolvimento de fototoxicidade e fotoalergenicidade. Para que qualquer reação de natureza fotoquímica aconteça a luz precisa ser absorvida pelo composto químico. Depois da absorção pela pele do agente fotoalergênico, a excitação causada pela luz com comprimento de onda adequado leva à formação de espécies reativas (radicais livres), que podem reagir com proteínas encontradas na pele e induzir o desenvolvimento de uma inflamação cutânea (PENDLINGTON; BARRATT, 1990).

Estrutura de uma diarilcetona

Figura 49. Diarilcetona presente no composto 15, responsável pelo alerta de fototoxicidade gerado.

5. CONCLUSÕES

A busca de eficientes terapias para as doenças que acometem a humanidade é uma constante no meio científico. Há sempre a necessidade de introdução de novos fármacos no arsenal terapêutico, seja pela falta de eficiência dos fármacos atuais, pelo alto nível de toxicidade dos mesmos, pelo surgimento de novos processos patológicos, ou até mesmo pelo aumento do número de casos de uma doença considerada “antiga” em uma determinada população. Mas, talvez, o maior desafio não seja sempre a descoberta de novas moléculas promissoras, e sim a descoberta de novas propriedades e potenciais de moléculas já existentes.

Na linha de frente dessa batalha está o câncer, que se caracteriza por ser uma doença de caráter heterogêneo, por acometer diferentes sistemas com diferentes graus de crescimento, proliferação e periculosidade. O contexto atual para a busca de novos fármacos, seja para o câncer ou qualquer outra doença, exige um conhecimento apurado acerca da fisiopatologia da doença e do alvo terapêutico a que se deseja intervir, constituindo-se no planejamento racional, que é abordagem mais utilizada no presente momento.

Seguindo esse contexto, o projeto brasileiro Genoma Humano Câncer buscou identificar os genes expressos nos tipos de câncer com maior incidência no país. Os resultados levaram à identificação de milhares de genes, com destaque para câncer de cabeça e pescoço, mama e cólon. Dentro desse projeto se inclui o Projeto Genoma Clínico, que visa, justamente, o desenvolvimento de novas formas de diagnóstico e tratamento para o câncer, tendo como base inicial o conhecimento dos genes expressos. A partir do estudo aprofundado acerca desses genes, várias informações relevantes puderam ser obtidas, como, por exemplo, a identificação de marcadores de vários tipos de câncer, que podem se tornar atrativos alvos terapêuticos para o desenvolvimento de fármacos.

A aplicação da química computacional tem oferecido um excelente suporte para o desenvolvimento de novos fármacos. Com o poder computacional e a tecnologia disponível atualmente, pode ser realizado um direcionamento nos estudos, facilitado pela capacidade de predição virtual de interações e propriedades. As técnicas de química computacional, aliadas à bioinformática, permitem uma análise criteriosa do alvo terapêutico a ser estudado, bem como a construção de modelos por homologia para os alvos que ainda não têm estrutura resolvida.

A proteína hnRNP K foi identificada, recentemente, como um marcador para câncer, sendo super-expressa em câncer de cabeça e pescoço. Ela apresenta diversas funções e é encontrada nos mais diversos compartimentos celulares, interferindo, basicamente, no sistema de expressão gênica.

A primeira fase deste estudo foi composta por um levantamento bibliográfico extenso, com o intuito de identificar aspectos estruturais e funcionais relevantes da proteína hnRNP K, para a aplicação de quimioinformática e bioinformática na construção de modelos, como suporte estrutural para a identificação de potenciais ligantes.

Após a conclusão do levantamento bibliográfico, que continua a ser atualizado constantemente, lançou-se mão das técnicas de bioinformática, que foram de extrema importância em uma segunda fase desse estudo, ou seja, a própria construção de modelos para cada um dos três domínios KH da proteína hnRNP K, pois apenas o domínio KH3 de uma isoforma teve sua estrutura resolvida e depositada no PDB. O domínio KH3 chegou a ser considerado o domínio mais importante para as interações da proteína com o DNA, mas evidências recentes sugerem que as interações da proteína hnRNP K com ácidos nucleicos são mediadas cooperativamente pelos três domínios KH. Daí a importância de se trabalhar em uma abordagem com os três domínios.

Para a construção dos modelos, um trabalho prévio foi realizado, tal como a identificação de seqüências homólogas às dos domínios KH e o alinhamento das mesmas. Uma vez obtido o alinhamento entre as seqüências, iniciou-se, efetivamente, a construção dos modelos por homologia estrutural, tendo sido utilizadas apenas seqüências homólogas com estruturas resolvidas e depositadas no PDB. Dessa forma, foram gerados três modelos de baixa energia para cada um dos domínios estudados. Uma vez gerados os modelos, um extenso trabalho de validação para escolha do melhor modelo a ser utilizado em simulações posteriores foi realizado, em que foram levados em consideração parâmetros estereoquímicos, de contatos atômicos e o enovelamento.

A terceira fase deste estudo, que se encontra em andamento, seria, justamente, a realização de “screening” virtual com os modelos construídos e a estrutura do domínio KH3 disponível no PDB (códigoPDB: 1J5K) para identificação de moléculas que apresentam um bom perfil teórico de interação com os domínios KH. Até o presente momento, simulações de “screening” virtual foram realizadas com as seguintes bases de dados: Ilibidiverse, Chembridge e IResearch Library. Todas essas bases de dados

apresentam compostos ativos, fármacos, produtos naturais e até moléculas com propriedades “drug-like” validadas *in silico*. O número de compostos presentes nas bases de dados e utilizados nas simulações é de aproximadamente 330.000.

Para uma seleção mais refinada, os compostos identificados nas simulações de “screening” virtual passaram por um processo de “rescore”, utilizando a abordagem de “docking” flexível. Dessa forma, quinze compostos foram selecionados e aqui foram apresentados. Dos quinze compostos selecionados, apenas o composto 2 não apresentou nenhum átomo nas posições sugeridas pelos cálculos dos campos de interação molecular realizados no sítio ligante do domínio KH3, em relação aos três grupos químicos de prova analisados. Os compostos 3 e 5 apresentaram sobreposição com as superfícies dos MIFs em pelo menos dois dos grupos de provas analisados. Assim, em relação aos MIFs, quatorze dos quinze compostos apresentados se mostraram com potencial, do ponto de vista energético, de realizar interações com os resíduos do domínio KH3 responsáveis pelo reconhecimento de seqüências nucleotídicas.

As simulações de dinâmica molecular revelaram que as orientações dos compostos 4, 13 e 15 se mostraram estáveis dos pontos de vista energético, conformacional e das interações com R59 e R40, constituindo-se em compostos bastante promissores em relação à capacidade de ligação e manutenção de interações com o domínio KH3. Pode-se destacar também a estabilidade conformacional dos compostos 5, 6 e 12 no sítio ligante do domínio KH3. Em relação ao composto 12, o mesmo apresentou uma interação estável com o resíduo de R59.

Em relação à análise de toxicidade, somente nos compostos 7 e 8 não foram identificados grupamentos com características toxicofóricas. Os alertas de toxicidade mais alarmantes foram gerados para os compostos 3 (potencial carcinogênico), 4 (potencial de causar tumor em órgão urinários), 5 (potencial em liberar cianido no metabolismo) e 14 (potencial de causar metahemoglobinemia). Vários alertas foram gerados em relação à presença de grupamentos capazes de desenvolver efeitos na pele, como por exemplo, hipersensibilidade cutânea e fotoalergenicidade. Embora, considerando que o desenvolvimento de fármacos contra o câncer, em sua grande maioria, vise a obtenção de formulações administradas por via oral ou parenteral, efeitos de toxicidade na pele não são totalmente descartados. Uma vez que o medicamento seja administrado ao paciente e seja absorvido (no caso de formulações por via oral) e, dessa forma, se torne biodisponível e seja distribuído pela corrente sanguínea por todo o

organismo, nada impede, a não ser as próprias características físico-químicas do fármaco, que as moléculas alcancem glândulas sudoríparas e sejam expelidas na pele, tornando-se, assim, aptas a desenvolver efeitos de toxicidade cutânea.

Além da identificação de quinze compostos com potencial de interagir com o domínio KH3 da proteína hnRNP K, com maior destaque para os compostos 4, 12, 13 e 15, que se mostraram promissores em simulações de “docking”, campos de interação molecular e dinâmica molecular, o presente trabalho também foi apto a identificar possíveis subestruturas capazes de realizar interações com o domínio. Uma delas é o anel oxazina, que se mostrou capaz de realizar interações com o resíduo de R59, e está presente em quatro dos quinze compostos apresentados, incluindo o composto 13 que é um dos mais promissores do ponto de vista de suas características estruturais. Vale lembrar também que a maioria dos compostos apresenta baixo número de grupos doadores de ligações de hidrogênio, e que são formados por estruturas mais rígidas (ricas em anéis) com extremidades polares. Outro grupamento presente em várias estruturas é a dioxoisindolina, que dentre os mais promissores está presente nos compostos 12, 13 e 15.

Seguindo este contexto, após a realização das etapas de “screening” virtual baseado no receptor, já estão sendo investigados “screenings” virtuais do ponto de vista comum a esses ligantes, i.e. o padrão farmacofórico, com o objetivo de selecionar mais moléculas promissoras. Assim, a grande perspectiva se encontra na realização futura de ensaios biológicos de atividade das quinze moléculas selecionadas, e das que ainda serão selecionadas nas etapas de “screening” utilizando padrão farmacofórico, com a proteína hnRNP K. Dessa forma, espera-se que resultados pioneiros sejam obtidos em relação a possíveis ligantes específicos da proteína hnRNP K. O processo de aquisição dos compostos selecionados já foi iniciado e, em breve, os ensaios biológicos serão realizados. Em paralelo, vem sendo realizada a clonagem, expressão e purificação da proteína hnRNP K, com o auxílio do Laboratório de Bioquímica Clínica da Faculdade de Ciências Farmacêuticas de Ribeirão Preto, sob supervisão da Profa. Dra. Andréia Machado Leopoldino. Ensaios de afinidade com oligonucleotídeos de fita simples de DNA já foram padronizados, e vêm sendo realizados para as duas isoformas da proteína.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, V. L.; LEITÃO, A.; REINA, L. C. B.; MONTANARI, C. A.; DONNICI, C. L.; LOPES, M. T. P. Câncer e agentes antineoplásicos ciclo-celular específicos e ciclo-celular não específicos que interagem com o DNA: uma introdução. **Química Nova**, v. 28, p. 118-129, 2005.

ALONSO, H.; BLIZNYUK, A. A.; GREASY. Combining docking and molecular dynamic simulations in drug design. **Medical Research Reviews**, v. 26, p. 531-568, 2006.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MEYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, p. 403-410, 1990.

ARCOS, J. C.; ARGUS, M. F. **Chemical induction of cancer**. New York: Academic Press, Volume 2B, 1974.

BABER, J. L.; LIBUTTI, D.; LEVENS, D.; TJANDRA, N. High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a *c-myc* transcription factor. **Journal of Molecular Biology**, v. 289, p. 949-962, 1999.

BACKE, P. H.; MESSIAS, A. C.; RAVELLI, R. B. G.; SATTTLER, M.; CUSACK, S. X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. **Structure**, v. 13, p. 1055-1067, 2005.

BAJORATH, J. Understanding chemoinformatics: a unifying approach. **Drug Discovery Today**, v. 09, p. 13-14, 2004.

BARRIL, X.; GELPI, J. L.; LÓPEZ, J. M.; OROZCO, M.; LUQUE, F. J. How accurate can molecular dynamics/linear response and Poisson-Boltzmann/solvent accesible surface calculations be for predicting relative binding affinities? Acetylcholinesterase huprine inhibitors as a test case. **Theoretical Chemistry Accounts**, v. 106, p. 2-9, 2001.

BARTON, G. J.; STERNBERG, M. J. E. A strategy for the rapid multiple alignment of protein sequences. **Journal of Molecular Biology**, v. 198, p. 327-337, 1987.

BOMSZTYK, K.; DENISENKO, O.; OSTROWSKI, J. HnRNP K: one protein multiple processes. **Bioessays**, v. 26, p. 629-638, 2004.

BOMSZTYK, K.; SEUNINGEN, I. V.; SUZUKI, H.; DENISENKO, O.; OSTROWSKI, J. Diverse molecular interactions of the hnRNP K protein. **FEBS Letters**, v. 403, p. 113-115, 1997.

BRADLEY, P. J.; ZUTSHI, B.; NUTTING, C. M. An audit of clinical resources available for the care of head and neck cancer patients in England. **Clinical Oncology**, v. 17, p. 604-609, 2005.

BRADDOCK, D. T.; BABER, J. L.; LEVENS, D.; CLORE, G. M. Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: solution structure of a complex between hnRNP K KH3 and single-stranded DNA. **The EMBO Journal**, v. 21, p. 3476-3485, 2002.

BRENK, R.; NAERUM, L.; GRAEDLER, U.; GERBER, H.; GARCIA, G. A. Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. **Journal of Medicinal Chemistry**, v. 46, p. 1133-1143, 2003.

BUS, J. S.; POPP, J. A. Perspectives on the mechanism of action of the splenic toxicity of aniline and structurally-related compounds. **Food and Chemical Toxicology**, v. 25, p. 619-626, 1987.

CARLSON, H.; MASUKAWA, K. M.; McCAMMON, J. A. Method for including the dynamic fluctuations of a protein in a computer-aided drug design. **Journal of Physical Chemistry A**, v. 103, p. 10213-10219, 1999.

CARPENTER, B.; MACKAY, C.; ALNABULSI, A, MACKAY, M.; TELFER, C.; MELVIN, W.T.; MURRAY, G.I. The roles of heterogeneous nuclear ribonucleoproteins in tumour development and progression. **Biochimica et Biophysica Acta – Reviews on Cancer**, v. 1765, p. 85-100, 2006.

CHEN, G. S.; CHANG, C. S.; KAN, W. M.; CHANG, C. L.; WANG, K. C.; CHERN, J. W. Novel lead generation through hypothetical pharmacophore three-dimensional database searching: discovery of isoflavonoids as nonsteroidal inhibitors of rat 5 α -reductase. **Journal of Medicinal Chemistry**, v. 44, p. 3759-3763, 2001.

COHEN, M. S.; ZHANG, C.; SHOKAT, K. M.; TAUNTON, J. Structural bioinformatics-based design of selective irreversible kinase inhibitors. **Science**, v. 308, p.1318-1321, 2005.

CRONIN, M. T. D.; BASKETTER, D. A. Multivariate QSAR analysis of a skin sensitization database. **SAR and QSAR in Environmental Research**, v. 02, p. 159-179, 1994.

De BONO, J. S.; ROWINSKY, E. K. The ErbB receptor family: a therapeutic target for cancer. **Trends in Molecular Medicine**, v. 08, n. 4(Suppl.), S. 19-26, 2002.

D'ALFONSO, G.; TRAMONTANO, A.; LAHM, A. Structural conservation in single-domain proteins: implications for homology modeling. **Journal of Structural Biology**, v. 134, p. 246-256, 2001.

DEANE, C. M.; BLUNDELL, T. L. Protein comparative modelling and drug discovery. In : Wermuth, C. G. **The Practice of Medicinal Chemistry**. London : Elsevier Academic Press, 2003, p. 445-458.

DEJGAARD, K.; LEFENS, H. Characterisation of the nucleic-acid-binding activity of KH domains. Different properties for different domains. **European Journal of Biochemistry**, v. 241, p. 425-431, 1996.

DESANY, B.; ZHANG, Z. Bioinformatics and cancer target discovery. **Drug Discovery Today**, v. 09, p. 795-802, 2004.

de SOUZA, S. J.; CAMARGO, A. A.; BRIONES, M. R. S.; COSTA, F. F.; NAGAI, M. A.; ALMEIDA, S. V.; ZAGO, M. A.; ANDRADE, L. E. C.; CARRER, H.; EL-DORRY, H. F. A.; ESPREAFICO, E. M.; HABR-GAMA, A.; GIANELLA-NETO, D.; GOLDMAN, G. H.; GRUBER, A.; HACKEL, C.; KIMURA, E. T.; MACIEL, R. M. B.; MARIE, S. K. N.; MARTINS, E. A. L.; NÓBREGA, M. P.; PAÇÓ-LARSON, M. L.; PARDINI, M. I. M. C.; PEREIRA, G. G.; PESQUERO, J. B.; RODRIGUES, V.; ROGATTO, S. R.; DA SILVA, I. D. C. G.; SOGAYAR, M. C.; SONATI, M. F.; TAJARA, E. H.; VALENTINI, S. R.; ACENCIO, M.; ALBERTO, F. L.; AMARAL, M. E. J.; ANEAS, I.; BENGTSON, M. H.; CARRARO, D. M.; CARVALHO, A. F.; CARVALHO, L. H.; CERUTTI, J. M.; CORRÊA, M. L. C.; COSTA, M. C. R.; CURCIO, C.; GUSHIKEN, T.; HO, P. L.; KIMURA, E.; LEITE, L. C. C.; MAIA, G.; MAJUMDER, P.; MARINS, M.; MATSUKUMA, A.; MELO, A. S. A.; MESTRINER, C. A.; IRACCA, E. C.; MIRANDA, D. C.; NASCIMENTO, A. L. T. O.; NÓBREGA, F. G.; OJOPI, E. P. B.; PANDOLFI, J. R. C.; PESSOA, L. G.; RAHAL, P.; RAINHO, C. A.; RO'S, N.; DE SÁ, R. G.; SALES, M. M.; DA SILVA, M. P.; SILVA, T. C.; JUNIOR, W. S.; SIMÃO, D. F.; SOUSA, J. F.; STECCONI, D.; TSUKUMO, F.; VALENTE, V.; ZALCBERG, H.; BRENTANI, R. R.; REIS, L. F. L.; DIAS-NETO, E.; SIMPSON, A. J. G. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. **Proceedings of the National Academy of Sciences**, v. 97, p. 12690-12693, 2000.

DUNHAM, I.; SHIMIZU, N.; ROE, B. A.; CHISSOE, S.; HUNT, A. R.; COLLINS, J. E.; BRUSKIEWICH, R.; BEARE, D. M.; CLAMP, M.; SMINK, L. J.; AINSCOUGH, R.; ALMEIDA, J. P.; BABBAGE, A.; BAGGULEY, C.; BAILEY, J.; BARLOW, K.; BATES, K. N.; BEASLEY, O.; BIRD, C. P.; BLAKEY, S.; BRIDGEMAN, A. M.; BUCK, D.; BURGESS, J.; BURRILL, W. D.; O'BRIEN, K. P. The DNA sequence of human chromosome 22. **Nature**, v.402, p. 489-495, 1999.

DILLER, D. J.; LI, R. Kinases, homology models, and high throughput docking. **Journal of Medicinal Chemistry**, v. 46, p. 4638-4647, 2003.

Discovery Studio ViewerPRO, Accelrys Inc, San Diego, CA, USA, 2002.

ENYEDY, I. J.; LEE, S. L.; KUO, A. H.; DICKSON, R. B.; LIN, C. Y.; WANG, S. Structure-based approach for the discovery of Bis-benzamidines as novel inhibitors of matriptase. **Journal of Medicinal Chemistry**, v. 44, p. 1349-1355, 2001.

EKINS, S.; ROSE, J. In silico ADME/Tox: the state of the art. **Journal of Molecular Graphics and Modelling**, v. 20, p. 305-309, 2002.

EVERS, A.; KLABUNDE, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha 1A adrenergic receptor. **Journal of Medicinal Chemistry**, v. 48, p. 1088-1097, 2005.

FOLKERS, G. SAR, scope and limitations of molecular design approaches. In: CODDING, P. W. **Structure-based drug design: experimental and computational approaches**. Dordrecht: Kluwer Academic Publishers, 1998. p. 27-40.

GRIDELLI, C.; BARESCINO, M. A.; SCHETTINO, C.; ROSSI, A.; MAIONE, P.; CIARDIELLO, F. Erlotinib in non-small cell lung cancer treatment: current status and future development. **The Oncologist**, v. 12, p. 840-849, 2007.

FRADERA, X.; DE LA CRUZ, X.; SILVA, C. H. T. P.; GELPI, J. L.; LUQUE, F. J.; OROZCO, M. Ligand-induced changes in the binding sites of proteins. **Bioinformatics**, v. 18, p. 939-948, 2002.

GILSON, M.; SHARP, K.; HONIG, B. J. Calculating the electrostatic potential of molecules in solution: method and error assessment. **Journal of Computational Chemistry**, v. 09, n. 04, p. 327-335, 1988.

GOLDENBERG, D.; LEE, J.; KOCH, W. M.; KIM, M. M.; TRINK, B.; SIDRANSKY, D.; MOON, C. Habitual risk factors for head and neck cancer. **Otolaryngology – Head and Neck surgery**, v. 131, p. 986-993, 2004.

GOODFORD, P. J. A Computational procedure for determining energetically favorable binding sites on biologically important macromolecules. **Journal of Medicinal Chemistry**, v. 28, n. 07, p. 849-857, 1985.

GRISHIN, N.V. KH domain: one motif, two folds. **Nucleic Acid Research**, v. 29, p. 638-643, 2001.

HÖLTJE, H. -D.; SIPPL, W.; ROGNAN, D.; FOLKERS, G. Introduction to comparative protein modeling. In: **Molecular Modeling: Basic Principles and Applications**. Weinheim: Wiley-VCH, 2003a, p. 87-143.

HÖLTJE, H. -D.; SIPPL, W.; ROGNAN, D.; FOLKERS, G. Small molecules. In: **Molecular Modeling: Basic Principles and Applications**. Weinheim: Wiley-VCH, 2003b, p. 9-72.

HUTH, J. R.; YU, L.; COLLINS, I.; MACK, J.; MENDOZA, R.; ISAAC, B.; BRADDOCK, D. T.; MUCHMORE, S. W.; COMESS, K. M.; FESIK, S. W.; CLORE, G. M.; LEVENS, D.; HAJDUK, P. J. NMR-driven discovery of benzoylanthranilic acid inhibitors of far upstream element binding protein binding to the human oncogene c-myc promoter. **Journal of Medicinal Chemistry**, v. 47, p. 4851-4857, 2004.

Insight II User Guide, version 2005, Accelrys: CA, USA, 2005.

ITO, K.; SATO, K.; ENDO, H. Cloning and characterisation of a single-stranded DNA binding protein that specifically recognizes deoxycytidine stretch. **Nucleic Acids Research**, v. 22, p. 53-58, 1994.

ITOH, M. Sensitization potency of some phenolic compounds. **Journal of Dermatology**, v. 09, p. 223-233, 1982.

KLEBE, G. Virtual ligand screening: strategies, perspectives and limitations. **Drug Discovery Today**, v. 11, p. 580-594, 2006.

LANIG, H. Molecular dynamics. In: GASTEIGER, J.; ENGEL, T. **Chemoinformatics - A Textbook**. Weinheim: Wiley-VCH, 2003, p. 359-375.

LASKOWSKI, R.A.; MACARTHUR, M.W.; THORNTON, J.M. Procheck: a program to check the stereochemical quality of protein structures. **Journal of Applied Crystallography**, v. 26, p. 283-291, 1993.

LEAVESLEY, H. B.; LI, L.; PRABHAKARAN, K.; BOROWITZ, J. L.; ISOM, G. E. Interaction of cyanide and nitric oxide with cytochrome c oxidase: Implications for acute cyanide toxicity. **Toxicological Sciences**, v. 101, p. 101-111, 2008.

LEOPOLDINO, A. M.; CARREGARO, F.; SILVA, C. H. T. P.; FEITOSA, O.; MANCINI, U. M.; FREITAS, J. M.; TAJARA, E. H. Sequence and transcriptional study of hnRNP K pseudogenes, and expression and molecular modeling analysis of hnRNP K isoforms. **Genome**, v. 50, p. 451-462, 2007.

LIPINSKI, C. A. Lead- and drug-like compounds: the rule-of-five revolution. **Drug Discovery Today: Technologies**, v. 01, p. 337-341, 2004.

LIPINSKI, C. A.; HOPKINS, A. Navigating chemical space for biology and medicine. **Nature**, v. 432, p. 855-861, 2004.

LIPINSKI, C. A.; LOMBARDO, F.; DOMINY, B. W.; FEENEY, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **Advanced Drug Delivery Reviews**, v. 23, p. 3-25, 1997.

LUTHY, R.; BOWIE, J. U.; EISENBERG, D. Assessment of protein models with three-dimensional profiles. **Nature**, v.356, p.83-85, 1992.

MARCU, L.; DOORN, T.; OLVER, I. Cisplatin and radiotherapy in the treatment of locally advanced head and neck cancer: a review of their cooperation. **Acta Oncologica**, v. 42, p. 315-325, 2003.

MARSHALL, G. R. Introduction to chemoinformatics in drug discovery – A personal view. In: OPREA, T. I. **Chemoinformatics in drug discovery**. Weinheim: WILEY-VHC, 2004. p. 1-22.

MCGURK, M.; GOODGER, N. M. Head and neck cancer and its treatment: historical review. **British Journal of Oral and Maxillofacial Surgery**, v. 38, p. 209-220, 2000.

MUNIZ, J. R. C. **Aplicação da bioinformática nos estudos dos genes e enzimas envolvidos na síntese da goma fastidiana produzida pela *Xylela fastidiosa***. 2003. 124f. Dissertação (Mestrado em Ciências: Física Aplicada) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2003.

MUSUNURU, K.; DARNELL, R.B. Determination and augmentation of RNA sequence specificity of the Nova K-homology domains. **Nucleic Acids Research**, v. 32, p. 4852-4861, 2004.

NAYEEM, A.; SITKOFF, D.; JUNIOR, S.K. A comparative study of available software for high accuracy homology modeling: from sequence alignments to structural models. **Protein Science**, v.15, p. 808-824, 2006.

O'BRIEN, S. E.; GROOT, M. J. Greater than the sum of its parts: combining models for useful ADMET prediction. **Journal of Medicinal Chemistry**, v. 48, p. 1287-1291, 2005.

OSTARECK, D.H. ; OSTARECK-LEDERER, A. ; WILM, M. ; THIELE, B.J. ; MANN, M. ; HENTZE, M.W. mRNA silencing in erythroid differentiation: hnRNP K nad hnRNP E1 regulate 15-lipoxygenase translation from the 3' end. **Cell**, v. 89, p. 597-606, 1997.

OSTROWSKI, J. ; BOMSZTYK, K. Nuclear shift of hnRNP K protein in neoplasms and other states of enhanced cell proliferation. **British Journal of Cancer**, v. 89, p. 1493-1501, 2003.

PASTOR, M. ; CRUCIANI, G. ; McLAY, I. ; PICKETT, S. ; CLEMENTI, S. GRID-Independent Descriptors (GRIND) : a novel class of alignment-independent three-dimensional descriptors. **Journal of Medicinal Chemistry**, v. 43, p. 3233-3243, 2000.

PATRICK, G. L. The why and the wherefore: drug targets. In: **An Introduction to Medicinal Chemistry**. New York: Oxford University Press, 2005, p. 8-23.

PAZIEWSKA, A.; WYRWICS, L.S.; BUJNICKI, J.M.; BOMSZTYK, K.; OSTROWSKI, J. Cooperative binding of the hnRNP K three KH domains to mRNA targets. **Federation of European Biochemical Societies Letters**, v. 577, p. 134-140, 2004.

PEITSCH, M. C. Manuel Peitsch discusses knowledge management and informatics in drug discovery. **Drug Discovery Today: BIOSILICO**, v. 02, p. 94-96, 2004.

PENDLINGTON, R. U.; BARRATT, M. D. Molecular basis of photocontact allergy. **International Journal of Cosmetic Science**, v. 12, p. 91-103, 1990.

PINO, I.; PIO, R.; TOLEDO, G.; ZABALEGUI, N.; VINCENT, S.; REY, N.; LOZANO, M.D.; TORRE, W.; GARCIA-FONCILIAS, J.; MONTUENGA, L.M. Altered patterns of expression of members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family in lung cancer. **British Journal of Cancer**, v. 95, p. 921-927, 2006.

RANG, H. P.; DALE, M. M.; RITTER, J. M. Quimioterapia do câncer. In: **Farmacologia**. Rio de Janeiro: Guanabara Koogan, 2001, p. 557-575.

REIS, E. M.; OJOPI, E. P. B.; ALBERTO, F. L.; RAHAL, P.; TSUKUMO, F.; MANCINI, U. M.; GIMARÃES, G. S.; THOMPSON, G. M. A.; CAMACHO, C.; MIRACCA, E.; CARVALHO, A. L.; MACHADO, A. A.; PAQUOLA, A. C. M.; CERUTTI, J. M.; DA SILVA, A. M.; PEREIRA, G. G.; VALENTINI, S. R.; NAGAI,

M. A.; KOWALSKI, L. P.; VERJOVSKI-ALMEIDA, P.; TAJARA, E. H.; DIAS-NETO, E. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck and thyroid cancer. **Cancer Research**, v. 65, p. 1693-1699, 2005.

Revista Época, v. 380, Ago. 2005.

Revista Pesquisa FAPESP, v. 56, Ago. 2000

Revista Pesquisa FAPESP, Edição Especial FAPESP 40 anos, Jun. 2002.

RIDINGS, J. E.; BARRATT, M. D.; CARY, R.; EARNSHAW, C. G.; EGGINGTON, C. E.; ELLIS, M. K.; JUDSON, P. N.; LANGOWSKI, J. J.; MARCHANT, C. A.; PAYNE, M. P.; WATSON, W. P.; YIH, T. D. Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. **Toxicology**, v. 106, p. 267-279, 1996.

RING, C. S.; SUN, E.; McKERROW, J. H.; LEE, G. K.; ROSENTHAL, P. J.; KUNTZ, I. D.; COHEN, F. E. Structure-based inhibitor design by using proteins models for the development of antiparasitic agents. **Proceedings of the National Academy of Sciences**, v. 90, p. 3583-3587, 1993.

RYCROFT, R. J. G.; WILKINSON, J. D. Irritants and sensitisers. In: CHAMPION, R. H.; BURTON, J. L.; EBLING, F. J. G. **Textbook of Dermatology**. Oxford: Blackwell, 1991, p. 717-754.

SALI, A. 100,000 protein structures for the biologist. **Nature Structural & Molecular Biology**, v. 05, p. 1029-1032, 1998.

SALI, A.; BLUNDELL, T. L. Comparative protein modeling by satisfaction of spatial restraints. **Journal of Molecular Biology**, v. 234, p. 779-815, 1993.

SANDERSON, D. M.; EARNSHAW, C. G. Computer prediction of possible toxic action from chemical structure: The DEREK system. **Human & Experimental Toxicology**, v. 10, p. 261-273, 1991.

SCHAFFERHANS, A.; KLEBE, G. Docking ligands into binding site representations derived from proteins. **Journal of Molecular Biology**, v. 307, p. 407-427, 2001.

SCHNEIDER, G.; BÖHM, H. Virtual screening and fast automated docking methods. **Drug Discovery Today**, v. 07, p. 64-70, 2002.

SCHNEIDER, G.; FECHNER, U. Computer-based *de novo* design of drug-like molecules. **Nature Reviews: Drug Discovery**, v. 04, p. 649-663, 2005.

SIDIQI, M.; WILCE, J. A.; VIVIAN, J. P.; PORTER, C. J.; BARKER, A.; LEEDMAN, P.J.; WILCE, M. C. J. Structure and RNA binding of the third KH domain of poly(C)-binding protein 1. **Nucleic Acids Research**, v. 33, p. 1213-1221, 2005.

SILVA, C. H. T. P. **Planejamento racional de inibidores de enzimas-alvo aplicado a diferentes doenças: modelagem, síntese, bioquímica e Qsar**. 1999. 161f. Tese (Doutorado) – Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, 1999.

SILVA, V. B.; SILVA, C. H. T. P. Modelagem molecular de proteínas-alvo por homologia estrutural. **Revista Eletrônica de Farmácia**, v. 04, p. 15-26, 2007.

SIOMI, H.; CHOI, M.; SIOMI, M.C.; NUSSBAUM, R.; DREYFUSS, G. Essential role for KH domains in RNA binding: impaired RNA binding by a mutation in the KH domain of FMR1 that causes fragile X syndrome. **Cell**, v. 77, p. 33-39, 1994.

SNYDER, R. D.; PEARL, G. S.; MANDAKAS, G.; CHOY, W. N.; GOODSID, F.; ROSENBLUM, I. Y. Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of

pharmaceutical molecules. **Environmental and Molecular Mutagenesis**, v. 43, p. 143-158, 2004.

Spartan User's Guide, version 0.6, Wavefunction, Inc: CA, USA, 2006.

Sybyl User Guide, version 7.1, Tripos Inc: CA, USA, 2005.

TAYLOR, R. D.; JEWSBURY, P. J.; ESSEX, J. W. A review of protein-small molecule docking methods. **Journal of Computer-Aided Molecular Design**, v. 16, p. 151-166, 2002.

TOMONAGA, T.; LEVENS, D. Heterogeneous nuclear ribonucleoprotein K is a DNA-binding transactivator. **Journal of Biological Chemistry**, v. 270, p. 4875-4881, 1995.

VANGREVELINGHE, E.; ZIMMERMANN, K.; SCHOEPFER, J.; PORTMANN, R.; FABBRO, D.; FURET, P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. **Journal of Medicinal Chemistry**, v. 46, p. 2656-2662, 2003.

VERDONK, M. L.; COLE, J. C.; HARTSHORN, M. J.; MULRRAY, C. W.; TAYLOR, R. D. Improved protein-ligand docking using GOLD. **Proteins: structure, function and genetics**, v. 52, p. 609-603, 2003.

VITKUP, D.; MELAMUD, E.; MOULT, J.; SANDER, C. Completeness in structural genomics. **Nature Structural & Molecular Biology**, v. 08, p. 559-566, 2001.

VRIEND, G.; SANDER, C. Quality control of protein models: directional atomic contact analysis. **Journal of Applied Crystallography**, v. 26, p. 47-60, 1993.

WADE, R. C. Calculation and application of molecular interaction fields. In : CRUCIANI, G. **Molecular Interaction Fields**. Weinheim: Wiley-VCH, 2006, p. 27-42.