Universidade Estadual de Campinas

Instituto de Química

Departamento de Físico-Química

Tese de Doutorado

# Estudo QSAR de inibidores da secreção gástrica e simulação molecular da inibição.



Aluno:
Edilson Grünheidt Borges

Orientador:
Prof. Dr. Yuji Takahata

Campinas, 21 de Janeiro de 2004

**FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA UNICAMP**

B644e

Borges, Edilson Grünheidt.
Estudo QSAR de inibidores da secreção gástrica e simulação molecular da inibição / Edilson Grünheidt Borges. -- Campinas, SP: [s.n], 2004.

Orientador: Yuji Takahata

Tese (Doutorado) – Universidade Estadual de Campinas, Instituto de Química.

1. Mecânica-quântica. 2. Regressão multivariada. 3. ATP-ase gástrica. 4. Mecanismo de reação. I. Takahata, Yuji. II. Universidade Estadual de Campinas. III. Título.

UNIDADE IQ
Nº CHAMADA
T/UNICAMP
B644e
V ____ EX ____
TOMBO BC/ 59995
PROC. 16.117-04
C ☐ D ☒
PREÇO 11,00
DATA 23-9-04
Nº CPD

# Agradecimentos

Agradecer nominalmente a todos seria enfadonho. Agradeço áqueles que de alguma forma colaboraram para a realização deste tabalho. Em especial gostaria de agradecer ao meu orientador, Yuji Takahata pelo seu apoio. Também quero agradecer meus pais, Wilson e Elza, por muito mais do que pode ser dito e escrito; e ao meu amor, Silvana, por estar ao meu lado.

As instituições que apoiaram a realização do trabalho foram o Instituto de Química da UNICAMP, que concedeu as instalações físicas, a Fundação para o Amparo da Pesquisa do Estado de São Paulo, que concedeu o apoio financeiro através do Processo 98/06065-5 e o Centro Nacional de Processamento de Alto Desempenho em São Paulo, que concedeu recursos computacionais.

# Curriculum Vitae
# Edilson Grünheidt Borges

## Atuação Profissional

1 Universidade Estadual de Campinas - UNICAMP

Vínculo institucional

1999 - 1999 Vínculo: Colaborador , Enquadramento funcional: Estágio não remunerado , Carga horária: 20, Regime : Parcial

Atividades

1/1999 - 6/1999 Estágio , Instituto de Química, Departamento de Química Orgânica

Estágio
1. Auxiliar didático

2 Universidade Estadual do Rio Grande do Sul - UERGS

Vínculo institucional

2002 - 2003 Vínculo: Celetista , Enquadramento funcional: Professor titular , Carga horária: 30, Regime : Dedicação Exclusiva

Atividades

4/2002 - 4/2003 Docência, graduação,

1. QUI001 Química Geral
2. QUI002 Química Orgânica 1

## Áreas de atuação

1 Química Teórica
2 Linguagens de Programação
3 Métodos e Técnicas de Ensino

## Produção bibliográfica

### Artigos completos publicados em periódicos

1 BORGES, E. G., TAKAHATA, Y.
Estudo QSPR sobre os coeficientes de partição: Descritores quânticos e análise multivariada. Química Nova. , v.25, p. 1061 - 1066, 2002.
Palavras-chave : partial least squares aquatic polutants neural net
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Português. Meio de divulgação: Impresso

2 BORGES, E. G., TAKAHATA, Y.
The 4-indolyl-2-guanidinothiazoles QSAR study using quantum mechanical descriptors. Journal Of Molecular Structure Theochem. , v.580, p.253 - 270, 2002.
Áreas do conhecimento : Química Teórica
Referências adicionais : Irlanda do Norte/Inglês. Meio de divulgação: Impresso

3 BORGES, E. G., TAKAHATA, Y.
QSAR study of anti-ulcer compounds using calculated parameters. Journal Of Molecular Structure Theochem. , v.539, p.245 - 251, 2001.
Palavras-chave : Guanidinothiazoles Semi-empirical Partial-least-sq
Áreas do conhecimento : Química Teórica
Referências adicionais : Irlanda do Norte/Inglês. Meio de divulgação: Impresso

Trabalhos publicados em anais de evento

1 BORGES, E. G., OKAMOTO, A. K., TAKAHATA, Y., VAZQUEZ, P. A. M.
Estudo do estado de transição de um composto anti úlcera com método ab-initio usando potencial efetivo no núcleo (ECP) In: 25ª reunião anual da Sociedade Brasileira de Química, 2002, Poços de Caldas.
Livro de resumos. São Paulo: Editora da SBQ, 2002. v.1.
Palavras-chave : Coordenada de reação, potencial efetivo no núcleo,
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Português. Meio de divulgação: Impresso

2 BORGES, E. G.
The 4 indolyl-2-guanidinothiazoles QSAR study of antiulcer activity using quantum chemical descriptors In: 14th European symposium on Quantitative Structure -Activity Relationships, 2002, Bournemouth.
EURO QSAR 2002. , 2002.
Palavras-chave : antiulcer, quantum descriptors, partial least squares
Áreas do conhecimento : Química Teórica
Setores de atividade : Desenvolvimento de produtos tecnológicos voltados para a saúde humana
Referências adicionais : Inglaterra/Inglês. Meio de divulgação: Impresso

3 BORGES, E. G., TAKAHATA, Y.
Estudo QSPR sobre o coeficiente de partição octanol/água In: Reunião Anual da SBQ, 2001, Poços de Caldas.
Livro de Resumos. São Paulo: Sociedade Brasileira de Química, 2001. v.24. p.md-05 -
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Português. Meio de divulgação: Impresso

4 BORGES, E. G., TAKAHATA, Y.
Molecular analysis of Putative SCH28080 binding sites of H+,K(+)-ATPase and H+,Na(+)-ATPase In: 1st Brazilian Symposium on Medicinal Chemistry, 2001, Caxambú.
Caderno de resumos. , 2001. p.SYB11 -
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Inglês. Meio de divulgação: Impresso

5 BORGES, E. G., TAKAHATA, Y.
QSPR studies on partition coefficients: Quantum descriptors and multivariate analysis In: 23ª Reunião anual da Sociedade Brasileira de Química, 2000, Poços de Caldas.
Livro de Resumos da 23ª SBQ. São Paulo: Sociedade Brasileira de Química, 2000. v.2. p.MD080 -
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Inglês. Meio de divulgação: Impresso

6 BORGES, E. G., TAKAHATA, Y.
Estudo QSAR de compostos 4-fenil-2-guanidinotiazóis com atividade biológica In: 22ª Reunião anual da Sociedade Brasileira de Química, 1999, Poços de Caldas.
Caderno de resumos. São Paulo: Sociedade Brasileira de Química, 1999. p.MD028 -
Áreas do conhecimento : Química Teórica
Referências adicionais : Brasil/Português. Meio de divulgação: Impresso

# Resumo

O ESTUDO QSAR DE INIBIDORES DA SECREÇÃO GÁSTRICA E SIMULAÇÃO MOLECULAR DA INIBIÇÃO CONSISTE EM QUATRO FASES. NA FASE UM BUSCA-SE VALIDAÇÃO DOS MÉTODOS DE SIMULAÇÃO DE SOLVATAÇÃO. NA FASE DOIS, A ELABORAÇÃO DE MODELOS DE RELAÇÕES QUANTITATIVAS ENTRE ESTRUTURA QUÍMICA E ATIVIDADE BIOLÓGICA (QSAR) PARA COMPOSTOS MODULADORES DA SECREÇÃO GÁTRICA COM MECANISMO DE AÇÃO LIGADO À ENZIMA $H^+$,$K^+$-ATPASE. A FASE TRÊS É A MODELAGEM PARCIAL DO MECANISMO DE ATIVAÇÃO DE PRÓ-DROGAS. A FASE QUATRO É A UTILIZAÇÃO DOS DADOS DE MECÂNICA QUÂNTICA PARA DESCREVER A TOPOLOGIA DE LIGANTES, CONSTRUÇÃO DE ESTRUTURAS DE FRAGMENTOS DAS PROTEÍNAS E O MAPEAMENTO DE SÍTIOS LIGANTES NAS PROTEÍNAS. A PRIMEIRA FASE USA DESCRITORES MECÂNICO-QUÂNTICOS PARA CRIAR MODELOS DO COEFICIENTE DE PARTIÇÃO OCTANOL/ÁGUA (LOG $P_{O/W}$), COM RESULTADOS DE ESTIMATIVA DE ERRO DE PREVISÃO DE PRÓXIMOS ÀS OBTIDAS COM MÉTODO EXPERIMENTAL DA MEDIÇÃO DE LOG $P_{O/W}$. NA SEGUNDA FASE FORAM OBTIDOS MODELOS QSAR CAPAZES DE ESTIMAR AS ATIVIDADES BIOLÓGICAS $IC_{50}$ MEDIDAS *IN VITRO*. O VALOR DE $q^2 = 0,67$ DO COEFICIENTE DE CORRELAÇÃO EM VALIDAÇÃO CRUZADA PARA UM CONJUNTO DE 124 COMPOSTOS É O RESULTADO MAIS SIGNIFICATIVO. OS DESCRITORES UTILIZADOS PARA MODELAGEM DAS ATIVIDADES DAS DROGAS SÃO CALCULADAS COM MÉTODOS MECÂNICO-QUÂNTICOS. OS MÉTODOS UTILIZADOS PARA OBTER CONJUNTOS DESCRITORES SÃO CÁLCULOS SEMI-EMPÍRICOS E MECÂNICA MOLECULAR PARA BUSCA CONFORMACIONAL E *AB INITIO* PARA A OTIMIZAÇÃO DE GEOMETRIA MOLECULAR E CÁLCULO DAS PROPRIEDADES DESCRITORAS. NA TERCEIRA FASE, DA MODELAGEM DO MECANISMO DE ATIVAÇÃO DE DROGAS QUE SOFREM MODIFICAÇÕES DENTRO DO ORGANISMO PARA TRANSFORMAREM-SE NA ESPÉCIE ATIVA, FORAM MODELADOS ESTADOS DE TRANSIÇÃO E INTERMEDIÁRIOS META-ESTÁVEIS QUE INDICAM UM MECANISMO QUE CONCORDA PARCIALMENTE COM OS RESULTADOS PUBLICADOS E MOSTRA A IMPORTÂNCIA DA ESTABILIZAÇÃO DOS ESTADOS META-ESTÁVEIS NA ATIVIDADE DESTES COMPOSTOS. NA QUARTA FASE, DA MODELAGEM DA INTERAÇÃO DROGA–ENZIMA, OS COMPOSTOS PREVIAMENTE ESTUDADOS EM QSAR SÃO UTILIZADOS PARA MAPEAMENTO DE SÍTIOS DE LIGAÇÃO COM AS PROTEÍNAS DE TRANSPORTE TIPO $H^+$-ATPASE E $H^+$,$K^+$-ATPASE. A DESCRIÇÃO DA TOPOLOGIA DOS LIGANTES LEVA EM CONTA AS INFORMAÇÕES OBTIDAS NOS ESTUDOS QSAR E PERMITE O MAPEAMENTO DE REGIÕES DAS PROTEÍNAS ONDE HÁ MELHOR INTERAÇÃO ELETROSTÁTICA E ESTÉRICA. PARA SORTEIO DE ALINHAMENTOS FOI UTILIZADO O MÉTODO DE *MONTE CARLO SIMULATED ANNEALING*, E PARA CALCULAR A ENERGIA DE ALINHAMENTO FOI UTILIZADA MECÂNICA MOLECULAR COM O CAMPO DE FORÇAS AMBER.

# Abstract

THE QSAR STUDY OF GASTRIC ACID SECRETION INHIBITORS AND MOLECULAR SIMULATION OF THEIR ACTIVITY CONCERN FOUR STEPS. THE FIRST STEP IS TO VALIDATE THE METHODS FOR SOLVENT EFFECTS SIMULATION. THE SECOND STEP IS TO ELLABORATE MODELS OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) FOR COMPOUNDS WHICH ARE GASTRIC SECRETION MODULATORS WITH ACTIVITY MECHANISM BASED ON THE ENZYME $H^+$,$K^+$-ATPASE. THE THIRD STEP IS THE PARTIAL MODELLING OF THE MECHANISM OF PRO-DRUGS ACTIVATION. THE FORTH STEP IS TO APLLY THE QUANTUM-MECHANICS DATA SET TO DESCRIBE MOLECULAR TOPOLOGY, BUILD THE STRUCTURE OF THE PROTEIN FRAGMENTS AND MAP THE LIGAND SITES ON THE PROTEIN. THE FIRST STEP APPLY QUANTUM-MECHANIC DESCRIPTORS FOR MODELLING THE PARTITION COEFFICIENT (LOG $P_{O/W}$), FINDING MODELS WHICH HAS THE CAPABILITY OF ESTIMATE LOG $P_{O/W}$ WITH STANDARD ERROR OF PREDICTION AS SMALL AS THE VALUES OF EXPERIMENTAL METHODS. THE SECOND STEP FOUND QSAR MODELS WITH THE CAPABILITY OF ESTIMATING THE *IN VITRO* BIOLOGICAL ACTIVITY $IC_{50}$. THE VALUE $q^2 = 0.67$ FOR THE CROSS VALIDATED LINEAR CORRELATION TEST FOR A 124 COMPOUNDS DATA SET IS THE MOST SIGNIFICANT RESULT. THE DESCRIPTORS THAT WERE USED TO MODELING THE DRUG ACTIVITIES WERE CALCULATED WITH QUANTUM-MECHANICS METHODS. SEMI-EMPIRCAL OR MOLECULAR MECHANICS CALCULATIONS WERE USED FOR CONFORMATIONAL SEARCH AND *AB INITIO* FOR GEOMETRY OPTIMIZATION AND FOR MOLECULAR PROPERTIES CALCULATIONS. THE THIRD STEP, THE MODELLING OF THE ACTIVATION MECHANISM OF THE DRUGS WHICH NEEDS TO BE TRANSFORMED INTO THE ORGANISM TO THE ACTIVE STATE, TRANSITION STATES AND META-STABLE STATES WHICH POINT TO A MECHANISM WHICH PARTIALLY AGREE WITH THE PUBLISHED ONE AND DEMONSTRATE THE NEEDS OF THE META-STABLE STATE FOR HIGH ACTIVITY. THE FOURTH STEP IS MODELLING THE DRUG-ENZYME INTERACTIONS, USING THE PREVIUSLY STUDIED COMPOUNDS AS LIGANDS TO MAP THE INTERACTION SITES ON THE PROTEINS OF $H^+$-ATPASE AND $H^+$,$K^+$-ATPASE KIND. THE LIGAND TOPOLOGY DESCRIPTION APLLY THE INFORMATION WHICH WERE OBTAINED ON THE QSAR STUDY TO ALLOW THE PROCEDURE. TO GENERATE EACH TRIAL ALIGNMENT THE MONTE CARLO SUMULATED ANNEALING WERE USED TO, AND TO CALCULATE EACH ALIGNMENT ENERGY MOLECULAR MECHANICS, WITH AMBER FORCE FIELD, WERE USED TO.

# Sumário

# Lista de abreviações

AM1: *Austim model* versão um
AQ: aquoso
ASC: carga superficial aparente
BPN: *back-propagation neural network* (rede neural de retropropagação)
CHELPG: carga líquida derivada do potencial eletrostático
CI: interação de configurações
CNDO: *complete negect of diatomic overlap* (negligenciamento completo do recobrimento diatômico)
CV: cavitação
DP: dispersão
D-PCM: *dieletric polarizable continum method* (método do continum polarizável dielétrico)
ECP: potencial efetivo no cerne
EGD: esofagogastroduodenoscopia
ES: eletrostático
$E_{HOMO}$: energia do orbital HOMO
GERD: síndrome do refluxo gastro-esofágico
GTO: *Gaussian type orbital* (orbital Gaussiana)
HF: Hartree–Fock
HOMO: *highest occupied molecular orbital* (orbital molecular ocupado de mais alta energia)
INDO: *intermediate neglect of diatomic overlap* (negligenciamento intermediário da sobreposição diatômica)
LM: *low mode* (modo de menor energia)
LO: *leave one out* (deixe um de fora)
$logP_{O/W}$: coeficiente de partição octanol-água
LUMO: *lowest unoccupied molecular orbital* (orbital molecular desocupado de mais baixa energia)
MD: dinâmica molecular
MINDO: *modified intermediate neglect of diatomic overlap* (negligenciamento intermediário modificado do recobrimento diatômica)
MNDO: *modified neglect of diatomic overlap* (negligenciamento modificado do recobrimento diatômico)
MPE: expansão multipolar
NDDO: *neglect of diatomic differential overlap* (negligenciamento das diferenciais de recobrimento diatômico)
NDO: *neglect of diatomic overlap* (negligenciamento do recobrimento diatômico)
NSAIDs: *non steroidal anti-inflamatory drugs* (drogas anti-inflamatórias não esteroidais)
OM: orbital molecular
PCA: *principal component analysis* (análise por componentes principais)
PC: componente principal
PCM: *polarizable continum method* (método do continum polarizável)
PLS: *partial least squares* (mínimos quadrados parciais)
PRESS: *prediction error sum of squares* (soma dos quadrados dos erros de previsão)
QSAR: *quantitative structure-activity relationship* (relações quantitativas entre estrutura e atividade)
QSPR: *quantitative structure-property relationship* (relações quantitativas entre estrutura e propriedade)
RMSEP: *round mean square error of prediction* (valor médio quadrado dos erros de previsão)
SAR: *structure-activity relationship* (relacções entre estrutura e atividade)
SCF: *self-consistent field* (campo autoconsistente)
SDEP: *standard deviation error of prediction* (desvio padrão de previsão)
SIMCA: *soft independent modelling class analogy* (modelagem por analogia simples entre classes)
STO: *Slater type orbitals* (orbitais de Slater)
SVD: *single value decomposition* (decomposição por valores singulares)
TM: transmembrana
VL: variável latente

ZDO: *zero differential overlap* (diferencial de recobrimento nula)

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução geral e dos métodos

## 1.1 Motivação para o estudo dos inibidores da secreção gástrica

O termo dispepsia refere-se à presença de dor ou desconforto episódico ou persistente, localizado no epigástrio ou andar superior do abdome, definida como desconforto ou dor no andar superior do abdome; queimação retroesternal ou náusea; intermitentes ou contínuos, com pelo menos duas semanas de duração [1,2]. Trata-se de uma das queixas mais comumente referidas por pacientes atendidos em ambulatório, sendo responsável por dois a cinco porcento de todas as consultas ao médico generalista [3].

Dispepsia está associada tanto a um quadro funcional sem maiores repercussões clínicas quanto, na maioria dos casos, a doenças orgânicas importantes que necessitam de pronto reconhecimento através do procedimento diagnóstico de escolha, que é a endoscopia digestiva alta (esofagogastroduodenoscopia: EGD).

Em estudo realizado nos ambulatórios do Departamento de Medicina Clínica e Unidade de Epidemiologia Clínica da Universidade Federal do Ceará com cerca de duzentos indivíduos, mostram que em cerca de cinqüenta porcento dos casos de relato de dispepsia são relacionados a alguma enfermidade [4]. Dispepsia mostrou-se como problema clínico de grande impacto sobre serviços básicos de saúde. Na Tabela 1.1 estão listadas as principais enfermidades detectadas nos indivíduos após exame ambulatorial com EGD.

Considerando dados da Faculdade de Ciências Médicas da Universidade Estadual de Campinas [5], temos que a incidência de dispesia afeta entre trinta e quarenta porcento da comunidade, e que apenas vinte e cinco porcento dos pacientes procuram serviço médico em função da dispepsia. A dispepsia funcional corresponde a cerca de cinqüenta a sessenta porcento dos casos, havendo leve predomínio de pacientes do sexo feminino; A úlcera péptica corresponde a cerca de quinze a vinte porcento dos casos, sendo mais comum em pacientes idosos; A doença do refluxo gastro-esofágico corresponde a cerca de vinte a vinte e cinco porcento dos casos, comum nos obesos. A neoplasia gástrica corresponde a apenas maeio a dois porcento dos casos, concentrando-se na faixa etária mais alta. Fatores associados ao aumento da incidência de todos os tipos de dispesia são tabagismo, obesidade, ingestão abusiva de sal e conservas, uso de medicamentos anti-

Tabela 1.1: Principais enfermidades associadas ao relato de dispepsia em estudo de duzentos casos realizado na cidade de Fortaleza durante dez meses.

| Características | Gastrite Crônica | Úlcera Duodenal | Úlcera Gástrica | Esofagite | Total entre categorias[a] |
|---|---|---|---|---|---|
| Idade Média (faixa) | 42 (16-69) | 48 (27-76) | 54 (27-73) | 41 (20-67) | 46 (16-76) |
| Sexo masculino (%) | 28,1 | 70,6 | 44,4 | 53,3 | 20 |
| Baixo nível de escolaridade[b] (%) | 26,3 | 52,9 | 33 | 26,7 | 15 |
| Aposentadoria (%) | 10,5 | 17,6 | 22 | 6,7 | 6 |
| SRQ maior de 7 (%) | 71,9 | 64,7 | 44,4 | 60 | 32,5 |
| Visitou médico nos últimos 12 meses (%) | 84,2 | 88,2 | 100 | 80 | 42 |
| Deixou de trabalhar devido à dispepsia (%) | 43,8 | 29,4 | 44,4 | 53 | 21 |
| Indivíduos | n=57 | n=17 | n=9 | n=15 | n=98 |
| Porcentual | 28,5% | 8,5% | 4,5% | 7,5% | 49% |

[a]As categorias diagnósticas neoplasia gástrica (n=2), duodenite (n=4), gastrite erosiva aguda (n=1), e outros: pólipo, teleangiectasia e alteração anatômica de estômago (n=4), não foram incluídas devido ao número reduzido de pacientes que apresentaram as mesmas.

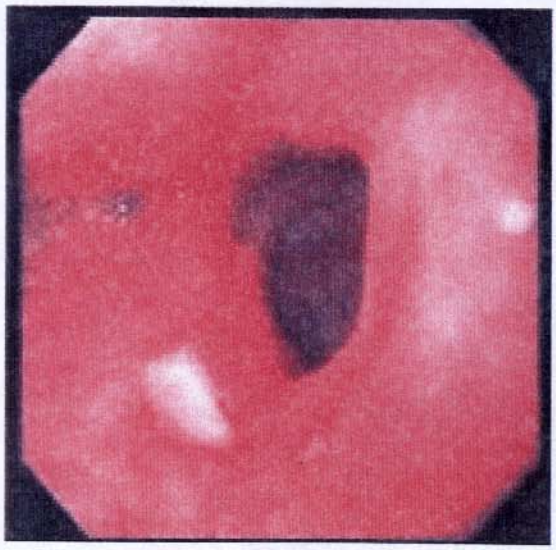
[b]Indivíduos que cursaram no máximo o primeiro grau.

inflamatórios, infecção por *Helicobacter pylori*, diabetes mellitus (longa evolução) e vários distúrbios de motilidade gastrointestinal.

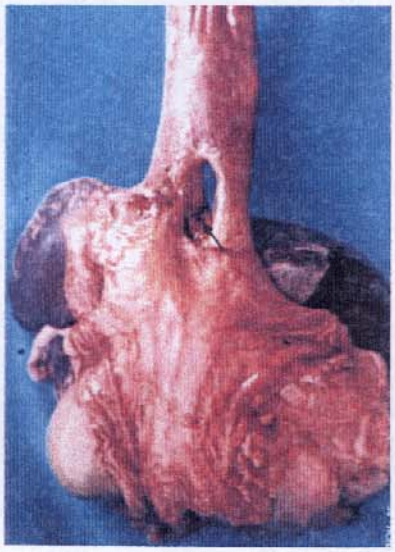
Dados em escala global mostram números que afastam a idéia de que trata-se de uma doença de menor importância. Segundo dados do *Centers for Disease Control and Prevention* dos Estados Unidos da América, mais de vinte e cinco milhões de cidadões norte-americanos padecem de úlcera durante sua vida [6]. Uma pesquisa realizada na Dinamarca entre 1981 até 1993 mostra que os avanços no tratamento da úlcera péptica não estão relacionados com uma diminuição da hospitalizção e diminuição sensível da mortalidade decorrente de complicações da doença [7]. Dados da Inglaterra obtidos através do *Office of National Statistics* mostram que apesar da diminuição total da entrada de pacientes no sistema hospitalar devido a complicações da doença no período entre 1950 e 1980, houve um aumento nos casos da mortalidade em mulheres de idade avançada, provavelmente associado ao aumento no consumo de anti-inflamatórios. Mostram também que o aumento das prescrições de inibidores da bomba de prótons gástrica foi de cerca de cinco mil porcento somente no período entre 1989 até 1999, ainda assim acompanhada da crescente incidência de complicações gástricas em idosos [8].

Na Figura 1.1 **A** podemos observar uma imagem obtida por endoscopia de uma pequena ferida ulcerativa na região do piloro, o músculo que separa o estômago do intestino. Na Figura 1.1 **B** temos uma imagem muito impressionante do sistema digestivo retirado em autópsia, mostrando um caso de úlcera perfurada que levou o paciente a óbito. Fica claro que a doença causa dor muito intensa em estágios avançados, como o mostrado. Entretanto, pode acontecer da úlcera não causar sintomas e a primeira manifestação ser

uma complicação da doença como o sangramento e a perfuração. Quando ocorre o sangramento o paciente nota fezes pretas, brilhantes, moles e particularmente mal cheirosas (melena) e/ou vômito com sangue vivo ou tipo borra de café (hematêmese) [9].



A



B

Figura 1.1: **A**: Imagem obtida por endoscopia da região do piloro do estômago mostrando uma ferida ulcerativa. O ponto preto na região superior esquerda da figura é o centro da ferida, o a cor escura é devida ao contato do sangue com o suco gástrico. Esta cor preta presente nas fezes é um dos sintomas de sangramento da úlcera. **B**: Foto mostrando o estômago e o duodeno (o fígado, escuro, ao fundo) de um paciente falecido em decorrência da perfuração na região do duodeno, mostrada claramente pela visualização da cor do fundo da foto. Na região da perfuração podemos observar a coloração escura devida ao sangue coagulado em contato com o suco gástrico, apontada por uma pequena flecha.

As principais drogas utilizadas no combate da úlcera são antibióticos, para erradicar a infecção por *H. Pylori*, os antiácidos e os reguladores da secreção ácida. Temos drogas derivadas de Omeprazol, que inibem a bomba des próton do estômago, e os inibidores do receptor H2 de histamina, como as mais receitadas para regulação da secreção gástrica. Porém, a utilização destas drogas por longos períodos tem problemas sérios. Os inibidores da bomba de prótons causam hiperplasia das células enterocromafínicas da parede do estômago e aumentam a chance do desenvolvimento de glaucoma. Também a utilização de inibidores do receptor H2 de histamina pode não ser suficiente para controlar a secreção gástrica, havendo desenvolvimento de tolerância a estas drogas e um perigoso efeito de rebote na abstinência [10].

Por outro lado, a venda de produtos farmacêuticos é um negócio gigantesco que movimento cerca de quatrocentos e trinta bilhões de dólares ao ano no mundo todo, mostrando um crescimento de cerca de oito porcento no biênio 2001-2002. Os produtos relacionados ao combate da úlcera respondem por cerca de seis porcento do faturamento total da indústria farmacêutica, algo em torno de vinte e dois bilhões de dólares. Esta classe de drogas é a líder do mercado à mais de uma década. Drogas como o Prilosec e Losec,

da marca AstraZeneca só têm tido uma diminuição acentuada da sua venda porque são pressionadas pela ascenção dos equivalentes genéricos e do equivalente de nova geração da própria AstraZeneca, denominado Nexium [11].

Considerando os pontos citados, fica bastante clara a importância de estudar compostos reguladores da secreção ácida. Cientificamente é um campo bastante amplo. O mecanismo de ação das drogas que regulam os canais de membrana envolvidos no processo de secreção ácida é ainda pouco conhecido [12, 13], havendo grandes possibilidades para o desenvolvimento racional de drogas a partir da elucidação iminente das estruturas terciárias das enzimas alvo. Do ponto de vista tecnológico e comercial, temos que novas drogas estão sendo lançadas no mercado ao mesmo tempo em que as patentes dos medicamentos tradicionais estão espirando, criando brechas para o desenvolvimento de produtos similares e originais, de grande potencial econômico.

## 1.2 Metodologias para modelagem empírica SAR e QSAR baseada na droga

O objetivo do desenvolvimento de drogas é descobrir compostos químicos e otimizá-los, para que interajam efetivamente com um organismo vivo, e como conseqüência compensem ou revertam uma doença. O desafio é grande. Como selecionar num universo químico enorme, contendo algo entre $10^6$ moléculas (bases de dados) ou mais de $10^{100}$ pequenas moléculas orgânicas possíveis de serem sintetizadas, os possíveis candidatos a ter alguma atividade biológica desejável?

A abordagem corrente é obter e testar um grande número de compostos o mais rápido possível e, tão rápido quanto possível, focalizar-se num dos raros compostos capazes de desencadear processos bioquímicos interessantes, uma abordagem conhecida como *high-throughput screening*. A eficiência típica do processo desta natureza pode chegar até um porcento [14].

A Química Teórica preocupa-se, entre outras coisas, em predizer exatamente propriedades físico-químicas baseando-se em suas estruturas moleculares. Através de cálculos é possível obter qualquer valor observável, pela aplicação do operador adequado sobre uma função de onda. Este conjunto enorme de propriedades que podem ser calculadas pode ser utilizado no reconhecimento de compostos com atividade bioquímica desejável.

Criando modelos estatísticos de classificação e reconhecimento de padrões com propriedades moleculares de origem mecânico-quântica, pode-se estabelecer procedimentos de classificação para caracterização da semelhança química e da reatividade em um determinado tipo de reação. Finalmente, a descrição química da interação entre a droga e um alvo específico no organismo pode ser modelada [15], permitindo um grande avanço no entendimento do processo e dos mecanismos que controlam sua efetividade para o objetivo desejado.

O início do trabalho de modelagem é calcular estruturas moleculares. Se não sabemos qual conformação de uma dada molécula é a bioativa, devemos arbitrar uma estrutura, baseando-se em análise conformacional. Sabemos que se uma droga não se liga ao seu receptor em uma conformação de baixa energia, não será um bom ligante [16]. Podemos

utilizar diversos métodos para gerar conformações, e otimizar as geometrias [17]. A geração de conformações pode ser realizada com método sistemático, método estocástico com matriz de distâncias métricas ou método de Monte Carlo, entre outros. As estruturas obtidas assim freqüentemente necessitam ter sua geometria otimizada para minimização da energia total da molécula. A minimização quase sempre é realizada com métodos de mecânica molecular ou mecânico-quânticos, em um nível de cálculo considerado satisfatório na relação entre custo e exatidão. A determinação do nível de cálculo mais adequado leva em conta a presença de elementos pesados, a extensão da série de compostos estudada e a disponibilidade de recursos computacionais.

Uma vez definida a estrutura, são calculadas as propriedades moleculares consideradas adequadas à descrição do sistema químico. Para cada composto são calculadas propriedades como cargas atômicas líquidas, momentos de dipolo, densidades eletrônicas de alguns orbitais e energias de orbitais. Estas propriedades são freqüentemente utilizadas como descritores eletrônicos, mas o conjunto pode estender-se a diversas outras propriedades. Podem também ser obtidas grandezas derivadas destas propriedades eletrônicas. O volume molecular pode ser calculado através da criação de uma cavidade de solvente com método do *Continuum* polarizável, e as hidrofilicidades podem ser obtidas por cálculos dos efeitos de solvatação em diferentes fases. Portanto, um conjunto de propriedades mecânico-quânticas pode abranger as interações eletrônicas, estéricas e de hidrofilicidade tradicionalmente utilizadas em QSAR, segundo a abordagem de Hansch.

É importante perceber o quanto a escolha dos descritores usados representa uma parte crucial. O quanto um descritor pode codificar da estrutura das moléculas, depende do tipo do descritor, da maneira como ele é obtido e da confiabilidade dos dados. Um quadro bastante abrangente dos descritores que podem ser obtidos com cálculos quânticos é mostrado por Karelson e co. [18].

Os descritores podem ser também parâmetros físico-químicos obtidos experimentalmente (hidrofilicidade, pK), indicadores de fragmentos (ocorrência de uma sub-estrutura), características topológicas (índices de conectividade ou distâncias), relações geométricas (forma) ou qualquer outra forma razoável de decodificação [19].

Descritores derivados de química quântica são fundamentalmente diferentes de valores obtidos experimentalmente, apesar da sobreposição natural. Diferentemente de medidas experimentais, não existe erro de associado à precisão das medidas em valores obtidos em cálculos quânticos. Há, sim, um erro associado às aproximações usadas para facilitar (ou permitir) os cálculos. Um fator que minimiza o efeito do desconhecimento da magnitude do erro associado ao método de cálculo, é que em séries análogas submetidas ao mesmo nível de cálculo, estes erros seguem a mesma tendência. As metodologias estatísticas utilizadas na elaboração dos modelos podem tornar este tipo de erro sistemático de menor importância.

Fatores importantes na escolha de descritores adequados, para utilização em SAR e QSAR são a reversibilidade do código, a invariância rotacional (e translacional) e a obtenção de um conjunto descritor unívoco e de tamanho uniforme.

A reversibilidade do código significa a possibilidade de construir uma estrutura molecular completa, baseando-se somente na análise do conjunto descritor. A unicidade relaciona-se à reversibilidade, e significa que um dado estrutural será codificado por ape-

nas um conjunto descritor, e que todo conjunto descritor codificará apenas uma estrutura.

O tamanho uniforme significa que, independente do tamanho da molécula, o tamanho do conjunto descritor terá o mesmo número de valores. A invariância rotacional (ou translacional) significa que o conjunto descritor independe do alinhamento entre moléculas, ou mesmo da geometria molecular. [20–23]. É fácil perceber que a maioria dos descritores freqüentemente utilizados não atende a todos estes requisitos simultaneamente.

## 1.3 Métodos mecânico-quânticos utilizados em SAR e QSAR

Uma parte da pesquisa em química preocupa-se em predizer propriedades, ou classificar acuradamente as estruturas das moléculas baseando-se em suas propriedades, com base somente na sua estrutura. A primeira racionalização para estas previsões é que as propriedades de um composto químico são função de sua estrutura molecular.

O objetivo principal é obter uma ponte para o entendimento das relações entre a propriedade de interesse e um conjunto de estruturas químicas. Pré-requisitos indispensáveis para aplicação de um modelo de estrutura–propriedade são a definição da propriedade e o reconhecimento do que a propriedade representa em termos estruturais.

Métodos de química quântica e técnicas de modelagem molecular permitiram a definição de um grande número de quantidades locais e moleculares que caracterizam a reatividade, forma e afinidade de ligação de uma molécula completa, assim como de fragmentos e substituintes.

- **Cargas atômicas líquidas.** Cargas elétricas nas molécula originam forças quando na presençao de um campo elétrico que direcionam, atraem ou repelem as partes carregadas. Portanto descritores baseados em cargas têm sido amplamente usados como índices de reatividade química ou como medida de interações intermoleculares fracas. Infelizmente as cargas não são quantidades que podem ser obtidas diretamente pela aplicação de um operador à uma função de onda. Os métodos utilizados para calcular cargas são variados. Este descritor tem sido muito utilizado em SAR e QSAR e corresponde ao fator determinante em alguns casos de ligação droga–receptor [24–26]

- **Energias de orbitais moleculares.** As energias de HOMO e LUMO são descritores quânticos muito populares. É demonstrado que estes orbitais têm papel fundamental na reatividade e estereosseletividade de reações químicas. O potencial de ionização pode ser expresso como $-\varepsilon_{HOMO}$, e o seu valor indica se um composto deve agir como nucleófilo em reações químicas, e mesmo sua reatividade em reações bioquímicas [27].

- **Densidades nos orbitais de fronteira.** De acordo com a teoria de reatividade do elétron de fronteira, muitas reações iniciam-se no sítio onde a sobreposição entre HOMO e LUMO pode ser máxima. No caso de uma molécula doadora de elétrons a densidade em HOMO é crítica para melhor transferência de carga. Num aceptor de

elétrons a densidade no LUMO é importante. A energia relativa entre os orbitais também é um fator importante. Os valores devem ser escalados pelo valor do potencial de ionização para comparação entre diferentes moléculas [28–31].

- **Polarizabilidade molecular.** É o efeito de polarização da molécula por um campo elétrico externo dado em termos dos tensores de suscetibilidade de ordem $n$ da molécula. Seu valor é relacionado ao volume molecular, e também à hidrofilicidade de moléculas. O tensor de primeira ordem está relacionado a possíveis interações do tipo indutivo na molécula, enquanto o termo de segunda ordem caracteriza propriedades aceptoras de elétrons da molécula. Este descritor tem sido muito utilizado em estudos QSAR [32–35].

- **Momento de dipolo e índices de polaridade.** Há relativo consenso na sua importância em fenômenos físico-químicos. O momento de dipolo total, entretanto, reflete apenas a polaridade global da molécula. Calcular valores dos momentos de dipolo para regiões específicas, resulta grandezas que são fortemente dependentes da orientação. Sua utilização em modelos de QSAR mostra resultados positivos em casos específicos [36].

- **Energia de protonação**. É definida como a diferença do calor de formação calculado para a espécie protonada e não protonada [37] pode ser usada para caracterizar a formação de pontes de hidrogênio na região da protonação, que tem grande importância na ligação entre drogas e receptores [38].

Estes descritores têm como vantagem o fato de serem possíveis de se obter para a maioria dos casos. Têm como desvantagem problemas relacionados a dependência de alinhamento, e não reversibilidade dos dados obtidos em informação estrutural. O problema da reversibilidade pode ser tratado realizando todo o processo de aquisição de dados para cada nova estrutura proposta. O problema de alinhamento só pode ser tratado dentro de séries restritas, ou considerando apenas o alinhamento de fragmentos bioisósteres no caso de atividades biológicas.

Outro problema relacionado a estes descritores quânticos é que o tratamento dos conjuntos de dados assim obtidos freqüentemente envolve técnicas matemáticas sofisticadas. Uma molécula orgânica simples com trinta e cinco ou quarenta átomos produz conjuntos de dados com centenas ou milhares de variáveis. A maneira de simplificar o problema é reconhecer conjuntos de variáveis que têm grande potencial baseando-se a análise no conhecimento químico das reações e das interações envolvidas no sistema estudado, ou de sistemas tão semelhantes quanto possível. As técnicas estatísticas de análise multivariada de dados podem ser de grande valia como instrumento auxiliar de reconhecimento dos descritores adequados.

## 1.4 A Mecânica Molecular (MM) e os campos de força

Os métodos baseados na Mecânica Molecular usam as leis da física clássica para predizer propriedades estruturais e moleculares. Existem vários métodos disponíveis em

diversos programas como GROMACS [39], MacroModel [40], TINKER [41], AutoDock [42], entre outros. Cada método é caracterizado por um campo de força específico, que leva em conta parâmetros que o torna mais adequado para certas aplicações. Um campo de força é escrito como um somatório de termos que descrevem a energia necessária para distorção da molécula num campo específico. De maneira geral um campo de força pode ser descrito pela Equação 1.1, onde $E_{str}$ representa a energia para o estiramento de uma ligação entre dois átomos. $E_{bend}$ representa a energia para torsão de um ângulo. $E_{tors}$ é a energia de torsão para uma rotação ao redor da ligação. $E_{vdw}$ e $E_{el}$ descrevem as interações átomo–átomo não ligados. $E_{cross}$ descreve o acoplamento entre os três primeiros termos, e o termo $E_{constr}$ refere-se as restrições que devem ser aplicadas ao sistema.

$$E_{FF} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{el} + E_{cross} + E_{constr} \tag{1.1}$$

Algumas vantagens e desvantagens dos métodos de MM podem ser destacadas, de maneira bem geral:

- Os cálculos de MM são computacionalmente baratos;
- Cada campo de força apresenta bons resultados para uma classe limitada de moléculas;
- Por desprezarem as interações eletrônicas, os métodos não podem tratar problemas químicos onde efeitos eletrônicos sejam predominantes;
- Dependem da disponibilidade de parâmetros para cada tipo de átomo.

Um campo de força típico leva em consideração a Equação 1.1 na elaboração dos seus parâmetros, agrupando termos que contém as interações entre átomos ligados, os termos que contém interações entre átomos não ligados e os termos relacionados às restrições impostas ao sistema, com uma variedade relativamente grande de funções para descrever cada grupo. Algumas funções que representam estas interações são muito comuns, ocorrendo em grande parte dos campos de força existentes e são apresentadas a seguir.

### 1.4.1 As interações entre átomos não ligados

Os campos de força desenvolvidos para utilização em cálculos de mecânica molecular aplicam diversos tipos de funções para representar as interações entre átomos não ligados covalentemente. Cada campo de força, particularmente, pode implementar os tipos de interação aqui citados de maneira distinta, tanto por motivos de aumento de eficiência computacional ou para redução do erro nos cálculos. Os detalhes da implementação da cada tipo de interação são dependentes do campo aplicado, permanecendo o sentido da sua utilização no campo de força.

#### Potencial de Lennard–Jones

O potencial de Lennard–Jones entre um par de átomos $i$ e $j$ é definido pela Equação 1.2, onde $C_{ij}^{(12)}$ e $C_{ij}^{(6)}$ são parâmetros definidos para cada tipos de par de átomos e

$r_{ij}$ é a distância entre o par.

$$V_{LJ}(r_{ij}) = \frac{C_{ij}^{(6)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \tag{1.2}$$

### Potencial de Buckingham

O potencial de Buckingham tem um termo de repulsão mais realista que o utilizado em Lennard–Jones, com custo computacional maior. A função que define o potencial é mostrada na Equação 1.3, onde $A_{ij}$, $B_{ij}$ e $C_{ij}$ são parâmetros e $r_{ij}$ é o raio entre o par de átomos $i$ e $j$.

$$V_{Buck}(r_{ij}) = A_{ij} \exp(-B_{ij} r_{ij}) - \frac{C_{ij}}{r_{ij}^{6}} \tag{1.3}$$

### Potencial de Coulomb

A energia de associada à interação entre duas partículas $i$ e $j$ carregadas é dada pela Equação 1.4, onde $f = \frac{1}{4\pi\varepsilon_0}$, $q_i$ e $q_j$ são as cargas das partículas, $\varepsilon_f$ é a constante dielétrica do meio e $r_{ij}$ é a distância entre as partículas.

$$V_{coul}(r_{ij}) = f \frac{q_i q_j}{\varepsilon_f r_{ij}} \tag{1.4}$$

O potencial de Coulomb freqüentemente é modificado nos diversos campos de força. São aplicados correções na forma de fatores e de parcelas, principalmente para melhor representar a constante dielétrica em meios contínuos. A introdução de limites de corte (*cutoff*) permite que este potencial tenha representação em forma adequada para distâncias curtas e longas, evita problemas na derivação de funções truncadas e aumenta a eficiência computacional na implementação dos diversos campos de força.

## 1.4.2 As interações entre átomos ligados

Os átomos ligados covalentemente estão sujeitos às interações que vão depender da influência de mais de dois átomos, em interações que representam a energia vibracional e nas interações que representam a energia conformacional. Também no caso das interações entre átomos ligados há grande variedade na aplicação de funções que melhor representam os sistemas em estudo, pelos diversos tipos de campos de força desenvolvidos.

### Potencial harmônico

A variação de energia causada pelo estiramento de uma ligação pode ser representada pela função de potencial harmônico, mostrada na Equação 1.5, onde $b_{ij}^0$ é a distância de equilíbrio entre os átomos $i$ e $j$, $k_{ij}^h$ é a constante de força do oscilador harmônico e $r_{ij}$ é a distância modificada entre o par de átomos.

$$V_{har}(r_{ij}) = \frac{1}{2} k_{ij}^h (r_{ij} - b_{ij}^0)^2 \tag{1.5}$$

### Potencial de quarta ordem

O potencial de estiramento pode ser modificado para utilizar a função definida pela Equação 1.6, onde $2k_{ij}^{4th}b^2{}_{ij} = k_{ij}^h$. Esta função permite maior eficiência computacional por que não há necessidade de se obter a raiz quadrada.

$$V_{4th}(r_{ij}) = \frac{1}{4}k_{ij}^{4th}(r_{ij}^2 - b_{ij}^2)^2 \tag{1.6}$$

### Potencial de Morse

A energia associada ao estiramento de uma ligação pode também ser representada pelo potencial de Morse. Este potencial é utilizado em substituição ao potencial de estiramento obtido de um oscilador harmônico por ter energia igual à zero se as distâncias interatômicas são infinitas, e por ter um perfil não simétrico em relação ao ponto de mínimo da função. Este potencial pode ser obtido pela Equação 1.7, onde $D_{ij}$ define a profundidade do poço de potencial, $\beta_{ij}$ define a largura do poço e $b_{ij}$ é a distância interatômica de equilíbrio entre o par de átomos $i$ e $j$.

$$V_{morse}(r_{ij}) = D_{ij}[1 - \exp(-\beta_{ij}(r_{ij} - b_{ij}))]^2 \tag{1.7}$$

### Potencial cúbico

Esta é outra representação de um potencial não harmônico, mais simples que o potencial de Morse, obtido basicamente com a adição de um termo cúbico ao potencial harmônico. A função que representa este potencial é mostrada na Equação 1.8, onde $k_{ij}^h$ é a constante de força utilizada no potencial harmônico, $k_{ij}^{cub}$ é uma nova constante de força introduzida e $b_{ij}^0$ é a distância de equilíbrio entre o par de átomos $i$ e $j$.

$$V_{cub}(r_{ij}) = k_{ij}^h(r_{ij} - b_{ij}^0)^2 + k_{ij}^h k_{ij}^{cub}(r_{ij} - b_{ij}^0)^3 \tag{1.8}$$

### Potencial vibracional angular cossenóide

O potencial associado á vibração angular de três átomos $i$, $j$ e $k$ consecutivamente ligados, com o átomo $j$ no centro, pode ser aproximando por uma função cosseno do ângulo $\theta_{ijk}$, conforme mostrado na Equação 1.9, onde $k_{ijk}^\theta$ é uma constante de força e $\theta_{ijk}^0$ é o ângulo de equilíbrio entre os átomos.

$$V_{vcos}(\theta_{ijk}) = \frac{1}{2}k_{ijk}^\theta(\cos(\theta_{ijk}) - \cos(\theta_{ijk}^0))^2 \tag{1.9}$$

### Potencial vibracional angular harmônico

Outra representação do potencial de vibração angular utiliza-se de função harmônica, mostrada na Equação 1.10, onde os termos têm a mesma notação que na Equação 1.9.

$$V_{vhar}(\theta_{ijk}) = \frac{1}{2}k_{ijk}^\theta(\theta_{ijk} - \theta_{ijk}^0)^2 \tag{1.10}$$

### Potencial de Ryckaert–Bellemans para diedros

O potencial associado à rotação de diedro em uma ligação entre os átomos $j$ e $k$ no diedro $ijkl$, pode ser aproximado pela Equação 1.11, onde os valores de $C_n$ são tabelados, $\theta = 0°$ na conformação *cis* e $\Psi = \theta - 180°$.

$$V_{rb}(\theta_{ijkl}) = \sum_{n=0}^{5} C_n(\cos(\Psi))^n \tag{1.11}$$

A expressão da Equação 1.11 pode ser utilizada para diferentes tipos de átomos, modificando os valores tabelados de $C_n$.

### Potencial periódico para diedros

O potencial periódico pode ser aplicado na representação da energia rotacional segundo a Equação 1.12, onde $k_\theta$ é uma constante de força, $\theta_0$ é utilizado para ajustar a posição dos mínimos de energia e $n$ o número de mínimos da função.

$$V_{per}(\theta_{ijkl}) = k_\theta(1 + \cos(n\theta - \theta_0)) \tag{1.12}$$

### Potencial de diedros impróprios

Um potencial deve ser aplicado para evitar que grupos planares sofram deformação, e para evitar que moléculas invertam sua configuração, para a imagem especular. Este potencial pode ser do tipo harmônico, dado por uma expressão simples como a Equação 1.13, onde $\xi^0$ é o valor que deve ser mantido sem deformação, $\xi_{ijkl}$ o ângulo atual e $k_\xi$ é uma constante de força. O potencial na forma da Equação 1.13 é harmônico e não periódico, então o valor de $\xi^0$ deve ser definido diferente de $\pm 180°$.

$$V_{id}(\xi_{ijkl}) = k_\xi(\xi_{ijkl} - \xi^0)^2 \tag{1.13}$$

## 1.4.3 Interações especiais

Pontes de hidrogênio são formadas, e sua descrição involve atração eletrostática, transferência de carga e formação de ligação. Interações especiais são adicionadas para reproduzir este comportamento complexo. Também é necessário introduzir outros tipos de interação sem significado físico algum, para impor restrições ao movimento de partículas. Estas interações são introduzidas para incluir o conhecimento de dados experimentais e evitar distorções do sistema durante etapas de equilibração dos sistemas (comuns na utilização explícita de solventes). Estas interações não fazem parte do campo de força, e sim da implementação de cada programa, havendo alguns tipos de interação mais comuns, mostradas aqui como exemplo.

### Pontes de hidrogênio

A origem da ponte de hidrogênio é alvo de debate. Sabe-se que a pequenas distâncias há um balanço entre atração eletrostática, transferência de cargas, polarização, dispersão e repulsão eletrônica. A distâncias maiores a força é quase exclusivamente eletrostática.

Em mecânica molecular freqüentemente adicionam o termo de correção para representar pontes de hidrogênio como termos no potencial eletrostático e no potencial de van der Waals. O termo adicionado ao potencial eletrostático pode ser exponencial ou uma interação tipo Lennard–Jones modificada. Um potencial, implementado no método GRID [43], leva em conta a distância entre o átomo doador de elétrons da ponte e o átomo de hidrogênio e a orientação. A função dependente da distância implementada, $E_{phr}$, é mostrada na Equação 1.14, onde '$C$' e '$D$' são parâmetros e '$r$' é a distância entre os átomos.

$$E_{phr} = \frac{C}{r^8} - \frac{D}{r^6} \tag{1.14}$$

Para o termo de orientação há dependência do tipo de átomo doador de elétrons, e do átomo em que o hidrogênio está ligado, não havendo uma formulação geral para a contribuição.

O campo de forças MM3 tem a expressão mostrada na Equação 1.15 para a energia associada à formação de uma ponte de hidrogênio [44], onde '$K_{HB}$' é um parâmetro de ligação de hidrogênio, '$r$' é a distância de equilíbrio da ponte de hidrogênio, '$R^0_{XH}$' é a distância de equilíbrio da ligação $X$—$H$ e '$\varepsilon$' é a constante dielétrica do meio. Os valores de '$R_{YH}$', '$R_{XH}$' e '$\beta$' são definidos de acordo com a Figura 1.2.

$$E_{HB} = K_{HB}\left(184000 * \exp\left(-12 * \left(\frac{R_{YH}}{r}\right)\right)\right) - \left(\frac{2,25\left(\frac{r}{R_{YH}}\right)^6 \cos\beta\left(\frac{R_{XH}}{R^0_{XH}}\right)}{\varepsilon}\right) \tag{1.15}$$

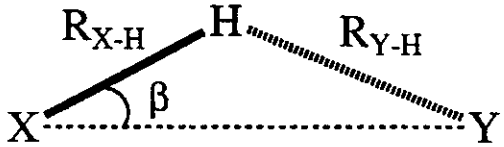



Figura 1.2: Esquema utilizado para definir os valores das variáveis da Equação 1.15 para representação das pontes de hidrogênio com o método MM3.

### Restrições de posição

Um potencial pode ser introduzido para manter duas partículas quaisquer a uma dada distância de referência $R_i$ através de uma função como a mostrada na Equação 1.16, onde

$k_{rpos}$ é uma constante de força geralmente definida no momento da aplicação da restrição e $r_i$ é a distância atual entre as partículas.

$$V_{rpos}(r_i) = \frac{1}{2} k_{rpos} \parallel r_i - R_i \parallel^2 \tag{1.16}$$

Outras implementações podem incluir também funções definidas por partes para representar as restrições de distância entre partículas, tornando a representação um pouco mais complicada. Na Equação 1.17 é mostrado um exemplo para definição de potencial por partes.

$$V_{rpp}(r_{ij}) = \begin{cases} \frac{1}{2} k_{rpp}(r_{ij} - r_0)^2 & se \quad r_{ij} < r_0 \\ 0 & se \quad r_0 \leq r_{ij} < r_1 \\ \frac{1}{2} k_{rpp}(r_{ij} - r_1)^2 & se \quad r_1 \leq r_{ij} < r_2 \\ \frac{1}{2} k_{rpp}(r_2 - r_1)(2r_{ij} - r_2 - r_1) & se \quad r_2 \leq r_{ij} \end{cases} \tag{1.17}$$

### Restrições de ângulo

Podemos aplicar restrições á posição relativa entre dois pares de átomos com coordenadas $\mathbf{r}_i$, $\mathbf{r}_j$ e $\mathbf{r}_k$, $\mathbf{r}_l$ através de uma função semelhante àquela utilizada para definir um diedro próprio. Na Equação 1.18 temos uma possível implementação, onde a constante de força $k_{rang}$ deve ser definida além das coordenadas das quatro partículas. O termo $m$ refere-se á multiplicidade, assumindo valor $m = 2$ se houver necessidade de distinção entre vetores paralelos e antiparalelos.

$$\begin{aligned} V_{rang}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) &= k_{rang}(1 - \cos(m(\theta - \theta^0))) \\ \theta &= \arccos\left( \frac{\mathbf{r}_j - \mathbf{r}_i}{\parallel \mathbf{r}_j - \mathbf{r}_i \parallel} \bullet \frac{\mathbf{r}_l - \mathbf{r}_k}{\parallel \mathbf{r}_l - \mathbf{r}_k \parallel} \right) \end{aligned} \tag{1.18}$$

## 1.5 Métodos de cálculo mecânico-quântico baseados na estrutura eletrônica

A mecânica quântica utiliza conceitos 'inovadores' que incluem dualidade onda–partícula e efeitos relativísticos para resolver problemas em escala atômica. O desenvolvimento de uma equação diferencial capaz de resultar uma função que pode prever a probabilidade de localização de uma partícula como o elétron, e dos métodos para solucionar esta equação são a base destes métodos. A equação de Schrödinger, que resulta uma função de onda de uma partícula, pode ser expressa como na Equação 1.19, onde $m_e$ é a massa do elétron, $h$ é a constante de Planck, $\mathbf{V}$ é o potencial no qual a partícula se move, $\Psi$ é a função de onda. $\|\Psi\|^2$ é interpretado como a densidade de probabilidade.

$$\left( \frac{-h^2}{8\pi^2 m_e} \nabla^2 + \mathbf{V} \right) \Psi(\tilde{\mathbf{r}}, t) = \frac{ih}{2\pi} \frac{\partial \Psi(\tilde{\mathbf{r}}, t)}{\partial t} \tag{1.19}$$

### 1.5.1 A Equação de Schrödinger Independente do Tempo

Se **V** não é função do tempo, então a Equação 1.19 pode ser simplifica usando técnicas matemáticas, como a separação de variáveis.

$$\Psi(\tilde{\mathbf{r}},t) = \Psi(\tilde{\mathbf{r}})\tau(t) \tag{1.20}$$

Substituindo a Equação 1.20 na Equação 1.19 temos:

$$\hat{H}\Psi(\tilde{\mathbf{r}}) = \varepsilon\Psi(\tilde{\mathbf{r}}) \tag{1.21}$$

Onde $E$ é a energia da partícula e $\hat{H}$ é o operador Hamiltoniano:

$$\hat{H} = -\frac{h^2}{8\pi^2 m_e}\nabla^2 + \hat{V} \tag{1.22}$$

#### O Hamiltoniano molecular

Para um sistema molecular, o operador Hamiltoniano depende da posição dos elétrons e núcleos dentro da molécula. O Hamiltoniano é construído em termos dos operadores correspondentes á energia cinética ($\hat{T}$) e potencial ($\hat{V}$) como:

$$\hat{H} = \hat{T} + \hat{V} \tag{1.23}$$

Nesta equação o termo de energia cinética depende apenas do movimento dos elétrons e dos núcleos. O termo de energia potencial depende das interações eletrostáticas entre elétrons–núcleos (atração), elétrons–elétrons (repulsão) e núcleos–núcleos(repulsão). A aproximação de Born–Oppenheimer permite que se considere a posição dos núcleos constante. Esta aproximação permite alguma simplificação nos termos do Hamiltoniano Molecular (mostrado na forma resumida na Equação 1.23), resultando no chamado Hamiltoniano Eletrônico. Mesmo considerando esta aproximação o termo de repulsão elétron–elétron resulta em uma expressão que não tem solução analítica para sistemas multieletrônicos.

As equações de Hartree–Fock na forma mostrada para sistemas atômicos podem ser estendidas para sistemas moleculares, mostradas nas Equações 1.24. Se $\chi$ são as funções de onda Hartree–Fock obtidas com o determinante de Slater, e $V_{NN}$ não envolve coordenadas eletrônicas, temos que $\langle\chi|V_{NN}|\chi\rangle = V_{NN}\langle\chi|\chi\rangle = V_{NN}$. O operador $\hat{H}_{el}$ é a soma dos operadores de um elétron $\hat{f}_i$ e do operador de dois elétrons $\hat{g}_{ij}$. Temos assim que $\hat{H}_{el} = \sum_i \hat{f}_i + \sum_j \sum_{i>j} \hat{g}_{ij}$, onde $\hat{f}_i = -\frac{h}{8\pi^2 m_e}\nabla_i^2 - \sum_\alpha \frac{Z_\alpha}{r_{i\alpha}}$ e $\hat{g}_{ij} = \frac{1}{r_{ij}}$. Os termos $J_{ij}$ e $K_{ij}$ são chamados de integrais de **Coulomb** e de **troca**, respectivamente.

$$\begin{aligned}
E_{HF} &= 2\sum_{i=1}^{\frac{n}{2}} H_{ii}^{core} + \sum_{i=1}^{\frac{n}{2}}\sum_{j=1}^{\frac{n}{2}}(2J_{ij} - K_{ij}) + V_{NN} \\
H_{ii}^{core} &\equiv \langle\phi_i(1)|\hat{H}_{(1)}^{core}|\phi_i(1)\rangle \\
\hat{H}_{ii}^{core} &\equiv -\frac{h}{8\pi^2 m_e}\nabla_1^2 - \sum_\alpha \frac{Z_\alpha}{r_{1\alpha}} \\
J_{ij} &\equiv \langle\phi_i(1)\phi_j(2)|\frac{1}{r_{12}}|\phi_i(1)\phi_j(2)\rangle \\
K_{ij} &\equiv \langle\phi_i(1)\phi_j(2)|\frac{1}{r_{12}}|\phi_j(1)\phi_i(2)\rangle
\end{aligned} \tag{1.24}$$

A utilização do método de Hartree–Fock para sistemas multieletrônicos implica o uso do método denominado do Campo Autoconsistente (SCF).

### O método variacional

Slater e Gaunt utilizaram o Método Variacional para resolver numericamente a equação de Schrödinger pela primeira vez em 1928. O método consiste na minimização da energia aproximada, $\varepsilon_{ap}$, calculada pelas Equações 1.25. Para isto é utilizada a derivada parcial da energia em relação à função de onda. Sabe-se que a energia aproximada calculada será sempre maior (menos negativa) que a energia exata ($\varepsilon_{ex}$). O valor mais negativo da energia exata pode ser aproximado com utilização de bases numéricas mais completas. A energia $\varepsilon_{ex}$ ainda não é o valor real obtido experimentalmente. O valor real não pode ser atingido por causa da aproximação do operador monoeletrônico: A energia de correlação eletrônica não pode ser calculada utilizando esta aproximação.

$$\begin{array}{rcl} \varepsilon_{ap} & = & \frac{\langle \Psi_{ap}|\mathcal{H}|\Psi_{ap}\rangle}{\langle \Psi_{ap}|\Psi_{ap}\rangle} \\ \frac{\partial \varepsilon_{ap}}{\partial \Psi_{ap}} & = & 0 \\ \varepsilon_{ex} & \leq & \varepsilon_{ap} \end{array} \tag{1.25}$$

O critério para que se considere a 'convergência' do sistema geralmente é a dupla desigualdade mostrada nas Equações 1.26, onde $\delta_i$ e $\delta_\varepsilon$ são os valores estipulados arbitrariamente como critério de convergência para a variação da densidade eletrônica calculada pela função de onda e para a energia, respectivamente. Estes valores são normalmente da ordem de $10^{-5}$ até $10^{-8}$ para $\delta_i$, a densidade eletrônica, e da ordem de $10^{-5}$ para $\delta_\varepsilon$, a energia. Outros critérios podem ser adotados também, como a utilização da relação entre energia cinética e potencial (Teorema do Virial).

$$\begin{array}{rcl} \|\Psi_i^{n-1}\|^2 - \|\Psi_i^n\|^2 & < & \delta_i \\ \|\varepsilon^{n-1} - \varepsilon^n\| & < & \delta_\varepsilon \end{array} \tag{1.26}$$

### Determinante de Slater

O produto da equação de Hartree (Equação 1.36) para a função de onda deve incluir o *spin* explicitamente, e deve ser antissimétrica em relação à troca de elétrons entre os orbitais.

A inclusão do *spin* por Fock e Slater em 1930, resultou na utilização de função antissimetrizada, para obter *spin*-orbitais. Estas funções de onda podem ser obtidas através do determinante de Slater, mostrado nas Equações 1.27.

$$\begin{array}{rcl} \alpha\psi_1 & = & S_1 \\ \beta\psi_2 & = & S_2 \\ \Psi & = & \frac{1}{\sqrt{n!}} \begin{vmatrix} S_1(1) & \cdots & S_n(1) \\ \vdots & \ddots & \vdots \\ S_1(n) & \cdots & S_n(n) \end{vmatrix} \end{array} \tag{1.27}$$

Portanto foi necessário desenvolver uma técnica matemática para obter uma solução satisfatória para a Equação 1.19 em sistemas multieletrônicos.

$$\hat{H}\Psi = \varepsilon\Psi \tag{1.28}$$

O Hamiltoniano da Equação 1.28 pode ser escrito como na Equação 1.29, onde $z$ é a carga nuclear e $\vec{r}$ é a distância elétron–núcleo.

$$\hat{H} = \left( -\frac{h}{8\pi^2 m_e}\nabla^2 - \frac{z}{\vec{\mathbf{r}}} \right) \tag{1.29}$$

Utilizando a equação de Schrödinger para um átomo com dois elétrons pode-se escrever o Hamiltoniano como mostrado na Equação 1.30, onde $z$ é a carga nuclear e $\vec{\mathbf{r_1}}$ e $\vec{\mathbf{r_2}}$ são as distâncias elétron(1)–núcleo e elétron(2)–núcleo.

$$\hat{H} = \left( -\frac{h}{8\pi^2 m_e}\nabla_1^2 - \frac{z}{\vec{\mathbf{r_1}}} \right) + \left( -\frac{h}{8\pi^2 m_e}\nabla_2^2 - \frac{z}{\vec{\mathbf{r_2}}} \right) + \frac{1}{\vec{\mathbf{r_{12}}}} \tag{1.30}$$

Segundo Hartree, pode-se obter o operador multieletrônico por combinações do operador monoeletrônico:

$$\mathcal{H} = \sum_{i=1}^{n} \hat{H}_i \tag{1.31}$$

A aproximação da partícula independente permite que o elétron(2) seja transformado em um campo de forças, que pode então ser aplicado ao elétron(1). O sistema núcleo–elétron(1) é um sistema monoeletrônico, que pode ser resolvido. Desta forma podem ser obtidos operadores monoeletrônicos para cada elétron de um sistema multieletrônico. Os Hamiltonianos têm a seguinte forma, onde $V_1^{e2}$ e $V_2^{e1}$ são os potenciais eletrostáticos médios criados pelos elétrons $e(1)$ e $e(2)$, que afetam os elétrons $e(2)$ e $e(1)$, respectivamente.

$$\begin{aligned} \hat{H}_1 &= \left(-\tfrac{h}{8\pi^2 m_e}\nabla_1^2 - \tfrac{z}{\vec{\mathbf{r_1}}} + V_1^{e2}\right) \\ \hat{H}_2 &= \left(-\tfrac{h}{8\pi^2 m_e}\nabla_1^2 - \tfrac{z}{\vec{\mathbf{r_2}}} + V_2^{e1}\right) \end{aligned} \tag{1.32}$$

O operador multieletrônico pode ser descrito como a soma dos operadores monoeletrônicos:

$$\mathcal{H} = \hat{H}_1 + \hat{H}_2 \tag{1.33}$$

O potencial eletrostático $\mathbf{V_1^{e2}}$ pode ser obtido pela Equação 1.34, onde $d\tau$ corresponde à integração por todo o espaço. O potencial eletrostático $\mathbf{V_2^{e1}}$ é obtido com por uma expressão análoga.

$$\mathbf{V_1^{e2}} = \int_0^{\infty} \frac{\|\Psi\|^2}{\|\vec{r_1} - \vec{r_2}\|} d\tau \tag{1.34}$$

O método SCF consiste na solução iterativa da equação de Schrödinger, através de uma estimativa inicial das funções de onda $\Psi_1^{(0)}$ e $\Psi_2^{(0)}$. O processo iterativo substitui as funções de onda $\Psi_1^{(0)}$ e $\Psi_2^{(0)}$ na Equação 1.28 utilizando os Hamiltonianos monoeletrônicos da Equação 1.29, calcula $\Psi_1^{(1)}$ e $\Psi_2^{(1)}$, obtém novos potenciais com a Equação 1.34. O cálculo prossegue até haver uma variação suficientemente pequena no potencial eletrostático calculado.

### Método de Hartree–Fock

A equação de Schrödinger pode ser resolvida exatamente para o átomo de hidrogênio, um sistema com um núcleo e um elétron, e para mais alguns poucos casos. Para sistemas multieletrônicos o método de Hartree–Fock é o procedimento normalmente adotado. Este método é a base da utilização de orbitais atômicos em sistemas multieletrônicos. O Hamiltoniano da Equação 1.28 para um sistema atômico multieletrônico pode ser escrito na forma da Equação 1.35, onde se assume que toda a massa do núcleo está restrita a um único ponto infinitesimal. A primeira somatória da Equação 1.35 contém o operador de energia cinética ($T$) para os $n$ elétrons. A segunda somatória é a energia potencial de atração núcleo–elétron $V_{Ne}$. A última é a energia potencial de repulsão eletrônica $V_{ee}$.

$$\hat{H} = \underbrace{-\frac{h}{8\pi^2 m_e}\sum_{i=1}^{n}\nabla^2}_{T} \underbrace{-\sum_{i=1}^{n}\frac{Z_e^2}{r_i}}_{V_{Ne}} \underbrace{+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{e^2}{r_{ij}}}_{V_{ee}} \tag{1.35}$$

A idéia do método Hartree–Fock é reduzir o problema de N elétrons para um problema de um elétron, que interage com os núcleos e a nuvem eletrônica dos demais elétrons. A interação elétron–elétron é introduzida de uma forma média. Hartree propôs a separação em $n$ equações monoeletrônicas. A função de onda assim obtida é útil para estimativa qualitativa, mas falha na descrição quantitativa. A utilização da carga nuclear total não é adequada. Este valor de carga nuclear pode ser modificado para cada orbital para levar em conta e blindagem eletrônica das camadas mais internas. A utilização de carga nuclear efetiva melhora os resultados obtido com esta aproximação, porém ainda não é suficiente. O próximo passo foi a utilização de funções variáveis, relaxando a restrição às funções utilizadas na descrição hidrogenóide. Assim, são determinadas funções $g$ para a Equação 1.36 que minimizem as Equações 1.25.

$$\phi = g_1(r_1, \theta_1, \phi_1)g_2(r_2, \theta_2, \phi_2)\cdots g_n(r_n.\theta_n, \phi_n) \tag{1.36}$$

A melhor solução para a Equação 1.36 pode ser aproximada por orbitais que são o produto de uma parte radial e uma função harmônica esférica, $g_i = h_i(r_i)Y_{l_i}^{m_i}(\theta_i, \phi_i)$. Hartree utilizou funções radiais normalizadas na Equação 1.36. Fock introduziu o conceito de antissimetrização nas equações de Hartree, pela utilização do determinante de Slater (ítem 1.5.1).

### As equações de Roothaan–Hall

As equações de Hartree–Fock utilizando determinante de Slater para antissimetrização foram resolvidas utilizando combinações lineares de conjuntos de funções de base que não eram modificados. As modificações ocorriam nos coeficientes de combinação das bases através de método Autoconsistente. Os cientistas que propuseram isto foram Roothaan e Hall. Inicialmente foram utilizadas *Slater Type Orbitals* (STO's) para montar os conjuntos de base. Boys propôs a utilização de *Gaussian Type Orbitals* (GTO's) na década de 50.

A expansão proposta por Roothaan dos orbitais em conjuntos de base tem a forma da expressão matemática mostrada nas Equações 1.37. Os conjuntos de base podem ser

substituídos na equação de Hartree–Fock como está mostrado nas equações. O sistema linear de $b$ equações homogêneas descreve o orbital molecular $\phi_i$. Uma solução não trivial deve ser obtida calculando o determinante das equações, como mostrado. As equações de Hartree–Fock–Roothaan devem ser resolvidas por processo iterativo, já que as integrais $F_{rs}$ dependem dos orbitais $\phi_i$ (através da dependência em $\hat{H}$ dos orbitais $\phi_i$), que por sua vez depende dos coeficientes $c_{si}$.

$$\begin{array}{rcl} \phi_i & = & \sum_{s=1}^{b} c_{si}\chi_s \\ \sum_s C_{si}\hat{F}\chi_s & = & \varepsilon_i \sum_s c_{si}\chi_s \\ \sum_{s=1}^{b} c_{si}(F_{rs} - \varepsilon_i S_{rs}) & = & 0 \; ; \; r = 1, 2, \cdots, n \\ F_{rs} \equiv \langle \chi_r | \hat{F} | \chi_s \rangle & & S_{rs} \equiv \langle \chi_r | \chi_s \rangle \end{array} \tag{1.37}$$

## 1.5.2 As funções de base

### Orbitais tipo Slater (STO)

Estes orbitais são utilizados com a parte angular igual a dos orbitais atômicos, e com a parte radial modificada conforme a Equação 1.38, onde $\zeta$ é o expoente orbital, $Y_l^m$ é a parte angular e $n, l, m$ são os números quânticos. Em cálculos simplificados o valor de $\zeta = \frac{Z-\wp}{n}$, onde $\wp$ é um fator ajustável que representa a blindagem.

$$\frac{(\frac{2\zeta}{a_0})^{n+\frac{1}{2}}}{((2n)!)^{\frac{1}{2}}} r^{n-1} \exp\left(\frac{-\zeta r}{a_0}\right) Y_l^m(\theta, \phi) \tag{1.38}$$

### Orbitais tipo Gaussiana (GTO)

Orbitais baseadas em STO são adequadas para uma boa representação de da equação de Schrödinger. Porém a utilização de funções tipo GTO é vantajosa porque permite a solução analítica das integrais de repulsão eletrônica de três e quatro elétrons. Outro fator que traz vantagens técnicas para utilização de GTO's é que o produto de duas Gaussianas é uma outra Gaussiana. Estas funções, porém, não têm boa reprodução do comportamento na região muito próxima à origem, um problema conhecido como condição de cúspide, nem da região assintótica. Este problema é contornado pela utilização de várias Gaussianas para representar cada STO. Um exemplo da forma de uma Gaussiana centrada num átomo $a$, em coordenadas Cartesianas, é mostrado na Equação 1.39. O parâmetro $N$ é a constante de normalização, e $\alpha$ é um expoente orbital positivo. Os valores de $i, j, k$ são inteiros não negativos. Se $i = 0, j = 0, k = 0$ a Gaussiana é do tipo $s$. Se $i + j + k = 1$ temos uma Gaussiana do tipo $p$, que contém os fatores $x_a, y_a, z_a$. Quando $i + j + k = 2$ é Gaussiana do tipo $d$ e tem os fatores $x_a^2, y_a^2, z_a^2, x_a y_a, x_a z_a, y_a z_a$. Uma função Gaussiana também pode ter expressão na forma radial.

$$g_{ikj} = N x_a^i y_a^j z_a^k \exp(-\alpha r_a^2) \tag{1.39}$$

### Base Mínima

Uma base mínima para um cálculo SCF molecular é aquela que utiliza uma única função de base para cada camada interna em cada átomo, assim como para cada camada de valência de cada átomo, quando se trabalham com funões de Slater. Considera-se a base Gaussiana mínima a que utiliza três Gaussianas para representar cada camada, sendo que as duas funções adicionadas são para representar corretamente a região assintótica e de cúspide.

### Base 3-21G

As bases tipo GTO usadas em cálculos *ab initio* devem levar em conta o compromisso entre custo computacional e acuracidade. Além da base mínima que utiliza apenas uma função por orbital e da base *split-valence* com duas funções de base para cada orbital temos diversas bases que utilizam funções de alto número quântico angular, denominadas funções de polarização. As funções de base normalmente são do tipo contraída, em que cada função é uma combinação linear de um certo número de primitivas Gaussianas. Os expoentes das primitivas Gaussianas são compartilhados entre funções do tipo '$s$' e '$p$' nas funções de valência.

No caso particular da base 3-21G para os elementos da primeira linha, temos a função da camada interna (cerne) tipo '$s$' formada por três Gaussianas, uma camada interna de valência de funções tipo '$s$' e '$p$' formada por duas Gaussianas, e outra camada de valência externa tipo '$sp$' formada por uma Gaussiana, em que todas as funções do conjunto de valência compartilham expoentes. O hidrogênio tem apenas duas funções tipo '$s$' na camada de valência [45, 46].

### Base 6-311G

Este conjunto de base foi desenvolvido por Pople e colaboradores [47], otimizada para obter a menor energia no estado fundamental no nível de perturbação de segunda ordem de Møller–Plesset, o que incorpora uma parte considerável da correção para a correlação dos elétrons de valência.

Um conjunto de base 6-311G utiliza um conjunto de base com Gaussianas contraídas (CGTF) obtida pela combinação de seis funções tipo GTO (ítem 1.5.2) para representar os orbitais internos (o cerne), exceto nos átomos de hidrogênio.

No próximo nível, a camada de valência, são utilizadas três divisões para as funções (*triple split*), em cinco funções tipo '$d$' descontraídas nos átomos pesados. Este conjunto de funções $(d_{3z^2-r^2},\ d_{xz},\ d_{yz},\ d_{x^2-y^2},\ d_{xy})\exp(-\alpha r^2)$ pode ser obtido das seis Gaussianas durante o desenvolvimento das integrais nos cálculos.

### O potencial efetivo no cerne (ECP)

O potencial efetivo substitui os cálculos com os elétrons do cerne por cálculos com somente elétrons de valência, e são derivados dos cálculos *ab initio* com todos os elétrons [48]. A obtenção dos potenciais efetivos inicia-se com a equação de Hartree–Fock

atômica para um orbital de valência com momentum angular $l$, conforme mostrado na Equação 1.40, onde $V_{core}$ e $V_{val}$ representam os operadores de Coulomb e de troca, somados sobre o cerne e sobre os orbitais de valência ocupados com coeficientes de acoplamento adequados, respectivamente. O orbital de valência $\phi_{li}$ é a solução de menor energia da Equação 1.40 somente se não há orbitais no cerne com o mesmo momentum angular $l$. Caso contrário o orbital de valência deve ser ortogonal a todos os orbitais do cerne com baixa energia que solucionam a equação de Hartree–Fock, sendo portanto nodal.

$$\left(-\frac{1}{2}\nabla^2{}_r - \frac{Z}{r} + \frac{l(l+1)}{2r^2} + V_{val} + V_{core}\right)\phi_{li} = \epsilon_{li}\phi_{li} \tag{1.40}$$

Obtidos $\phi_{li}$ e $\epsilon_{li}$ da solução auto-consistente da equação de Hartree–Fock é possível construir uma equação do tipo Hartree-Fock apenas com os orbitais de valência, que usa potencial efetivo para garantir que o orbital de valência é a solução de energia mais baixa. A Equação 1.40 pode ser escrita na forma da Equação 1.41, onde $V_l^{eff}$ é o potencial efetivo, $\chi_{li}$ é um pseudo-orbital não-nodal derivado de $\phi_{li}$ e $\epsilon_{li}$ é o autovalor da equação de Hartree–Fock original.

$$\left(-\frac{1}{2}\nabla^2{}_r - \frac{Z_{eff}}{r} + \frac{l(l+1)}{2r^2} + V'_{val} + V_l^{eff}\right)\chi_{li} = \epsilon_{li}\chi_{li} \tag{1.41}$$

Portanto $V_l^{eff}$ substitui a ortogonalidade explícita da Equação 1.40 e os potenciais de Coulomb e troca ($V_{core}$). A parte blindada pelo cerne da atração nuclear também é absorvida pelo potencial efetivo, portanto $Z_{eff}$ é igual ao à carga nuclear menos o número de elétrons do cerne.

O potencial efetivo $V_l^{eff}$ deve ser dependente do momentum angular $l$, já que o potencial de troca do cerne e as condições de ortogonalidade são dependentes de $l$. Entretanto, se o momentum angular do orbital de valência exceder o maior valor de $l$ do cerne, não há restrição da ortogonalidade radial e somente o potencial de troca do cerne requer dependência de $l$. Já que para os elementos da primeira e segunda linha da tabela periódica a interação de troca com o cerne é pequena para os elétrons de valência com alto momenta angular, os potenciais efetivos para todo valor de $l$ maior que os do cerne são dominados pelo potencial de Coulomb do cerne, e são portanto muito parecidos.

O potencial efetivo total para cada átomo é dado pela Equação 1.42, onde a utilização de projeção garante que os potenciais estão conectados com a componente adequada de $l$ da função de onda de valência. O valor máximo de $l$ do cerne é designado $L$.

$$V^{eff}(r) = V_{L+1}^{eff}(r) + \sum_{l=0}^{L}[V_l^{eff}(r) - V_{L+1}^{eff}(r)]\sum_m |lm\rangle\langle lm| \tag{1.42}$$

Os pseudo-orbitais $\chi$ construídos devem ser tão semelhantes quanto possível aos orbitais $\phi$, para garantir uma correta sobreposição de orbitais de valência interatômicos e também que os termos de interação valência–valência atômicos que surgem em $V'_{val}$ na Equação 1.41 sejam similares aos existentes em $V_{val}$ na Equação 1.40.

Podemos definir orbitais que atendem estes requisitos através da Equação 1.43, onde para qualquer ponto entre $R_l$ e o infinito, os pseudo-orbitais são idênticos aos orbitais de

valência Hartree-Fock. Para distâncias radiais menores que $R_l$ o pseudo-orbital é definido por uma expansão polinomial que tende a zero suavemente. Os coeficientes do polinomial são definidos de modo que sejam reproduzidos o valor e as três primeiras derivadas de $\phi_{li}$ em $R_l$ e que $\chi_{li}$ seja normalizada. O valor de $N$ utilizado é $N = 3$, porém valores $N = l + 2$ também podem ser utilizados.

$$\begin{array}{rcll} \chi_{li}(r) & = & \sum_{k=0}^{4} c_k r^{N+k} & r \leq R_l \\ \chi_{li}(r) & = & \phi_{li}(r) & r \geq R_l \end{array} \tag{1.43}$$

Uma vez que os pseudo-orbitais e os autovalores são conhecidos é possível determinar o valor 'exato' do potencial efetivo invertendo a Equação 1.41, que resulta na Equação 1.44. A utilização da Equação 1.44 permite calcular valores numéricos para os pseudo-potenciais a partir de valores numéricos de pseudo-orbitais. Em cálculos moleculares os valores dos potenciais geralmente são ajustados por combinações lineares de funções analíticas. Um ajuste suficientemente acurado necessita de diversos termos na expansão da função analítica.

As funções de base construídas através do formalismo descrito são suficientemente acuradas para substituir, com vantagens na eficiência, a utilização das bases mínima STO-3G e mesmo 4-31G com todos os elétrons [49, 50].

$$V_l^{eff}(r) = \epsilon_{li} + \frac{Z_{eff}}{r} - \frac{l(l+1)}{2r^2} + \frac{(\frac{1}{2}\nabla^2 - V'_{val})\chi_{li}}{\chi_{li}} \tag{1.44}$$

### 1.5.3 A inclusão do efeito de solvatação

Inicialmente os modelos de solvatação trataram o solvente de forma implícita, direcionados aos aspectos dielétricos do efeito eletrostático, assumindo que o solvente é um meio contínuo e isotrópico caracterizado por somente uma constante dielétrica escalar e estática, $\epsilon$. Estes modelos ficaram conhecidos como *continuum models*. Modelos assim, desenvolvidos por Born, Onsanger e Kirkwood mostraram-se rapidamente aplicáveis apenas em algumas situações específicas.

Há duas vantagens principais no uso de modelos tipo *continuum*: A primeira é a redução do número de graus de liberdade do sistema. A representação explícita de duzentas moléculas de água adiciona mil e oitocentos graus de liberdade ao sistema. A segunda é que os modelos explícitos também são muito imperfeitos. A não ser que se utilize um alto grau de refinamento, não são capazes de reproduzir a polarização elétrica do solvente tão bem como os métodos implícitos que utilizam dados experimentais da constante dielétrica.

Um dos problemas cruciais no tratamento implícito do solvente diz respeito à primeira camada de solvatação. Suas propriedades diferenciadas dificultam a definição do ponto onde efetivamente inicia-se o *continuum*. Aspectos como as interações de dispersão, as pontes de hidrogênio, o efeito hidrofóbico, a transferência de carga soluto–solvente (particularmente em solvente não condutores), contribuem para que a representação correta dos efeitos da primeira camada de solvente sejam um enorme desafio para o desenvolvimento e a implementação qualquer tipo de modelo de solvatação [51].

### O modelo básico

A formulação do modelo quântico de *continuum* polarizável requer a definição simultânea de dois problemas: (1) o problema de mecânica quântica de calcular a distribuição eletrônica $\varrho_M$, definida como a soma das cargas nucleares discretizadas e da função de densidade eletrônica com núcleos fixos, na presença do campo eletrostático gerado pelo dielétrico polarizado, $\phi_\sigma$; (2) o problema eletrostático para determinar o potencial de reação eletrostática do solvente, $\Phi_\sigma$, e sua interação energética com a distribuição de cargas, $\varrho_M$.

Trata-se de um típico problema não-linear, porquê tanto $\Phi_\sigma$ quanto $\varrho_M$ dependem de $\mathcal{V}$. Um método iterativo pode ser utilizado, resolvendo alternadamente o problema quântico para determinar $\Phi_\sigma$ e o problema clássico para determinar $\varrho_M$, partindo de uma estimativa razoável de $\Phi_\sigma$ e $\varrho_M$.

Para elaboração do modelo de solvatação deve ser definido um Hamiltoniano molecular para a espécie solvatada. Na Equação 1.45 $\mathcal{H}_M$ é o Hamiltoniano do soluto, que depende das coordenadas dos núcleos (fixas) e dos elétrons, $\mathcal{H}_M^0$ é o Hamiltoniano do soluto no vácuo e $\mathcal{V}_f$ é um potencial que será definido de acordo com o modelo de solvatação utilizado. Toda a informação relevante sobre a interação soluto–solvente está contida na função de onda $\Psi^f$ e nos autovalores $\varepsilon^f$.

$$\begin{array}{rcl} \mathcal{H}_M & = & \mathcal{H}_M^0 + \mathcal{V}_f \\ \mathcal{H}_M \Psi^f & = & \varepsilon^f \Psi^f \end{array} \tag{1.45}$$

A carga molecular pode ser utilizada em substituição à função de onda conforme as Equações 1.46, onde $\varrho_M$ é a soma das cargas nucleares discretizadas, $\varrho_{nuc}$, e da função densidade eletrônica, $\varrho_{el}$. O valor de $Z_\alpha$ é dado pelas cargas nucleares e o índice $\alpha$ diz respeito a cada um dos núcleos. O sinal negativo na integração da densidade eletrônica é a carga do elétron, e *nel* refere-se ao número total de elétrons.

$$\begin{array}{rcl} \varrho_M^{(\mathbf{r},\mathbf{Q})} & = & \varrho_{nuc}^{(\mathbf{r},\mathbf{Q})} + \varrho_{el}^{(\mathbf{r},\mathbf{Q})} \\ \varrho_{nuc}^{(\mathbf{r},\mathbf{Q})} & = & \sum_\alpha Z_\alpha \delta(\mathbf{r} - \mathbf{Q}_\alpha) \\ \varrho_{el}^{(\mathbf{q_1},\mathbf{Q})} & = & -\int |\Psi^f(\mathbf{q},\mathbf{Q})|^2 d\mathbf{q_2}, \ldots, d\mathbf{q_{nel}} \end{array} \tag{1.46}$$

O termo $\mathcal{V}$ contém informação sobre a função de distribuição do solvente ao redor do soluto ($g_s$), relacionada á temperatura média. A extensão da informação incluída pode ser definida durante a elaboração da função, e do próprio modelo de solvatação. Considerando apenas os efeitos da polarização eletrostática podemos definir um potencial de interação aproximado, $\mathcal{V}_\sigma$, através das Equações 1.47, onde $\Phi_\sigma(\mathbf{r})$ é o valor do campo eletrostático gerado pelo dielétrico polarizado a uma distância $\mathbf{r}$ e $W_{MS}$ é a contribuição da interação soluto–solvente para a energia total $\varepsilon^\sigma$.

$$\begin{array}{rcl} \mathcal{V}_\sigma(\mathbf{q},\mathbf{Q}\varrho_M,\epsilon) & = & \sum_\alpha Z_\alpha \Phi_\sigma(\mathbf{Q}_\alpha) - \sum_i \Phi_\sigma(\mathbf{q_i}) \\ W_{MS} & = & \oint \varrho_M(\mathbf{r}) \Phi_\sigma(\mathbf{r}) d\tau \end{array} \tag{1.47}$$

O modelo *continuum* básico envolve um elevado grau de simplificação de $\mathcal{V}$, que fica reduzido á componente clássica da interação. Se considerado assim, $g_s$ descreve o *continuum*

linear isotrópico, caracterizado por $\epsilon$, a constante dielétrica do solvente, dependente da temperatura. Modelos construídos desta forma não são suficientemente acurados, sendo necessária uma maior complexidade na formulação do termo $\mathcal{V}$.

### O modelo D-PCM

O problema eletrostático para determinar $\Phi_\sigma$, e sua interação com a distribuição de cargas, $\varrho_M$ pode ser resolvido de maneiras diferentes. A implementação do efeito de solvatação no programa GAMESS inclui o método conhecido como *Dieletric Polarizable Continuum Method*, ou D-PCM, além de outros efeitos para calcular a energia de formação da cavidade.

Inicialmente deve ser definido o espaço ocupado pelo soluto, pela criação da cavidade no solvente. No método denominado D-PCM é utilizada uma cavidade do tipo molecular, definida como uma união de esferas com raios determinados empiricamente. A cavidade molecular utilizada deve ser diferenciável em todas as suas partes, o que pode ser conseguido com união de esferas e elipsóides, e deve ter tamanho suficiente para acomodar toda a distribuição de cargas $\varrho_M$, sem espaço vazio que poderia ser ocupado por solvente [52]. O raio de van der Waals multiplicado por um fator $f$ tal que $1,40 > f > 1,10$ foi estabelecido como sendo adequado para a maioria dos casos, por diversos autores [53].

Estabelecido o formato e o tamanho da cavidade, temos o problema eletrostático. A constante dielétrica pode assumir valores dentro ($V_{in}$) ou fora ($V_{out}$) da cavidade: $\epsilon(r) = 1 \Rightarrow r \in V_{in}$ ou $\epsilon(r) = \epsilon \Rightarrow r \in V_{out}$. A carga $\varrho_M$ está confinada na cavidade, $\varrho_M(r) = 0 \Rightarrow r \in V_{out}$. O potencial eletrostático total pode ser definido pelas Equações 1.48.

$$\begin{array}{rcll} \nabla^2\Phi(\mathbf{r}) & = & -4\pi\varrho_M & \Rightarrow \quad \mathbf{r} \in V_{in} \\ \nabla^2\Phi(\mathbf{r}) & = & 0 & \Rightarrow \quad \mathbf{r} \in V_{out} \end{array} \tag{1.48}$$

Podemos ainda definir que na região assintótica, muito distante do soluto $M$ e para pontos muito próximos do limite da cavidade, valem as condições de contorno mostradas nas Equações 1.49, onde $\alpha$ e $\beta$ têm valores finitos, os valores de $\Phi_{in}$ e de $\Phi_{out}$ são os potenciais eletrostáticos nos limites interno e externo da cavidade, $\partial/\partial\mathbf{n}$ é a derivada na direção perpendicular à superfície da cavidade e o vetor $n$ aponta para fora da cavidade.

$$\begin{array}{rcl} \lim_{r\to\infty} r\Phi(\mathbf{r}) & = & \alpha \\ \lim_{r\to\infty} r^2\nabla\Phi(\mathbf{r}) & = & \beta \\ \Phi_{in} & = & \Phi_{out} \\ \frac{\partial\Phi_{in}}{\partial\mathbf{n}} & = & \epsilon\frac{\partial\Phi_{out}}{\partial\mathbf{n}} \end{array} \tag{1.49}$$

Para uma dada distribuição de cargas $\varrho_M$, o potencial eletrostático $\Phi_\sigma$ difere daquele calculado no vácuo já que sofre influência do campo de reação gerado pela distribuição de cargas do solvente. A integral de interação soluto–solvente, $W_{MS}$, fica limitada ao volume da cavidade se $\varrho_M = 0$ fora da cavidade. O trabalho necessário para trazer uma carga $\varrho_M$ para dentro da cavidade em um modelo rígido é $\Delta G_{el}^{(0)} = \frac{W_{MS}}{2}$. O índice $(i)$ denota o nível de aproximação da descrição do efeito de polarização mútuo soluto–solvente: $i = 0$ supõe cargas rígidas; $i = 1$ supõe polarização do soluto em resposta ao campo do solvente; $i = f$ supõe que tanto o soluto como o solvente polarizam-se.

Claverie e Huron aplicaram as relações estabelecidas nas Equações 1.47 e 1.49 à geometria de uma cavidade do tipo molecular, obtida pela união de esferas, obtendo a definição do potencial $\Phi$ dentro e fora da cavidade. A formulação matemática foi desenvolvida na série de artigos publicada no início da década de setenta [52, 54, 55], e não será reproduzida aqui dada a sua grande complexidade. O desenvolvimento leva em conta várias aproximações e a inclusão de constantes para a calibração adequada dos raios das esferas utilizadas na definição da cavidade.

No interior da cavidade $\Phi_{in}$ foi definido pela expansão de um conjunto de base formado por harmônicos, $r^l Y_l^m(\theta, \phi)$, centrados num único ponto, e no exterior $\Phi_{out}$ é definido pela expansão de diversos conjuntos de harmônicos, centrados nos pontos onde são posicionadas as cargas localizadas. Este é o método conhecido como Expansão Multipolar (MPE).

### O potencial de reação eletrostática do solvente

A abordagem ao problema de mecânica clássica para determinar o potencial de reação eletrostática do solvente, $\Phi_\sigma$, adotada na implementação do método D-PCM no programa GAMESS, é conhecida como *Aparent Surface Charge* (ASC).

A cavidade de solvente é dividida em porções menores através de um algoritmo conhecido como GEPOL (*Generation of Polyhedra*) durante a discretização para definição das cargas aparentes. Estas porções menores são denominadas *tesserae*, às quais um conjunto de cargas aparentes $q_i$ são unívocamente associadas. Cada valor é definido por duas componentes distintas, $q_i = q_i^N + q_i^e$, onde $q_i^N$ é a contribuição das cargas nucleares na superfície e $q_i^e$ a contribuição da componente eletrônica. A soma destas cargas, $\mathcal{Q}_\sigma^T = \sum_i^t q_i$, geralmente não corresponde exatamente á carga total do soluto, $\mathbf{Q}_M$, uma vez que a distribuição de cargas do soluto não termina na cavidade, particularmente $q_i^e$ extende-se além desta e decai exponencialmente. Para compensar este erro um procedimento de renormalização deve ser aplicado, multiplicando os valores das cargas aparentes por um fator $\mathcal{F}$ para que o resultado da sua soma fique igual a carga total do soluto. O procedimento de normalização dado pelas Equações 1.50 mostrou-se simples e eficiente, exceto em casos onde há grande assimetria na distribuição da cargas. Notadamente as cavidade que envolvem *zwiterions* não são bem representados assim, pois é claro que o fator utilizado para normalização da carga aparente na porção da molécula onde situa-se a carga negativa não será adequado para a porção onde situa-se a carga positiva [56].

$$\begin{aligned} \mathcal{F}^N \textstyle\sum_i q_i^N &= -\frac{\epsilon-1}{\epsilon} \mathcal{Q}_M^N \\ \mathcal{F}^e \textstyle\sum_i q_i^e &= -\frac{\epsilon-1}{\epsilon} \mathcal{Q}_M^e \\ \mathbf{Q}_M &= \mathcal{F}^e \textstyle\sum_i q_i^e + \mathcal{F}^N \sum_i q_i^N \end{aligned} \tag{1.50}$$

Para considerar a porção da carga localizada na parte exterior da cavidade devemos adicionar outra fonte de cargas, $\rho_b^{out}(\mathbf{r})$, para corrigir a carga aparente na superfície de cada *tesserae*. A distribuição de cargas da superfície, $\sigma(s)$ deve ser somada a uma componente de carga posicionada no meio dielétrico. O tratamento completo do problema eletrostático da distribuição de carga difusa na presença de uma cavidade em um meio dielétrico uniforme e isotrópico é dado nas Equações 1.51, onde $\hat{n}_s$ é o vetor unitário na superfície que aponta na direção do meio e $\mathbf{E}$ é o campo elétrico total criado pelas

três fontes somadas: A distribuição de cargas do soluto $M$, da superfície $\sigma(s)$ e do meio $\rho_b^{out}(\mathbf{r})$, $\mathbf{E} = \mathbf{E}_{\sigma(s)} + \mathbf{E}_{\rho_b(\mathbf{r})} + \mathbf{E}_M$.

$$\begin{array}{rcll} \sigma(s) & = & -\frac{\epsilon-1}{4\pi\epsilon}\mathbf{E}(s)\cdot\hat{n} & \Rightarrow \text{na superfície} \\ \rho_b(\mathbf{r}) & = & -\frac{\epsilon-1}{4\pi}\nabla\cdot\mathbf{E}(\mathbf{r}) & \Rightarrow \text{no meio} \end{array} \tag{1.51}$$

Porém, a descrição do campo elétrico fora da cavidade utilizando o formalismo já descrito é problemática. O decaimento é muito rápido, e o valor tende a zero em alguns décimos de angstron, sendo necessária uma malha extremamente refinada para a correta descrição do efeito. No entanto, podemos considerar que o efeito da carga volumétrica aparente do meio é completamente descrita pelo seu potencial de reação correspondente, mostrado na Equação 1.52.

$$V_\rho(\mathbf{r}) = \int_{bulk} \frac{\rho_b(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \tag{1.52}$$

O intervalo de integração (*bulk*) extende-se por todo o espaço não ocupado pela cavidade do solvente, $\mathbf{r}'$. Aplicando-se o teorema de Green o volume pode ser substituído pela integração no intervalo da superfície $S(C)$, conforme a Equação 1.53. Assim o que é calculado na prática é um potencial efetivo, denominado $\sigma_{eff}(\mathbf{s})$.

$$\sigma_{eff}(\mathbf{s}) = \frac{\mathcal{Q}_b\rho(\mathbf{r})}{\int_{S(C)}\rho(\mathbf{s}')} d\mathbf{s}' \tag{1.53}$$

A carga volumétrica total aparente, $\mathcal{Q}_b$ pode ser obtida a partir das Equações 1.54 onde $\mathcal{Q}_{out}$ corresponde à carga total do soluto negligenciada pela cavidade, $\mathcal{Q}_{in}$ é a carga dentro da cavidade e $\rho(\mathbf{s})$ é a carga eletrônica do soluto calculada no ponto $\mathbf{s}$ da superfície.

$$\begin{array}{rcl} \mathcal{Q}_b & = & -\frac{\epsilon-1}{\epsilon}\mathcal{Q}_{out} \\ \mathcal{Q}_{out} & = & -(N_{el}-\mathcal{Q}_{in}) \\ & = & -\left(N_{el} + \frac{1}{4\pi}\int_{S(C)}\mathbf{E}_M(\mathbf{s})\cdot\hat{n}_s d\mathbf{s}\right) \end{array} \tag{1.54}$$

Uma vantagem da formulação ASC em relação à formulação MPE para o problema eletrostático é a possibilidade da determinação da relação entre a distribuição de cargas, $\varrho$, e a carga total do soluto $\mathbf{Q}_M$. Particularmente no caso da necessidade de incluir os termos de adição do potencial fora da cavidade na determinação das cargas superficiais, a formulação MPE tem solução muito complicada, tendo sido publicada apenas para o átomo de hidrogênio por Chipman [57].

### A determinação do problema quântico

O problema quântico, na formulação D-PCM pode ser tratado no nível de cálculo Hartree-Fock, ou pós HF por Interação de Configurações (CI) ou Campo Autoconsistente multiconfiguracional (MCSCF). A formulação no nível Hartree–Fock será apresentada de maneira resumida. Formalmente trata-se simplesmente de adicionar o termo $\mathcal{V}$ á equação

de Schrödinger, conforme mostrado nas Equações 1.45. A inclusão deste termo de interação soluto–solvente, $W_{MS}$, pode ser feita em partes como mostrado nas Equações 1.55.

$$\begin{aligned} W_{MS} &= W_{nuc,nuc} + W_{nuc,el} + W_{el,nuc} + W_{el,el} \\ W_{nuc,nuc} &= \oint \varrho_{nuc}(\mathbf{r},\mathbf{Q})\Phi_{\sigma,nuc}(\mathbf{r},\mathbf{Q})d\tau \\ W_{nuc,el} &= \oint \varrho_{nuc}(\mathbf{r},\mathbf{Q})\Phi_{\sigma,el}(\mathbf{r},\mathbf{Q})d\tau \\ W_{el,nuc} &= \oint \varrho_{el}(\mathbf{r},\mathbf{Q})\Phi_{\sigma,nuc}(\mathbf{r},\mathbf{Q})d\tau \\ W_{el,el} &= \oint \varrho_{el}(\mathbf{r},\mathbf{Q})\Phi_{\sigma,el}(\mathbf{r},\mathbf{Q})d\tau \end{aligned} \tag{1.55}$$

Utilizando o método de Hartree–Fock, resolvemos a Equação secular $\mathbf{FC} = \varepsilon\mathbf{SC}$ após a expansão das funções de um elétrons em bases $\{\cdots|\mu\rangle\cdots|\nu\rangle\}$. Os elementos $\mathbf{F}$ da matriz de Fock são modificados pela adição da contribuição de $\mathcal{V}_\sigma$ aos termos monoeletrônicos $\mathbf{h}$ e também pela dependência do termo dieletrônico $\mathbf{G}$ da densidade eletrônica $\varrho_{el}$ conforme a Equação 1.56

$$F_{\mu\nu} = h_{\mu\nu} + G_{\mu\nu}(\varrho_{el}) + \langle\mu|\mathcal{V}_\sigma(\varrho_{el})|\nu\rangle \tag{1.56}$$

### 1.5.4 Métodos Semi-empíricos

Os cálculos semi-empíricos apresentam a mesma estrutura dos cálculo HF, com algumas modificações, tais como as integrais de dois elétrons que são aproximadas ou completamente desprezadas. Para corrigir os erros introduzidos por esta omissão, o método é parametrizado para reproduzir da melhor forma os resultados experimentais [58].

#### Zero Differential Overlap (ZDO)

Esta aproximação cancela integrais com mais de 2 centros na matriz de Fock. As integrais de repulsão elétron–elétron seguem o esquema proposto nas Equações 1.57.

$$\begin{aligned} (\mu^A\mu^A|\lambda^A\lambda^A) &\implies \text{1 centro:} && \text{fica} \\ (\mu^A\mu^A|\lambda^A\sigma^B) &\implies \text{2 centros, } \lambda \neq \sigma: && \text{desaparece} \\ (\mu^A\mu^A|\lambda^B\lambda^B) &\implies \text{2 centros, } \lambda = \lambda: && \text{fica} \\ (\mu^A\mu^A|\mu^A\mu^A) &\implies \text{1 centro :} && \text{fica} \end{aligned} \tag{1.57}$$

Os elementos da matriz de Fock podem então ser descritos da forma mostrada nas Equações 1.58, onde $P$ representa a matriz de densidades, $t$ é o operador de energia cinética e $V_{ne} \equiv -\frac{Z_A}{R_{Ae}}$ ou $V_{ne} \equiv -\frac{Z_B}{R_{Ae}}$. Os elementos que não tiverem a mesma simetria orbital desaparecem.

$$\begin{aligned} \mathbf{FC} &= \mathbf{SCE} \\ F_{\mu\mu} &= H_{\mu\mu} + \textstyle\sum_\lambda P_{\lambda\lambda}[(\mu\mu|\lambda\lambda) - \frac{1}{2}\underbrace{(\mu\mu|\mu\mu)}_{\varsigma_{AA}}] \\ F_{\mu\nu} &= H_{\mu\nu} - \frac{1}{2}\underbrace{(\mu\mu|\nu\nu)}_{\varsigma_{AB}} P_{\mu\nu} \\ H_{\mu\mu} &= \underbrace{(\mu|t|\mu) - (\mu|V_A|\mu)}_{V_{\mu\mu}} - \textstyle\sum_{B\neq A}\underbrace{(\mu|V_B|\mu)}_{V_{AB}} \\ H_{\mu\nu} &= (\mu|t|\nu) - \underbrace{(\mu|V_A|\nu)}_{=0} - \underbrace{(\mu|V_B|\nu)}_{=0} - \textstyle\sum_{C\neq A,B}\underbrace{(\mu|V_C|\nu)}_{=0} \end{aligned} \tag{1.58}$$

O operador de energia cinética, $t \equiv \frac{\partial^2}{\partial f^2}$, pode ser modificado nas funções de base, o que faz com que as integrais não se anulem.

As integrais do tipo $p$ são substituídas por integrais do tipo $s$. Isso é feito porque as funções tipo $s$ têm, simetria esférica e não têm problema de variância rotacional. Por exemplo $(2s^A 2s^A|2s^B 2s^B) \equiv \varsigma_{AB}$ ; $(2p_x^A 2p_x^A|2p_x^B 2p_x^B) \equiv \varsigma_{AB}$.

Pople utilizou os Potenciais de Ionização (I) e as Afinidades Eletrônicas (A) para calcular o potencial eletrostático, além de outras aproximações para simplificar o sistema de equações. Estas aproximações estão mostradas nas Equações 1.59.

$$\begin{aligned}
V_{AB} &= Z_B \varsigma_{AB} \\
H_{\mu\nu} &= (\mu|t|\nu) \\
&= \beta_{AB} S_{\mu\nu} \\
\beta_{AB} &\equiv -\tfrac{1}{2}(\beta_A + \beta_B) \\
-I_\mu &= V_{\mu\mu} + (Z_A - 1)\varsigma_{AA} \\
-A_\mu &= V_{\mu\mu} + Z_A \varsigma_{AA} \\
F_{\mu\mu} &= -\tfrac{1}{2}(I_\mu + A_\mu) + (P_{AA} - Z_A) \\
&- \tfrac{1}{2}(P_{\mu\mu} - 1)\varsigma_{AA} + \textstyle\sum_{B \neq A}(P_{BB} - Z_B)\varsigma_{AB} \\
F_{\mu\nu} &= \beta_{AB} S_{\mu\nu} - \tfrac{1}{2} P_{\mu\nu} \varsigma_{AB}
\end{aligned} \tag{1.59}$$

### Neglect of Diatomic Differential Overlap (NDDO)

A aproximação conhecida como *Modified Neglect of Diatomic Overlap* (MNDO) [59] é uma das mais conhecidas, que utilizam o conceito de NDDO. Todos os métodos semi-empíricos fazem alguma simplificação dos elementos de Fock, e tentam recuperar isto com a adição de algum termo.

Nas aproximações anteriores à MNDO, conhecidas como CNDO e INDO, as integrais de repulsão $(\mu\mu, \nu\nu)$ entre qualquer Orbital Molecular (OM) em um átomo $A$ e qualquer OM em um átomo $B$ são consideradas iguais. Estas integrais não são iguais, e na aproximação denominada NDDO são consideradas diferentes, além de um número adicional de integrais de dois centros que são desconsideradas em CNDO e INDO. Uma abordagem denominada MINDO, uma modificação do INDO, certo nível de correlação é adicionada pela correta modificação da integrais de repulsão.

O tratamento dado com o método MNDO, reproduzido aqui, é restrito aos casos de camada fechado, aos elétrons da camada de valência. Os OMs $\Psi_i$ das camadas de valência são representados por uma combinação linear de Orbitais Atômicos (OAs) $\psi_i$ com base mínima, conforme a Equação 1.60.

$$\Psi_i = \sum_\nu C_{\nu i} \psi_i \tag{1.60}$$

Os coeficientes $C_{\mu i}$ são determinados pelas equações de Roothaan-Hall, e assumem a forma mostrada na Equação 1.61, onde $\delta_{\nu\mu}$ é o delta de Kronecker e $E_i$ é o autovalor da OM $\Psi_i$.

$$\sum_\nu (F_{\nu\mu} - E_i \delta_{\nu\mu}) C_{\nu i} = 0 \tag{1.61}$$

O valor de cada elemento $F_{\nu\mu}$ da matriz de Fock é uma soma de uma parte de um elétron correspondente ao cerne e de uma parte correspondente a dois elétrons. Considerando que os OAs com índice $\nu$ e $\mu$ são localizados sobre o átomo $A$ e que os que tem índice $\lambda$ e $\sigma$ são localizados no átomo $B$ podemos definir os elementos $F$ através das Equações 1.62.

$$\begin{aligned} F_{\mu\nu} &= \textstyle\sum_B V_{\mu\nu,B} + 1/2 P_{\mu\nu}[3(\mu\nu,\mu\nu) - (\mu\mu,\nu\nu)] \\ &+ \textstyle\sum_B \sum_{\lambda,\sigma}^{B} P_{\lambda\sigma}(\mu\nu,\lambda\sigma) \\ F_{\mu\mu} &= U_{\mu\mu} + \textstyle\sum_B V_{\mu\mu,B} + \sum_{\nu}^{A} P_{\nu\nu}[(\mu\mu,\nu\nu) - 1/2(\mu\nu,\mu\nu)] \\ &+ \textstyle\sum_B \sum_{\lambda,\sigma}^{B} P_{\lambda\sigma}(\mu\mu,\lambda\sigma) \\ F_{\mu\lambda} &= \beta_{\mu\lambda} - 1/2 \textstyle\sum_{\nu}^{A} \sum_{\sigma}^{B} P_{\nu\sigma}(\mu\nu,\lambda\sigma) \end{aligned} \tag{1.62}$$

Os elementos que aparecem na matriz de Fock são os seguintes:

- Energias de um centro e um elétron $U_{\mu\mu}$ que representam a soma da energia cinética de um elétron no OA $\psi_\mu$ do átomo $A$ e da energia potencial devida a atração pelo cerne do átomo $A$.
- Integrais de um centro e dois elétrons (Coulomb) $(\mu\mu,\nu\nu)$ e de troca $(\mu\nu,\mu\nu)$.
- Integrais de ressonância de dois centros e um elétron $\beta_{\mu\lambda}$.
- Atrações de dois centros e um elétron $V_{\nu\mu,B}$ entre um elétron na distribuição $\psi_\mu\psi_\nu$ do átomo $A$ e o cerne do átomo $B$.
- Integrais de repulsão de dois elétrons $(\nu\mu,\lambda\sigma)$.

Com as aproximações utilizadas no método MNDO vários termos da matriz de Fock e das repulsões do cerne não são calculados analiticamente. São determinadas empiricamente a partir de dados experimentais ou através de expressões semi-empíricas que contém parâmetros numéricos que podem ser ajustados para reprodução dos dados experimentais.

As integrais de repulsão de dois centros $(\nu\mu,\lambda\sigma)$ representam a energia de interação entre a distribuição de carga $e\psi_\mu\psi_\nu$ no átomo $A$ e a distribuição $e\psi_\lambda\psi_\sigma$ no átomo $B$, considerando que $e$ é a carga elementar. Classicamente estes termos são iguais a soma de todas as interações entre os momenta de multipolo $M_{lm}$ das duas distribuições de carga, onde os subscritos $l$ e $m$ correspondem à ordem e à orientação dos multipolos. No MNDO as interações entre os multipolos são calculadas através da expressão das Equações 1.63, que substitui as integrais $(\nu\mu,\lambda\sigma)$.

$$\begin{aligned} (\nu\mu,\lambda\sigma) &= \textstyle\sum_{l1}\sum_{l2}\sum_m [M^A_{l1m}, M^B_{l2m}] \\ \textstyle\sum_{l1}\sum_{l2}\sum_m [M^A_{l1m}, M^B_{l2m}] &= \frac{e^2}{2^{l_1+l_2}} \textstyle\sum_{i=1}^{2^{l_1}} \sum_{j=1}^{2^{l_2}} f_1(R_{ij}) \end{aligned} \tag{1.63}$$

Nas Equações 1.63 $f_1(R_{ij})$ é uma função definida para ter o comportamento correto no intervalo de $R \to (0,\infty)$. Para calcular corretamente as distâncias $R_{ij}$ para uma dada distância interatômica $R_{AB}$ as configurações de cargas puntuais mostradas na Figura 1.3. As separações de carga $D_l$ nas configurações de dipolo e quadrupolo são determinadas pela

condição que o momento multipolar de cada configuração fique igual ao da distribuição de carga correspondente. Para átomos da primeira linha as expressão são mostradas nas Equações 1.64, onde $\zeta_{2s}$ e $\zeta_{2p}$ são os expoentes da Slater para os OMs $2s$ e $2p$.

$$\begin{aligned} D_1 &= \frac{5}{\sqrt{3}} \frac{(4\zeta_{2s}\zeta_{2p})^{5/2}}{(\zeta_{2s}+\zeta+2p)^6} \\ D_2 &= \frac{\sqrt{3/2}}{\zeta_{2p}} \end{aligned} \tag{1.64}$$

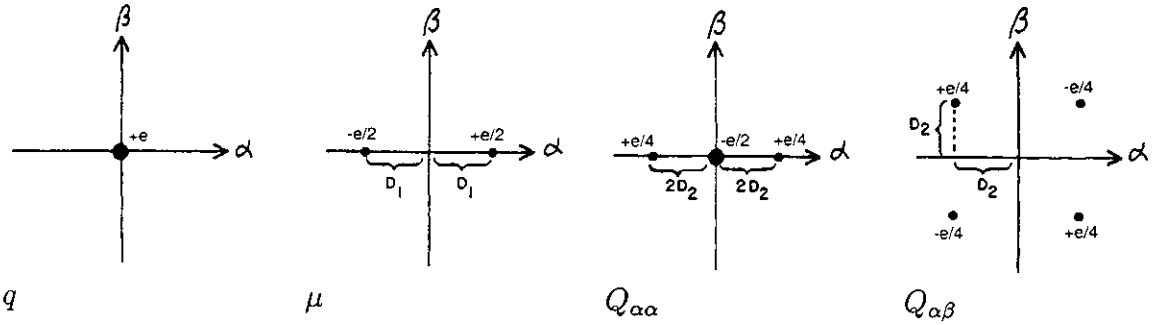
A função semi-empírica $F_1(R_{ij})$ é definida pela Equação 1.65, onde os termos aditivos $\rho_l$ são determinados de forma numérica para monopolos ($l = 0$), dipolos ($l = 1$) ou quadrupolos ($l = 2$) de maneira que as Equações 1.63 tenham o valor correto do limite de um centro para a interação entre dois monopolos ($g_{ss}$), dois dipolos ($h_{sp}$) ou dois quadrupolos ($h_{pp}$).

$$f_1(R_{ij}) = [{R_{ij}}^2 + (\rho_{l_1}^A + \rho_{l_2}^B)^2]^{-1/2} \tag{1.65}$$

Combinando as equações para encontramos as expressões para calcular as integrais, e determinar o que não é calculado. Despreza-se os orbitais que estão em centros diferentes, e utiliza valores experimentais para estimar as integrais do tipo $(\mu\mu|\nu\nu)$ e $(\mu\nu|\mu\nu)$, como mostrado nas Equações 1.66 onde '$I$' é o potencial de ionização e '$A$' é afinidade eletrônica.

$$\begin{aligned} (2s^A|2s^B) &= 0 \\ (2s^A|2p_z^A) &\neq 0 \\ (s^A s^A|s^A s^A) &= (I - A) \quad \Rightarrow 1 \text{ centro} \end{aligned} \tag{1.66}$$

No caso das integrais de dois centros as contribuições são parametrizadas em termos de potenciais eletrostáticos, conforme os esquemas da Figura 1.3, nas Equações 1.62 para chegar às Equações 1.67. Estas equações mostram dois casos possíveis, para as quais as simplificações adotadas resultam em dois operadores diferentes para encontrar o valor denominado $I1$.




$q$ $\mu$ $Q_{\alpha\alpha}$ $Q_{\alpha\beta}$

Figura 1.3: Representação do esquema de utilização dos momenta na parametrização do MNDO. O monopolo $q$ representa a distribuição de cargas entre as funções tipo $ss$ e $p_\alpha p_\alpha$. O dipolo $\mu_\alpha$ representa as distribuições $sp_\alpha$. O quadrupolo $Q_{\alpha\alpha}$ representa os quadrupolos lineares das distribuições $p_\alpha p_\alpha$. O quadrupolo $Q_{\alpha\beta}$ representa as distribuições dos quadrupolos quadrados $p_\alpha p_\beta$. Os índices $\alpha$ e $\beta$ correspondem aos eixos $x$, $y$ ou $z$.

Nas Equações 1.67 $V_{ee}$ é a repulsão elétron-elétron, a repulsão núcleo-núcleo, $V_{NN}$, é parametrizada utilizando um valor experimental, e $\beta_{\mu\nu}$ é parametrizado usando o valor das integrais de recobrimento $S_{\mu\nu}$ multiplicado por uma constante. O valor de $\beta$ é diferente para cada tipo de átomo.

A expressão utilizada para $V_{NN}^{MNDO}$ é mostrada na Equação 1.68. No caso do MNDO, Dewar parametrizou o sistema para reproduzir $\Delta H^o_{f,298}$ em fase gasosa. São utilizados cerca de doze parâmetros diferentes em $\alpha$. O método não MNDO tem boa representação para os valores de pontes de hidrogênio.

$$
\begin{aligned}
F_{\mu\nu} &= H_{\mu\nu}^{core} + V_{ee} \\
H_{\mu\nu}^{core} &= \langle \phi_\mu | -\tfrac{1}{2}\nabla^2 - \textstyle\sum_B \frac{-Z_B}{r_b} | \phi_\nu \rangle \\
H_{\mu\nu}^{core} &= \overbrace{\langle \psi_\mu | \underbrace{-\frac{1}{2}\nabla^2 - \frac{Z_A}{r_a}}_{\hat{h}_a} | \psi_\nu \rangle}^{I1} + \langle \psi_\mu | -\tfrac{1}{2}\nabla^2 - \tfrac{Z_B}{r_b} | \psi_\nu \rangle \\
\text{Caso 1} \Rightarrow \quad \psi_\mu^A &= \psi_\nu^A \\
I1 &= V_{NN} \\
\text{Caso 2} \Rightarrow \quad \psi_\mu^A &= \psi_\nu^B \\
I1 &= \langle \psi_\mu^A | \hat{h}_a | \psi_\nu^B \rangle \equiv \beta_{\mu\nu} \\
\beta_{\mu\nu} &= S_{\mu\nu}(\beta_\mu^A + \beta_\nu^B) \\
S_{\mu\nu} &= \textstyle\int_0^\infty \psi_\mu^* \psi_\nu d\tau
\end{aligned}
\tag{1.67}
$$

$$
V_{NN}^{MNDO} = Z_A Z_B (s_A s_A, s_B s_B)[1 + \exp(-\alpha_A R_{AB}) + (-\alpha_B R_{AB})] \tag{1.68}
$$

**Austin Model (AM1)**

O AM1 pertence à mesma família do MNDO. De fato, são quase idênticos, exceto pela modificação da Equação 1.68 que tem a forma mostrada nas Equações 1.69. Há um acréscimo de termos de repulsão nuclear, pelos parâmetros $K, L, M$. A expressão para $F(B)$ tem a forma similar à de $F(A)$, mostrada na Equação 1.69 [60], onde $n_a$ é um valor que depende do tipo de átomo utilizado, valendo três para carbonos, por exemplo.

$$
\begin{aligned}
V_{NN}^{MNDO} &= Z_A Z_B (s_A s_A, s_B s_B)[1 + F(A) + F(B)] \\
F(A) &= \exp(-\alpha R_{AB}) + \sum_{i=1}^{n_a} K_i^A \exp[L_i^A (R_{AB} - M_i^A)^2]
\end{aligned}
\tag{1.69}
$$

## 1.6 A análise conformacional na busca da conformação bioativa

A maioria dos métodos de busca conformacional partem da geração de uma geometria inicial, que é otimizada com um método rápido, como mecânica molecular. O confôrmero resultante é comparado com outros confôrmeros já encontrados para testar se trata-se de

uma geometria duplicada, e se é considerado um novo confôrmero, é adicionado à lista de confôrmeros únicos. O ciclo repete-se até que todas as geometrias iniciais tiverem sido otimizadas e testadas, ou até que novos confôrmeros não sejam mais encontrados.

Os métodos utilizados para gerar as geometrias iniciais podem ser divididos em duas grandes categorias: (i) Busca determinística que cobre todo o espaço conformacional sistematicamente ou (ii) busca estocástica utiliza elementos aleatórios para explorar o espaço conformacional [61].

### 1.6.1 Busca sistemática de geometria

Geralmente o que diferencia confôrmeros não são comprimentos ou ângulos de ligação, e sim os ângulos de diedro formados entre duas tríades de átomos. Isto deve-se á grande energia necessária para distorcer ângulos e comprimentos de ligação, se comparada à necessária para variar um ângulo de diedro.

A busca sistemática pode garantir a exploração de todo o espaço conformacional, porém resulta em um número excessivamente elevado de geometrias iniciais, igual a $(360/s)^n$ para passos de '$s$' graus em uma molécula com '$n$' diedros. Utilizar passos maiores que 60° em cada rotação de ligação pode não permitir uma resolução adequada na varredura do espaço conformacional [62], capaz de gerar geometrias iniciais na proximidade de todos os mínimos locais com a perda de conformações.

### 1.6.2 Métodos de busca aleatória de geometria

#### A busca estocástica de geometria

A busca estocástica, ou de Monte Carlo, não percorre todo o espaço conformacional. Limita-se a variar a geometria de confôrmeros estáveis em coordenadas internas ou externas em pequenos passos, e comparar a geometria minimizada com o banco de dados: Se a energia estiver dentro de uma faixa ou for menor que os valores presentes no banco é aceita. A possibilidade de encontrar novas conformações diminui com o aumento do banco, e o processo cessa quando não se encontra novas conformações partindo dos mínimos já encontrados em um determinado número de tentativas.

Uma implementação diferente é dada pelo algorítimo de Metrópolis, onde a probabilidade $P_T^a$ de uma nova conformação ser aceita é dada pela expressão da Equação 1.70, onde '$T$' é a temperatura absoluta, '$\Delta E_i$' é o valor de energia calculada para a conformação e '$k$' é a constante de Boltzman.

$$P_T^a = \exp\left(-\frac{\Delta E_i}{kT}\right) \tag{1.70}$$

O valor de $P_T^a$ obtido é comparado com um número aleatório entre [0;1], e se for maior a conformação é aceita. Se for menor uma nova conformação é gerada. Este algorítimo é considerado ineficiente para busca conformacional em proteínas devido a má probabilidade de aceitação de configurações causada pela anisotropia da superfície de energia potencial em um sistema com muitas ligações covalentes [62].

### Busca de geometria por dinâmica molecular (MD)

Utilizando Dinâmica Molecular podemos realizar busca conformacional. Apesar de ser efetiva na busca, requer um tempo maior para criar geometrias significativamente diferentes e é lento para atravessar barreiras rotacionais altas. Estes problemas podem ser contornados com a utilização de temperaturas altas na simulação, e passos maiores de tempo [61].

### Busca de geometria com o método *Low Mode* (LM)

O método LM segue os autovetores de uma análise dos modos normais de vibração de um dos mínimos de energia, através da qual as interconexões entre os mínimos em uma superfície de energia potencial são traçadas aleatoriamente. A conformação inicial é um mínimo local em que um cálculo dos modos normais de vibração é utilizado para identificar quais as direções de movimento que conduzem às energias menores. A geometria inicial é então perturbada em passos fixos pelo caminho indicado por um dos vetores de vibração normal escolhido aleatoriamente. A minimização da estrutura obtida dá origem à outra geometria, que é associada à geometria inicial pelo ponto de sela. O processo prossegue de forma análoga para gerar mais mínimos locais. A eficiência do método reside no fato que um grau de liberdade é levado ao seu ponto de máximo, enquanto os demais são mantidos em seus mínimos [63].

### Busca de geometria com matriz de distâncias métricas

O método utilizado para gerar as estruturas iniciais na busca de geometria foi o da Matriz de Distâncias Métricas [64]. Um matriz de distâncias $D_{ij} \Rightarrow i \neq j \ ; \ i < n \ ; \ j < n$ entre os $n$ átomos da molécula é construída da seguinte forma: A parte triangular superior $(i < j)$ contendo valores das distâncias máximas entre o par de átomos $i$ e $j$. A parte triangular inferior $(i > j)$ contém as distâncias mínimas permitidas para cada par de átomos $i, j$. Uma representação tridimensional da molécula é construída de forma que satisfaça as restrições da matriz $D$ e as restrições de configuração obtidas em uma tabela de conectividade convenientemente transformada. Esta tabela de conectividade contém informação sobre como os átomos estão ligados e sobre sua conformação estereoquímica.

O preenchimento da matriz $D$ é realizado em diversos passos, enumerados a seguir:

1. As distâncias na matriz $D$ são obtidas dos comprimentos de ligação padrão, tabelados, para pares de átomos ligados $i, j$;

2. A distância entre um par de átomos $i, k$, ambos ligados a um terceiro átomo $j$ é definida pela Equação 1.71, onde $\theta$ é o ângulo formado pelos três átomos;

3. Anéis de três e quatro membros são identificados através do algorítimo de Welsh–Gibbs–Assembly, descrito por Dyott [65];

4. As distâncias mínima e máxima para átomos $i$ e $l$, formando um diedro, são computadas utilizando os ângulos de ligação relevantes e as distâncias previamente computadas;

5. As distâncias máxima ($u_{ij}$) e mínima ($l_{ij}$) entre cada par de átomos $i, j$ são computadas:

   - Para átomos considerados nos ítens (1) e (2) as distâncias $l_{ij} = u_{ij} = d_{ij}$;
   - Para os átomos considerados no ítem (4) em que há ligação dupla entre os átomos $j$ e $k$ os valores $l_{ij}$ e $u_{ij}$ podem ser definidos precisamente utilizando os ângulos e comprimentos de ligação tabelados, além da configuração *cis/trans*;
   - Para ligações simples e triplas no ítem (4) os valores mínimo e máximo das distâncias são obtidos com o ângulo $\varphi$ igual à 0° e 180° nas Equações 1.72. Os valores de $\theta_1$ e $\theta_2$ são os ângulos entre os átomos $i, j, k$ e $j, k, l$, e os valores de $k_{ij}$ e $k_{kl}$ são os comprimentos de ligação entre os átomos $i, j$ e $k, l$.

6. Na falta de outro valor padrão, os valores mínimos de distância entre átomos não ligados são estabelecidos em 2 Å e os valores máximos em $10n^{1/3}$, onde $n$ é o número de átomos da molécula;

7. O último passo é o refinamento da matriz de distâncias pela aplicação exaustiva e iterativa da desigualdade triangular até que não seja possível contrair mais o intervalo $(l_{ij}, u_{ij})$

$$d_{ik} = \sqrt{d_{ij}^2 + d_{jk}^2 - 2d_{ij}d_{jk}\cos\theta} \tag{1.71}$$

$$\begin{aligned} d_{il} &= \sqrt{p_1 - p_2\cos\varphi} \\ p_1 &= d_{ij}^2 + d_{jk}^2 + d_{kl}^2 - 2d_{ij}d_{jk}\cos\theta_1 \\ &- 2k_{jk}d_{jl}\cos\theta_2 + 2d_{ij}d_{kl}\cos\theta_1\cos\theta_2 \\ p_2 &= 2d_{ij}k_{kl}\,\mathrm{sen}\,\theta_1\,\mathrm{sen}\,\theta_2 \end{aligned} \tag{1.72}$$

Obtida a matriz de distâncias devemos computar as coordenadas iniciais do conjunto de átomos. Dada a matriz $D$ com as distâncias, podemos calcular as coordenadas a partir do centro de massa $O$ calculado com todas as massas atômicas unitárias utilizando a Equação 1.73. Uma nova matriz de distâncias $G$ é definida a partir dos centros de massa, onde cada elemento $g_{ij}$ é dado pela Equação 1.74. Esta matriz é simétrica por definição, e portanto é possível resolvê-la em seus autovalores e autovetores. Um conjunto de coordenadas iniciais para a molécula pode ser obtido a partir dos três primeiros autovetores com maior autovalor, [66]. Este passo corresponde à projeção da molécula no subespaço dos três primeiros autovetores, o que freqüentemente leva a violações das restrições de quiralidade e distância impostas.

$$d_{iO}^2 = n^{-1}\sum_{j=1}^{n} d_{ij}^2 - n^{-2}\sum_{j=2}^{n}\sum_{k=1}^{j-1} d_{jk}^2 \tag{1.73}$$

$$g_{ij} = 1/2(d^2{}_{iO} + d^2{}_{jO} - d^2{}_{ij}) \tag{1.74}$$

Uma maneira conveniente de obter novamente as coordenadas com as restrições corretamente consideradas é o método do gradiente conjugado proposto por Fletcher e Reeves,

que consiste em minimizar o erro $E_{err}$ nas Equações 1.75. O primeiro termo, $F$, é função somente das distâncias, com as somatórias realizadas apenas dentro dos intervalos definidos por $(l_{ij}, u_{ij})$. O segundo termo, $C$, é usado para impor as restrições de quiralidade. Corresponde à soma dos quadrados das diferenças entre os paralelogramos definidos pelos vetores $\vec{r_{ia}}$, $\vec{r_{ja}}$ e $\vec{r_{ka}}$ entre o átomo quiral $a$ três dos seus vizinhos, $i$, $j$ e $k$ e o volume orientado $\hat{v_a}$ quando a quiralidade e os ângulos de ligação estão corretos. A somatória é realizada sobre todos os átomos assimétricos.

$$\begin{aligned} E_{err} &= F + C \\ F &= \sum_{i<j}(d^2{}_{ij} - u^2{}_{ij})^2 + \sum_{i<j}(d^2{}_{ij} - l^2{}_{ij})^2 \\ C &= \sum_a(\vec{r_{ia}} \bullet (\vec{r_{ja}} \times \vec{r_{ka}} - \hat{v_a})^2) \end{aligned} \tag{1.75}$$

A implementação do algorítimo no programa `distgeom`, que é parte do pacote `TINKER`, segue os mesmos conceitos aqui apresentados, com pequenas modificações. O algorítimo proposto por Ponder (não publicado) utiliza uma distribuição Gaussiana que é otimizada iterativamente durante para a seleção da geometria inicial durante a metrização. A otimização da distribuição é feita em termos das médias e desvios padrão comparando propriedades da estrutura 'de trabalho' com as estimadas diretamente da matriz de distâncias. Outra modificação é que o algorítimo usa pares de átomos escolhidos da matriz de distâncias ao invés de escolher na molécula, o que aumenta o grau de aleatorização e uma melhor busca no espaço conformacional para um mesmo tempo de computação [67].

Várias implementações foram publicadas para este procedimento de geração do geometrias moleculares, e mais recentemente uma denominada *Geometrical Algorithm to Search the Conformational Space* (GASCOS) [68–71], que tem seu formalismo matemático brilhantemente descrito na serie de artigos em que é apresentada, e que é considerada pelos autores mais eficiente que as demais publicadas anteriormente.

## 1.7 Métodos estatísticos utilizados para obter modelos SAR e QSAR

Associado aos métodos para obtenção das propriedades moleculares utilizados, existe a necessidade de aplicação de técnicas matemáticas para estabelecer correlações robustas entre o conjunto de centenas de descritores e as atividades biológicas de algumas dezenas de compostos.

Diversas metodologias têm sido utilizadas, desde regressão multivariada simples até aplicação de redes neurais bastante complexas. A regressão multivariada é a base da análise de Hansh, a mais tradicional e aceita metodologia para estabelecimento de modelos de QSAR. Uma vantagem deste tipo de análise, além da sua simplicidade, é fato de permitir uma fácil identificação da influência de cada descritor no modelo obtido.

Se o conjunto descritor é muito extenso, outros métodos estatísticos são mais adequados. Os métodos que envolvem uso projeção (*Partial Least Squares* - PLS, *Principal Component Regression* - PCR) são freqüentemente utilizadas em conjuntos de descritores mecânico-quânticos [72, 73]. Têm como desvantagem o fato de dificultar uma análise

detalhada da influência de cada descritor no modelo obtido. A acuracidade do modelo é considerada em termos da capacidade de previsão obtida, para compostos fora do conjunto de treinamento. Os métodos que envolvem um tratamento estatístico mais sofisticado do conjunto de dados como PLS, PCA, *Back-Propagation Neural Networks* (BPN) têm muitos pontos comuns, podendo ser reduzidos a uma formulação geral muito semelhante quando considerados em seus aspectos matemáticos fundamentais [74].

Técnicas de classificação (PCA, SIMCA, BPN) e reconhecimento de padrões, podem ser usadas para classificar grupos de compostos a partir dos dados mecânico-quânticos correspondentes, criando modelos de Relação entre Estrutura e Atividade (SAR). Modelos de classificação também podem ser obtidos com o método PLS, utilizando uma técnica denominada Análise Discriminante PLS. A classificação entre grupos de substâncias muito ativas ou pouco ativas (ou inativas) pode ser suficiente para selecionar possíveis candidatos ao posto de *lead* em um processo de *screening* farmacológico.

Os conjuntos descritores capazes de modelar adequadamente valores de atividade (ou classificar grupos de alta e baixa atividade), e de prever com acerto os valores (ou classes) de atividade, podem também informar sobre as regiões da molécula responsáveis pelas interações que ativam o mecanismo de resposta biológica. A interpretação do significado químico dos dados mecânico-quânticos, pode fornecer informações sobre modificações na estrutura de um composto com potencial para aumentar sua resposta biológica.

A atribuição de regiões putativas de ligação de uma classe de drogas ao seu alvo biológico, baseada na análise das variáveis descritoras relacionadas com regiões da molécula, pode ser utilizada como aproximação da estrutura do farmacóforo. Sendo assim, as informações obtidas na análise multivariada das propriedades calculadas para um grupo de compostos, podem ajudar a prever a identidade do seu alvo biológico numa macromolécula, baseando-se na complementaridade entre substrato e enzima [75].

A elucidação completa do mecanismo de ação é um grande passo no caminho de um processo de desenho racional da droga. A localização exata das interações com a macromolécula, a natureza de cada interação e sua importância relativa para a ativação da resposta biológica, são um conhecimento essencial para tanto. Qualquer informação pode ser de valia, nos casos em que não se tem informações mais detalhadas sobre o mecanismo de ação biológica. Este freqüentemente é o caso dos estudos SAR ou QSAR.

### 1.7.1 Natureza das transformações nos dados de entrada em modelagem empírica

O modelo determinado por qualquer método de modelagem empírico pode ser representado como uma soma ponderada de funções de base como mostrado na Equação 1.76, onde $y_k$ é a k-ésima variável de resposta prevista, $\mu_n$ é a n-ésima base ou função de ativação, $b_{mk}$ é o peso final ou coeficiente de regressão relacionada à n-ésima função de base e k-ésimo valor de saída, $a$ é a matriz de parâmetros da função de base, $f_n$ representa uma transformação nos dados de entrada e $x_1, \ldots, x_j$ são os dados de entrada ou variáveis independentes. A variável obtida pela transformação nos dados de entrada $z_n = f_n(a, \mathbf{X})$ é comumente chamada de variável latente ou projeção dos dados de entrada.

Métodos específicos de modelagem empírica podem ser derivados da Equação 1.76

dependendo das decisões sobre a natureza da transformação dos dados de entrada, tipo de função de ativação ou função de base e critério de otimização. Na Tabela 1.2 são mostrados exemplos de combinações freqüentemente aplicadas em cada método específico.

$$y_k = \sum_{m=1}^{M} b_{mk}\mu_n(f_n(a; x_1, x_2, \ldots, x_j) \tag{1.76}$$

### Funções de transformação aplicadas na construção de modelos

A redução da dimensionalidade do espaço dos dados de entrada é essencial para um aumento da complexidade de modelagem. Técnicas de modelagem empírica podem capturar a relação entre dados de entrada com um número menor de variáveis latentes que o número de dados de entrada. Esta redução na dimensionalidade é geralmente acompanhada pela exploração das inter-relações dos dados de entrada, distribuição dos dados de treinamento no espaço dos dados de entrada ou relevância de variáveis de entrada para previsão dos dados de saída. Portanto podemos agrupar em três classes principais as técnicas de transformação de dados de entrada [74].

- **Métodos baseados em projeção linear.** Exploram as relações lineares entre valores de entrada projetando-os num hiperplano linear, como mostrado na Figura 1.4(a); antes de aplicar a função de base. Portanto os valores de entrada são transformados por combinação de um somatória linear ponderada para formar variáveis latentes.

- **Métodos baseados em projeção não linear.** Exploram as relações não lineares entre dados de entrada projetando-os numa hipersuperfície não linear, resultando variáveis latentes que são funções não lineares como mostrados nas Figuras 1.4(b) e (c). Se os dados de entrada são projetados em uma hipersuperfície localizada então as funções de base são locais como na Figura 1.4(b). Se as funções de base são de natureza não local a hipersuperfície de projeção não tem representação localizada como na Figura 1.4(c).

- **Métodos baseados em partição.** Reduzem a dimensionalidade do sistema selecionando as variáveis mais importantes para modelagem. O espaço de entrada é dividido em hiperplanos perpendiculares pelo menos à um dos eixos de entrada, como mostrado na Figura 1.4(d).

### Tipos de função de ativação

A função que relaciona a variável latente (z) com o valor de saída é naturalmente bidimensional, sendo chamada função de ativação. A função que relaciona os valores de entrada com os de saída, (a;**X**), é chamada função de base. É grande a variedade de funções de ativação usadas em métodos de modelagem empírica.

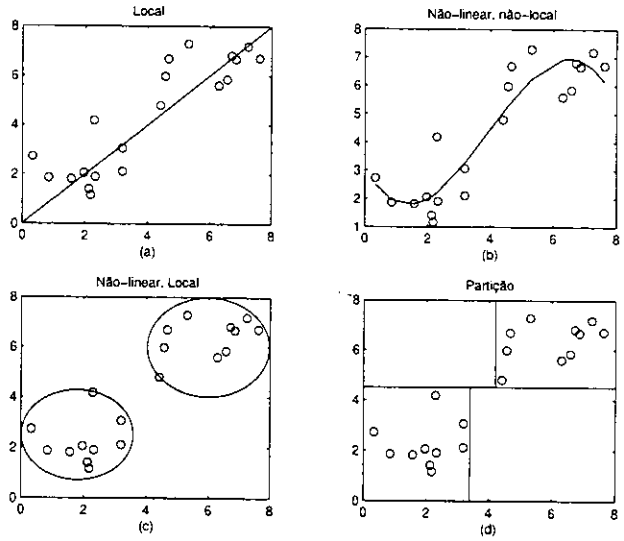



Figura 1.4: Transformações freqüentemente aplicadas em conjuntos de dados de entrada: Método baseado em projeção linear (a), projeção não-linear não-local (b), projeção não-linear local (c) e projeão baseada em partição (d).

- **Funções de ativação de formato fixo.** Podem ser lineares, sigmóides, gaussianas, *wavelet* ou sinusoidais, entre outras. Ajustando os parâmetros da função pode-se mudar seu tamanho e localização, porém não sua forma.

- **Funções de ativação de formato adaptativo.** São usadas por métodos que relaxam a condição de forma fixa da função de ativação. Este grau adicional de liberdade dá grande flexibilidade na determinação da superfície desconhecida de entrada e saída, e freqüentemente resulta em modelos mais compactos. Isso é obtido pela aplicação de técnicas de ajuste com *splines* e polinomiais para aproximar o espaço transformado de entrada e saída.

### Critério de otimização

O objetivo de qualquer método de modelagem empírica é extrair as relações subliminares entre entrada e saída a partir dos dados disponíveis, transformando os dados de entrada se necessário.

A transformação é determinada pela função $(f)$ e seus parâmetros $(a)$, enquanto a relação entre entrada transformada e saída é determinada pelos parâmetros $(b)$ e funções de base $(\mu)$.

A maioria dos métodos de modelagem empírica minimiza a média dos quadrados do erro de aproximação para determinar o conjunto de base $(\mu)$ e os coeficientes $(b)$. O critério para determinar os parâmetros de transformação da entrada, $(f)$ e $(a)$, difere para cada método dependendo da ênfase no aproveitamento da variância dos dados de entrada, ou na sua relação (covariância) com os dados de saída.

Tabela 1.2: Tabela de comparação para alguns métodos de modelagem empírica baseada nos tipos de funções de entrada, de base e nos critérios de otimização adotados freqüentemente para cada método.

| Método[b] | Transformação de entrada | Função de base | Critério de otimização |
|---|---|---|---|
| PCA | Projeção linear | Linear de formato fixo | i–Min. erro de aproximação dos dados de entrada |
| PLS | Projeção linear | Linear de formato fixo | i–Max. covariância entre projeção de entrada saída<br>ii–Min. erro de aproximação |
| PCR | Projeção linear | Linear de formato fixo | i–Max. variância das projeções de entrada<br>ii–Min. erro de aproximação |
| SIMCA | Projeção linear e não local | Linear de formato fixo | i–Min. erro de aproximação dos dados de entrada |
| BPN | Projeção linear | Sigmóide de formato fixo | i–Min. erro de aproximação |
| BPNM | Projeção não linear e não local | Sigmóide de formato fixo | i–Min. erro de aproximação |
| KUNN | Projeção linear e não local | Sigmóide de formato fixo | i–Proximidade entre pesos e entradas<br>ii–Número de ciclos |
| NLPCA | Projeção não linear e não local | Formato adaptativo | i–Min. erro de aproximação dos dados de entrada |
| RBFN | Projeção não linear e local | Radial de formato fixo | i–Min. distância entre valores de entrada e centro do bloco<br>ii–Min. erro de aproximação |
| CART | Partição dos dados de entrada | Formato adaptativo de modo constante | i–Min. erro de aproximação |
| MARS | Partição dos dados | Splines (adaptativo) | i–Min. erro de aproximação |

(*b*)—PCA–Análise por Componentes Principais; PLS–Mínimos Quadrados Parciais; PCR–Regressão por Componentes Principais; SIMCA–Soft Independent Modelling of Class Analogy; BPN–Rede Neural de retropropagação com uma camada oculta; BPNM–Rede neural de retropropagação com múltiplas camadas ocultas; KUNN–Rede neural de treinamento não supervisionado com arquitetura tipo Kohonen; NLPCA–Análise por Componentes principais Não Linear; RBFN–Rede de Funções de Base Radiais; CART–Árvores de Classificação e Regressão; MARS–Regressão Adaptativa Multivariada por Splines.

### 1.7.2 Análise por componentes principais

A análise de componentes principais (*Principal Component Analysis*, PCA) é a principal forma de compressão e extração de informação de dados multivariados. Matematicamente, a PCA consiste em uma decomposição em autovalores da matriz de covariância (ou correlação) das variáveis de independentes. Para uma dada matriz $\mathbf{X}$ de dados com $m$ linhas e $n$ colunas, em que cada coluna constitui uma variável e cada linha representa uma amostra, a matriz de covariância de $\mathbf{X}$ é definida pela Equação 1.77, desde que as colunas de $\mathbf{X}$ estejam centradas na média.[1]

$$\text{cov}(\mathbf{X}) = \frac{\mathbf{X}^T\mathbf{X}}{m-1} \tag{1.77}$$

A PCA decompõe a matriz de dados em uma soma de produtos vetoriais de vetores $\mathbf{t}_i$ e $\mathbf{p}_i$, mais uma matriz de resíduos $\mathbf{E}$.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} \tag{1.78}$$

O índice $k$, na Equação 1.78, deve ser menor ou igual à menor dimensão de $\mathbf{X}$. Os vetores $\mathbf{t}_i$ são denominados *scores* e contém informação sobre como as amostras relacionam-se umas às outras. Na Equação 1.79 os vetores $\mathbf{p}_i$ são os autovetores da matriz de covariância. Para cada $\mathbf{p}_i$, $\lambda_i$ é o autovalor associado ao autovetor $\mathbf{p}_i$.

$$\text{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_i \tag{1.79}$$

Na PCA os vetores $\mathbf{p}_i$ são conhecidos como *loadings* e contêm a informação de como as variáveis relacionam-se umas às outras. Os vetores $\mathbf{t}_i$ formam um conjunto ortogonal ($\mathbf{t}_i^T\mathbf{t}_j = \delta_{ij} \parallel \mathbf{t}_i^2 \parallel$), onde $\delta_{ij}$ é o Delta de Kronecker[2] e os vetores $\mathbf{p}_i$ formam um conjunto ortonormal ($\mathbf{p}_i^T\mathbf{p}_j = 0$ para $i \neq j$ e $\mathbf{p}_i^T\mathbf{p}_j = 1$ para $i = j$). Na Equação 1.80 a matriz $\mathbf{X}$ é decomposta no produto das matrizes $\mathbf{T}$ (matriz de *scores*) e $\mathbf{P}$ (matriz de *loadings*) onde $\mathbf{TP}^T = \sum_{i=1}^{k} \mathbf{t}_i\mathbf{p}_i^T$.

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \tag{1.80}$$

Para qualquer par $(\mathbf{p}_i; \mathbf{t}_i)$ o vetor de *score* $\mathbf{t}_i$ é uma combinação linear de $\mathbf{X}$ definida por $\mathbf{p}_i$. Geometricamente, verifica-se que $\mathbf{t}_i$ são as projeções de $\mathbf{X}$ em $\mathbf{p}_i$ conforme a Equação 1.81.

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i \tag{1.81}$$

Uma característica da decomposição em componentes principais é que os pares $(\mathbf{p}_i; \mathbf{t}_i)$ são arranjados em ordem decrescente de acordo com o autovalor $\lambda_i$ associado, que é uma medida da quantidade de variância, ou informação, descrita por cada par. Cada par $(\mathbf{p}_1; \mathbf{t}_1)$ captura a maior quantidade de informação possível dos dados que se encontre na direção de maior variância, e cada par subseqüente captura a maior quantidade possível da informação restante após subtrair $\mathbf{t}_i\mathbf{p}_i^T$ de $\mathbf{X}$.

[1] Caso as colunas de $\mathbf{X}$ estejam auto-escaladas, a Equação 1.77 calcula a matriz de correlação de $\mathbf{X}$.

[2] $\delta_{ij}$: Delta de Kronecker. $\delta_{ij} = \begin{cases} 1 & \text{para } i = j \\ 0 & \text{para } i \neq j \end{cases}$

Assim, a PCA agrupa as amostras que estão altamente correlacionadas em novas variáveis, ou fatores, formados por combinações lineares das variáveis originais, e dispostos nas direções de maior variância do conjunto de dados, como é mostrado na Figura 1.5. Os fatores são formados pelos pares $(\mathbf{p}_i; \mathbf{t}_i)$ e denominam-se Componentes Principais (PC). É importante salientar que a representação em outra base não afeta as relações entre as variáveis originais [76].

Graficamente, a PCA pode ser vista como uma rotação nos eixos de coordenadas originais, de forma a produzir novos eixos, que passem através das direções de maior variância, sempre ortogonais entre si. A Figura 1.5 apresenta a primeira componente principal na direção da maior variância das três variáveis originais; a segunda componente principal seria obrigatoriamente ortogonal a esta, passando pela segunda direção de maior variância; e assim sucessivamente, até o número máximo $k$ de fatores. O ponto $\mathrm{T}_i$ representa o valor (*score*) da amostra $\mathrm{X}_i$ no novo eixo de coordenadas $\mathrm{P}_1$. Matematicamente, exis-

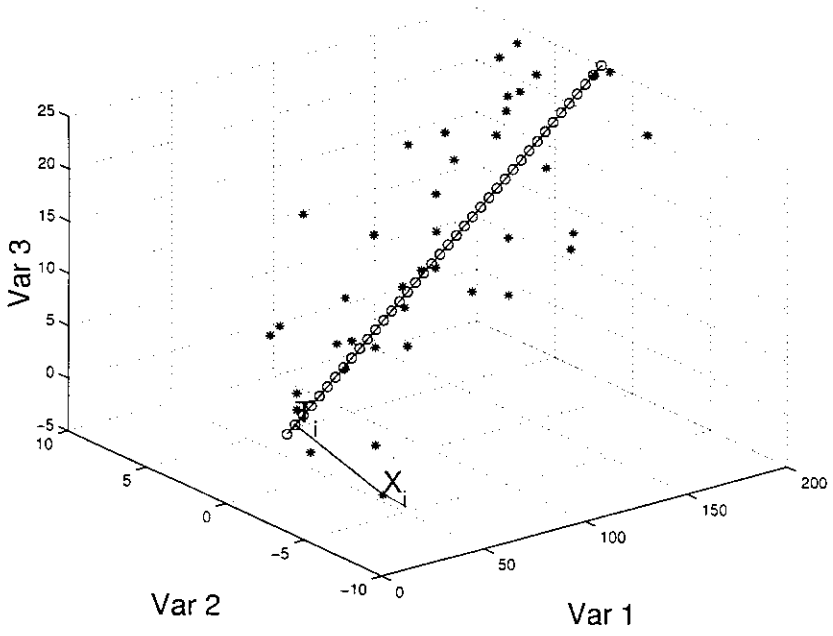



Figura 1.5: Descrição gráfica da PCA mostrando esquematicamente como a nova coordenada do ponto $X_i$ é projetada na PC obtida, no ponto $T_i$. Neste caso o sistema de três coordenadas cartesianas foi comprimido para apenas uma coordenada, a componente principal PC1.

tem várias formas de se obter as matrizes de *scores* e *loadings*, dentre as quais uma das mais empregadas é o método de Decomposição em Valor Singular [77] (*Singular Value Decomposition*, SVD), que consiste em uma eliminação LU de Gauss seguida por uma ortogonalização de Gram-Schmidt. Este método é eficiente para encontrar a solução e apresenta boa estabilidade numérica [78].

Uma vez que as primeiras componentes principais descrevem a maior parte da variância dos dados, é possível, com um número de fatores sensivelmente menor do que o número de variáveis originais, concentrar elevada fração da informação contida nos dados.

Como conseqüência, a PCA permite identificar e dar ênfase à informação relevante, bem como visualizar os dados em duas ou três dimensões, tornando-se uma ferramenta exploratória muito útil. Gráficos de *scores* permitem a identificação de amostras, verificando se são semelhantes ou não, típicas ou *outliers*[3], e se apresentam agrupamentos. Gráficos de *loadings* permitem a identificação das variáveis importantes, seleção e classificação dos objetos, entre outras melhorias proporcionadas à visualização dos dados [79–81].

Outra característica da PCA é que e as últimas componentes, em geral, não descrevem mais do que ruído, formado pela parcela com menor variância e sem correlação com outras variáveis. A Equação 1.80 pode ser reordenada resultando a Equação 1.82, onde $\overline{\mathbf{X}} = \mathbf{T}\mathbf{P}^T = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_k\mathbf{p}_k^T$ é a matriz $\mathbf{X}$ aproximada.

$$\mathbf{X} = \overline{\mathbf{X}} + \mathbf{E} \tag{1.82}$$

Na Equação 1.82, quando $k = n$, temos $\mathbf{E} = 0$, ou, em outras palavras, toda a informação de $\mathbf{X}$ é contemplada pela PCA, e $\overline{\mathbf{X}} = \mathbf{X}$. Por outro lado, $k$ pode ser truncado em um número menor de fatores suficiente para descrever a maior parte da variância de $\mathbf{X}$, situação em que $\mathbf{E} \neq 0$ e $\overline{\mathbf{X}} \neq \mathbf{X}$. Neste caso, a matriz aproximada não conterá o ruído presente nas componentes excluídas, que estará presente na matriz de resíduos. Desta forma, o emprego da PCA ao tratamento de dados multivariados permite redução na dimensionalidade dos dados e eliminação de ruído.

### 1.7.3 Regressão multivariada por mínimos quadrados parciais

O processo geral de calibração consiste de duas etapas: a modelagem, que estabelece uma relação matemática entre $X$ e $Y$ no conjunto de calibração, e a validação, que otimiza a relação no sentido de uma melhor descrição das espécies de interesse [79].

O método de regressão multivariada PLS (*Partial Least Squares*, também conhecido como *Projection to Latent Variables*) é um método de regressão que utiliza a decomposição em componentes principais. Dentre as principais características dos modelos produzidos através deste método, destaca-se a robustez: os parâmetros do modelo por ele produzido praticamente não se alteram com a inclusão de novas amostras no conjunto de calibração. Além disso, não é necessário que o número de amostras exceda o número de variáveis, e nem que estas tenham baixa dependência linear entre si, limitações típicas de métodos clássicos de calibração [79].

O modelo de calibração, construído a partir de um conjunto de amostras representativo da população, tem a forma mostrada na Equação 1.83, onde $\mathbf{Y}$ é a matriz (ou vetor) de variáveis dependentes, $\mathbf{X}$ é a matriz das variáveis independentes e $\beta$ é o vetor de regressão, resultado final da aplicação do método, que poderá ser utilizado na previsão de amostras desconhecidas.

$$\mathbf{Y} = \mathbf{X}\beta \tag{1.83}$$

Conforme apresentado na Equação 1.81, o vetor de *score* $\mathbf{t}_i$ é uma combinação linear de $\mathbf{X}$, de modo que é possível representar uma matriz de dados através de sua matriz de *scores*. O método PLS realiza uma decomposição em componentes principais em ambos

[3]Outlier: dado anômalo, que pode, ou não, estar errado.

os conjuntos $\mathbf{X}$ e $\mathbf{Y}$, conforme a Equação 1.84, onde $\mathbf{T}$ e $\mathbf{U}$ são as matrizes de *scores*, $\mathbf{P}$ e $\mathbf{Q}$ são as matrizes de *loadings*, e $\mathbf{E}$ e $\mathbf{F}$ são as matrizes de resíduos de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente.

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}' + \mathbf{E} \\ \mathbf{Y} &= \mathbf{UQ}' + \mathbf{F} \end{aligned} \tag{1.84}$$

Em uma etapa seguinte é realizada uma regressão linear entre as matrizes de *scores* de $\mathbf{X}$ e $\mathbf{Y}$ para os fatores que apresentarem variação em direções colineares, na forma da Equação 1.85, onde o índice $h$ representa cada fator obtido [82].

$$\mathbf{u}_h = b_h \mathbf{t}_h \tag{1.85}$$

Para otimizar essa relação, o método PLS utiliza iterativamente a informação dos blocos de dados $\mathbf{X}$ e $\mathbf{Y}$ durante a decomposição dos fatores, de maneira que ligeiras rotações produzam fatores em direções mais colineares. Assim, os fatores—doravante denominados variáveis latentes (variável latente)—são construídos de forma que a primeira variável latente de $\mathbf{X}$ descreva a direção de máxima variância que também se correlaciona com $\mathbf{Y}$; e as demais sucessivamente nas direções colineares subseqüentes. Por meio do equilíbrio entre as informações de $\mathbf{X}$ e $\mathbf{Y}$, o método reduz o impacto de grandes variações nas variáveis independentes que podem ser irrelevantes para a variável dependente [79, 80].

A relação mista entre $\mathbf{X}$ e $\mathbf{Y}$ tem então a forma $\mathbf{Y} = (\mathbf{TB})\mathbf{Q}' + \mathbf{F}$ onde se deseja minimizar $\mathbf{F}$. Para tanto, existe um número ótimo de fatores a incluir no modelo. Conforme apresentado na Seção 1.7.2, uma das vantagens de se realizar uma análise de componentes principais em uma matriz de dados é a possibilidade de filtrar ruído dos dados originais, pela exclusão de componentes que apresentarem menor quantidade de informação. Porém, um modelo de calibração deve incluir o máximo de informação das variáveis dependentes que seja relevante na predição da variável independente. A determinação do número ideal de fatores a incluir no modelo envolve técnicas de validação.

O modelo de regressão multivariada pode ser expresso em termos dos coeficientes de cada variável descritora em cada uma das variáveis latentes do modelo, definido pela matriz $M$ nas Equações 1.86. O vetor de regressão é $\mathbf{m}$.

$$\begin{aligned} \mathbf{M} &= (w * \mathrm{inv}(p' * w) * \mathrm{diag}(b))' \\ \mathbf{m} &= \mathrm{sum}(\mathbf{M}') \end{aligned} \tag{1.86}$$

Nas Equações 1.86, $b$ é a matriz das relações internas, $w$ são os pesos das variáveis independentes (descritores) e $p$ são os *loadings* do bloco das variáveis independentes, de acordo com o procedimento realizado pelo método NIPALS [83]. As equações consideram que há somente uma variável dependente (atividade), devendo ser modificadas para incluir os *loadings* do bloco das variáveis dependentes no cálculo com mais de uma variável dependente. As operações de álgebra matricial utilizam-se da sintaxe determinada para o programa `octave` [84], e realizam as operações de inversão de matriz (inv), a criação de uma matriz quadrada com o vetor dado como argumento de entrada na diagonal da matriz (diag), a soma das colunas da matriz (sum) e a transposição da matriz (').

### 1.7.4 As redes neurais de retropropagação

O grande interesse despertado pelas Redes Neurais de Retropropagação (BPN–*Back Propagation Neural Networks*) vem em parte da sua habilidade de aproximação universal e facilidade de processamento paralelo. Porém o modelo treinado com BPN é do tipo 'caixa preta' e freqüentemente necessita de uma grande relação entre o tamanho do conjunto de dados de treinamento e o conjunto de variáveis de entrada. A construção da rede é cara do ponto de vista computacional por causa da computação simultânea de todos os parâmetros.

Numa BPN a função de ativação é não linear e de formato fixo, geralmente tipo sigmóide. Esta função encontra as direções de projeção que minimizam o erro de previsão, ignorando qualquer relação interna nos dados de entrada. A representação matemática de um neurônio $y^1$, na primeira camada de uma BPN, pode ser dada por uma função como a da Equação 1.87. Os termos de *bias* das camadas ocultas e de saída, $w_i^1$ e $w_i^2$, são ajustados até a convergência do método. Os parâmetros do modelo numa BPN são computados para minimizar a média dos quadrados do erro de aproximação na saída da rede, conforme a Equação 1.88, onde $Y$ são os valores utilizados no treinamento e $\hat{Y}$ os valores obtidos na saída da rede para um dado conjunto de valores de *bias* $w_i^n$.

$$y_j^l = \sum_{i=1}^{x} w_{ji}^l x_i^l \tag{1.87}$$

$$\sigma^{BPN} = \min_{w}(Y - \hat{Y})^2 \tag{1.88}$$

Métodos para treinamento de BPN têm recebido considerável atenção. O mais comumente utilizado otimiza os pesos dos dados de entrada (direções de projeção) e pesos da saída (coeficientes de regressão) simultaneamente para toda rede pelo algorítimo de retropropagação. Este procedimento pode ser custoso porque o número ótimo de nós é determinado pelo treinamento de várias redes com diferentes números de nós.

Uma forma de simplificar o treinamento de uma BPN é reduzir o número de variáveis na camada de entrada. Pode-se fazer isto comprimindo o conjunto de variáveis de entrada, e utilizando os *scores* da PCA como variáveis de entrada na BPN.

### 1.7.5 Técnicas de validação e otimização de modelos

Um bom modelo de calibração significa um modelo que apresente bom desempenho na previsão de amostras desconhecidas. No método PLS, a inclusão de fatores no modelo, por um lado, leva a um incremento na capacidade preditiva, na medida em que aumenta a quantidade de informação; porém, a inclusão de um elevado número de variáveis latentes leva também à incorporação do ruído dos dados do conjunto de treinamento. Como resultado, pode-se obter um modelo que tenha um ajuste muito bom ao conjunto de treinamento, mas seja ruim na previsão de novas amostras [85].

A validação é uma forma de testar a consistência interna do modelo de regressão enquanto ele está sendo construído, permitindo avaliar sua habilidade preditiva, mediante

o uso de amostras que não participaram da sua elaboração e o emprego de técnicas estatísticas para avaliação do erro envolvido na previsão. Quando se dispõe de um conjunto grande de amostras, algumas amostras são excluídas do conjunto de treinamento para serem usadas apenas na validação. Porém, o que ocorre mais freqüentemente é que o número de amostras é escasso e não se podem desperdiçar amostras que seriam úteis na modelagem. Nesse caso, costumam-se utilizar métodos de validação cruzada, que consistem na remoção de subconjuntos de amostras para formação do conjunto de teste. Diversos subconjuntos são removidos, reconstruindo-se o modelo determinado número de vezes. A forma de escolha dos subconjuntos, a fração de amostras e o número de iterações variam [79, 86].

Dentre as técnicas de validação cruzada, uma das mais simples, e comumente empregada, é a validação do tipo *leave one out* (LO). Esta técnica consiste em retirar-se uma amostra do conjunto, construindo o modelo com as demais $m - 1$ amostras. O erro de predição é então avaliado para a amostra excluída. Esse procedimento é repetido excluindo-se, uma a uma, todas as amostras. Ao final, obtém-se $m$ modelos de $m - 1$ amostras, e avalia-se o erro de previsão através do somatório dos quadrados dos erros de predição de todas as amostras, denominado PRESS (*Prediction Residual Sum of Squares*).

$$PRESS = \sum_{i=1}^{m} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 \tag{1.89}$$

Na Equação 1.89, $\hat{\mathbf{Y}}$ é o valor de $\mathbf{Y}$ conforme predito pelo modelo. O PRESS pode ser avaliado para todos os modelos com um a $k$ fatores, geralmente resultando em uma curva côncava, cujo mínimo determina o número ideal de fatores a considerar [85].

A Figura 1.6 apresenta esquematicamente a curva do PRESS em função do número de variáveis latentes consideradas no modelo, e mostra a contribuição de dois tipos de erro ao erro total de previsão: o erro sistemático, devido a interferência não modelada nos dados, e o erro de estimativa, causado por ruídos aleatórios de medição de vários tipos. O erro sistemático diminui à medida em que se adiciona informação ao modelo, enquanto o erro de estimativa aumenta, pela adição de ruído do conjunto de treinamento. O erro total de previsão corresponde à soma das duas contribuições. É importante salientar que os modelos que consideram todos ou quase todos os fatores têm ótimo ajuste ao conjunto de treinamento, porém o que se deseja em um modelo de calibração é habilidade preditiva [85].

O erro padrão das previsões (*root mean standard error of prediction*, RMSEP) pode ser calculado a partir do PRESS e do número de amostras, $m$, pela seguinte equação:

$$RMSEP = \sqrt{\frac{PRESS}{m}} \tag{1.90}$$

Para um conjunto de dados auto-escalado, a magnitude do erro padrão pode ser julgada comparando-a com o desvio padrão das amostras, que neste caso é unitário.

Outro critério de avaliação introduzido em modelos PLS, diretamente derivado do RMSEP, denominado SDEP também leva em conta o número de variáveis latentes, $lv$,

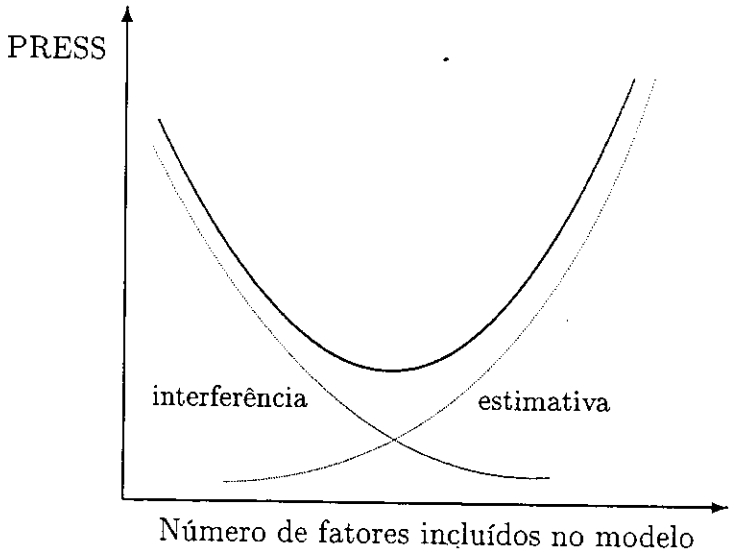



Figura 1.6: Erro de predição como função da complexidade do modelo de calibração.

do modelo conforme a Equação 1.91.

$$SDEP = \sqrt{\frac{PRESS}{m - lv - 1}} \tag{1.91}$$

Outro critério para avaliar a capacidade de previsão de um modelo após validação cruzado LO é o coeficiente de correlação em validação cruzada, $q^2$ [87,88], calculado pela expressão da Equação 1.92, onde $\bar{Y}$ é a média dos valores da variável dependente.

$$q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{m}(\hat{Y}_i - \bar{Y})^2} \tag{1.92}$$

O método de validação LO permite também verificar a integridade do conjunto de dados. Se a amostra excluída for atípica, ocorre que o modelo construído para as demais amostras não será adequado para predizê-la, o que será verificado por uma alta contribuição ao PRESS.

Uma forma de quantificar a influência das amostras no modelo é através do parâmetro denominado *leverage*[4], que é calculado segundo as Equações 1.93, onde $\mathbf{T}$ é a matriz dos *scores*, $\mathbf{x}_i$ é o vetor da amostra $i$ e $\overline{\mathbf{x}}$ é a média das amostras.

$$\begin{aligned} \mathbf{H_0} &= \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}' \\ h_{ii} &= \tfrac{1}{n} + (\mathbf{x}_i - \overline{\mathbf{x}})^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \end{aligned} \tag{1.93}$$

O *leverage* da amostra $i$ é o elemento diagonal $h_{ii}$ da matriz $\mathbf{H_0}$. Os valores de $h_{ii}$ estão compreendidos entre zero e um. Geometricamente, o *leverage* pode ser interpretado como a distância de uma amostra ao centróide do conjunto de dados. Amostras que possuem grande influência no modelo possuem alto *leverage*. Para distinguir amostras anômalas,

[4]Influência.

costuma-se adotar um valor crítico acima do qual a amostra é considerada suspeita, definido segundo a Equação 1.94, onde $k$ é o número de fatores considerados no modelo e $m$ é o número de amostras [76].

$$h_{crit} = \frac{3k}{m} \tag{1.94}$$

Outra forma de avaliação das amostras é a análise do resíduo de Student ($S$), que tem a forma mostrada nas Equações 1.95, que consiste em uma estimativa do resíduo para uma amostra assumindo que esta não tenha sido usada na formação do modelo, somada a alguns ajustes, de forma a obter uma distribuição normal de resíduos [78].

$$\begin{aligned} L_i &= \sqrt{\frac{(Y_i - \hat{Y}_i)^2}{(m-1)(1-h_i)^2}} \\ S_i &= \frac{(Y_i - \hat{Y}_i)}{L_i\sqrt{1-h_{ii}}} \end{aligned} \tag{1.95}$$

Supondo-se a hipótese de distribuição normal, pode-se aplicar um teste $t$ como indicativo, para verificar se a amostra está ou não dentro da distribuição, com um nível de confiança de 95 %. Como os resíduos de Student são definidos em unidades de desvio padrão do valor médio, valores além de $\pm 2,5$ são considerados altos sob as condições usuais da estatística. Amostras com valores de referência errôneos tenderão a possuir altos resíduos de Student. Observa-se que os resíduos aumentam quando se inclui excesso de fatores (inclusão de ruído) e quando se incluem poucos fatores (informação insuficiente).

A combinação de *leverage* e resíduo de Student é uma forma de identificação de amostras anômalas. Quando uma amostra está acima dos limites para ambos os parâmetros, é considerada um *outlier*, e sua exclusão do conjunto de treinamento é indicada [78, 79].

Outro bom subsídio para avaliar o modelo é matriz ou vetor de resíduos de $\mathbf{Y}$. O gráfico dos resíduos em função da amostra deve ter distribuição aleatória, não apresentando padrão observável, o que indicaria um ajuste inadequado.

Outros índices que medem a capacidade de previsão da regressão linear também são úteis para avaliar o modelo. O índice $Q$, como a soma dos quadrados das linhas (amostras) da matriz de resíduos $\mathbf{E}$, é calculado pela Equação 1.96, onde $\mathbf{e}_i$ é a $i$-ésima linha de $\mathbf{E}$ e $\mathbf{I}$ é a matriz identidade.

$$Q_i = \mathbf{e}_i\mathbf{e}_i^T = \mathbf{x}_i(\mathbf{I}_{n,n} - \mathbf{P}\mathbf{P}^T)\mathbf{x}_i^T \tag{1.96}$$

O valor estatístico de $Q$ indica o quanto cada amostra está em conformidade com o modelo, medindo a diferença entre uma amostra e sua projeção nas $k$ variáveis latentes retidas no modelo. $Q$ é uma medida da variação dos dados fora das variáveis latentes incluídas no modelo PLS. Geometricamente, $Q$ é a distância do plano formado por duas variáveis latentes e $\sqrt{Q}$ é a distância da amostra ao plano das variáveis latentes.

A soma dos quadrados normalizados de *scores*, conhecida como $T^2$ estatístico de Hotteling, é uma medida da variação de cada amostra dentro do modelo, definida pela Equação 1.97, onde $\Lambda^{-1}$ é a matriz que contém a inversa dos autovalores associados aos $k$ autovetores guardados pelo modelo. $T^2$ é a medida da distância da projeção da amostra no espaço vetorial das variáveis latentes ao ponto médio multivariado.

$$T_i^2 = \mathbf{t}_i\Lambda^{-1}\mathbf{t}_i^T = \mathbf{x}_i\mathbf{P}\Lambda^{-1}\mathbf{P}^T\mathbf{x}_i^T \tag{1.97}$$

O limite para o resíduo $Q$ com limite $\alpha$ de confiança é calculado pela Equação 1.98, onde $\Theta_i = \sum_{j=k+1}^{N} \lambda_j{}^i$, $i = 1, 2, 3 \ldots$; $h_0 = 1 - 2\Theta_1\Theta_3/\Theta_2{}^2$; e $N$ é o número de fatores suficiente para acomodar toda a variância dos dados.

$$Q_\alpha = \Theta_1 \left[ 1 + \frac{c_\alpha \sqrt{2\Theta_2 h_0{}^2}}{\Theta_1} + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1{}^2} \right]^{\frac{1}{h_0}} \tag{1.98}$$

Para $T^2$, o limite de confiança estatístico é calculado por meio da distribuição $F$ pela Equação 1.99. O limite de $Q$ define uma distância ao plano que pode ser considerada não-usual, baseada nos dados utilizados para a construção do modelo. O limite de $T^2$ define uma elipse no plano, dentro da qual os dados normalmente são projetados.

$$T^2_{k,m,\alpha} = \frac{k(m-1)}{m-k} F^{\alpha}_{k,m-k} \tag{1.99}$$

Existem vários testes envolvendo os parâmetros de um modelo de regressão múltipla ou o coeficiente de correlação múltipla. Um teste usual é o de análise de variância, no qual se compara a variação explicada com a variação não explicada da variável dependente. Essa relação tem distribuição $F$, com $k$ e $(n - k - 1)$ graus de liberdade, sendo k o número de regressores e n o tamanho da amostra. Então, compara-se o parâmetro estatístico calculado $F_{calc}$ com o tabelado $F^{int}_{[k,n-k-1]}$. Sendo $F_{calc} > F_{tab}$, rejeita-se a hipótese nula de não existência de relação linear, de acordo com as indicações de significância da norma de avaliações determinadas pelo parâmetro *int*, ou seja, aprova-se (aceita-se) a equação de regressão. Para o modelo obtido, este teste é realizado com $F_{calc}$ calculado pela Equação 1.100, onde $\sigma^2{}_{exp}$ e $\sigma^2_{val}$ são as estimativas da variância dos dados experimentais e dos dados obtidos por validação cruzada com o modelo de regressão, respectivamente.

$$F_{calc} = \frac{\sigma^2_{exp}/k}{\sigma^2_{val}/(n-k-1)} \tag{1.100}$$

# Capítulo 2

# Estudo QSPR sobre os coeficientes de partição octanol/água

## 2.1 O coeficiente de partição octanol/água

O coeficiente de partição octanol/água, $logP_{o/w}$, continua sendo um dos descritores mais utilizados em química medicinal desde o seu surgimento na década de 50. Entender o significado desta informação estrutural representa uma parte do estudo realizado em química medicinal desde então. Trata-se da medida, geralmente feita experimentalmente pelo método conhecido como *shake-flask*, do equilíbrio resultante da dissolução de um soluto qualquer uma solução bifásica com uma fase aquosa e outra contendo $n$-octanol. O termo *shake-flask* é uma alusão direta ao método analítico utilizado: Chacoalhar o balão contendo a solução até o equilíbrio entre fases do soluto, decantar, separar as fases e obter a medida da concentração do soluto em cada fase. O valor de $logP_{o/w}$ é dados pela expressão da Equação 2.1, onde $[oc]$ é a concentração do soluto na fase orgânica e $[aq]$ é a concentração na fase aquosa. O procedimento experimental para determinação de $logP_{o/w}$ por medição da concentração em fases diferentes pode ser substituído pelo método cromatográfico, considerado tão bom quanto o tradicional.

$$logP_{o/w} = \log \frac{[oc]}{[aq]} \tag{2.1}$$

Desenvolvendo QSAR a química medicinal acumulou uma excelente bagagem sobre as interações hidrofóbicas e hidrofílicas. As interações hidrofóbicas são das mais importantes e menos compreendidas forças não covalentes na formação de ligações e interações biomacromoleculares. O efeito hidrofóbico é considerado em grande parte de origem entrópica. Infelizmente os efeitos entrópicos não podem ser incluídos no formalismo da maioria dos métodos utilizados para estudo estrutural, como mecânica quântica ou molecular. A estimativa da entropia pode ser feita com modelos que aplicam ciclos termodinâmicos de perturbação, que utilizam métodos quânticos ou de mecânica molecular, porém são computacionalmente caros [89].

Portanto utilizar a química quântica para estimar o valor de $logP_{o/w}$ corresponde, em parte, à verificar como os modelos de mecânica quântica são capazes predizer as

grandezas associadas aos efeitos que governam o equilíbrio de um soluto em fases distintas. Considerando que os efeitos que governam $logP_{o/w}$ também são os que governam parte da interação de uma droga com biomacromoléculas, trata-se de um tema importante quando se pretende aplicar química quântica ao desenvolvimento de modelos QSAR. A Figura 2.1 mostra esquematicamente como agem as forças que governam a interação entre as moléculas durante o processo de ligação.

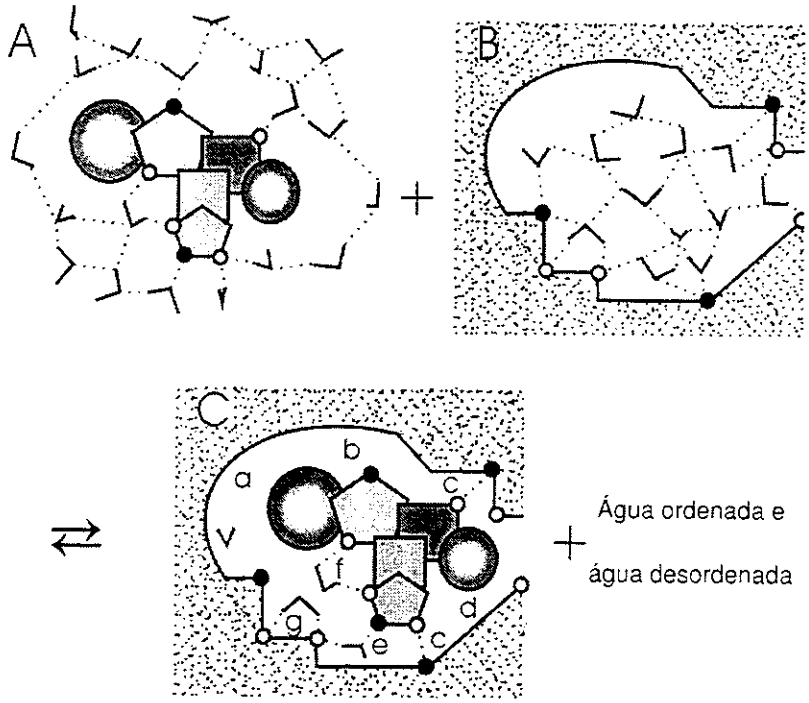



Figura 2.1: Esquema mostrando o processo de formação de uma ligação não covalente. (A) O ligante envolto por moléculas de água. As superfícies 'curvas' representam interações hidrofóbicas, e os vértices representam interações hidrofílicas. Círculos vazios são aceptores de pontes de hidrogênio e círculos cheios são doadores de elétrons para pontes de hidrogênio. (B) O sítio de ligação da macromolécula preenchido com moléculas de água. (C) O ligante alojado no receptor, com as seguintes características: (a) Interações do tipo hidrofóbica–hidrofóbica na superfície; (b) O ligante tem uma região doadora de elétrons para ponte de hidrogênio em uma região não aceptora hidrofóbica do receptor; (c) Pontes de hidrogênio formadas entre o ligante e o receptor; (d) Grupos hidrofóbicos do ligante em regiões polares do receptor; (e) Pontes de hidrogênio mediadas por água entre o ligante e o receptor; (f) Água ligada ao ligante, talvez trazida pelo ligante; (g) Água associada ao receptor, pouco afetada pela interação ligante–receptor.

Uma excelente revisão bibliográfica sobre os métodos de cálculo do coeficiente de partição octanol/água foi publicada por Albert J. Leo [90]. Cita como métodos principais (1) CLOGP (desenvolvido por Leo, Hansch et al.), (2) por 'substituintes' (desenvolvida por Fujita, Iwasa e Hansh), (3) por fragmentação usando $C_M$, (4) por contribuição atômica e/ou área superficial, (5) por propriedades moleculares e (6) por parâmetros solvatocrômicos.

A metodologia CLOGP (1) é basicamente um método por fragmentação. Estabelece que o valor de $logP_{o/w}$ pode ser estimado pela soma da contribuição de cada fragmento, multiplicado pelo número de vezes que o fragmento ocorre, mais uma soma de fatores multiplicada pela número de vezes que cada fator ocorre. O programa `CLOGP` tem um conjunto de regras simples de fragmentação, e um enorme banco de dados cobrindo os possíveis fragmentos e interações entre eles (fatores). O sucesso depende da aplicação correta e inequívoca da fragmentação, e da boa estimativa dos fatores que estabelecem as contribuições aditivas oriundas da proximidade de grupos. A obtenção do valor de $logP_{o/w}$ deste modo passa pela utilização do programa desenvolvido pelo grupo (`CLOGP`) porque é bastante complicado aplicar as regras de fragmentação e consultar as bibliotecas de dados. Este programa é vendido pelo autores.

O método 'por substituintes' (2) é também um método de fragmentação, que considera o valor de $logP_{o/w}$ como uma propriedade aditivo-constitutiva, relacionada à energia livre, que é numericamente igual à soma do $logP_{o/w}$ de um soluto ancestral mais um termo que representa a diferença entre o $logP_{o/w}$ entre um substituinte particular e o átomo de hidrogênio que foi substituído. O método foi desenvolvido para ser aplicado à substituição em anéis aromáticos. Investigações posteriores mostram que o método pode falhar se o hidrogênio substituído é parte de um grupo polar. Foram sugeridas modificações para corrigir esta anomalia.

Outro método por fragmentação, usando '$C_M$'(3) foi desenvolvido por Rekker et al. Este método usa a média da contribuição de fragmentos muito simples (C, CH, $CH_2$, $CH_3$, OH, $NH_2$. etc.), obtidas no maior banco de dados disponível. Não dá indicações sobre o que é um fragmento válido, e a quebra do soluto em fragmentos depende da escolha pessoal. O termo '$C_M$' refere-se à 'constante mágica' utilizada como fator no somatório de contribuições de fragmentos. Outro fator utilizado ($\omega$) por Waterbend e Testa em método semelhante é a quarta parte de $C_M$, e refere-se à hidratação. São fatores dificilmente relacionados a uma grandeza física bem definida.

Uma metodologia baseada em contribuições atômicas (4) foi desenvolvida por Broto et al. Usando técnicas de regressão e método de Monte Carlo foi criado um conjunto de duzentos e vinte e dois descritores. Este método não se presta ao cálculo de estruturas com potencial para produzir pontes de hidrogênio intermoleculares. Este método tem exatidão da ordem de 0,4 unidades logaritimicas na estimativa de $logP_{o/w}$. Ghose e Crippen conseguiram a mesma exatidão na estimativa, com um método semelhante usando cento e dez descritores. Esta mesma linha metodológica foi usada por Viswanadhan et al. para estender a contribuição para um número maior de átomos. A maior dificuldade nesta abordagem é exatamente a classificação dos tipos atômicos.

O programa `XLOGP`, que foi utilizado por nós para comparação entre os resultados obtidos, os publicados e os calculados pelo método aqui proposto, usa a metodologia de adição de átomos. Um total de noventa tipos atômicos são definidos pelos autores para classificar os diferentes ambientes químicos de carbono, nitrogênio, oxigênio, enxofre, fósforo e halogênios. São aplicados mais dez fatores de correção para lidar com algumas estruturas especiais. O valor de $logP_{o/w}$ é obtido da Equação 2.2, onde $A_i$ é a ocorrência do i-ésimo átomo e $B_j$ é a ocorrência do j-ésimo fator de correção; $a_i$ é a contribuição do

i-ésimo átomo e $b_j$ é a contribuição do j-ésimo fator de correção.

$$logP_{o/w} = \sum_i a_i A_i + \sum_j b_j B_j \tag{2.2}$$

Os coeficientes são derivados da regressão multivariada com um grande conjunto de treinamento ($n = 1853$; $r = 0,973$; $s = 0,349$). Este programa tem como grande atrativo o fato de ser gratuitamente disponibilizado à comunidade científica, mediante simples cadastramento junto aos autores.

Uma metodologia de cálculo de $logP_{o/w}$ baseada em áreas acessíveis ao solvente (SASA) foi proposta por Iwase, Moriguchi, et al. O método utiliza-se de interações com uma 'capa' de solvente ($S_A$) e de uma interação tipo hidrofóbica ($S_H$, semelhante à usada por Rekker). Para estruturas simples uma equação com apenas dois parâmetros produziu excelentes resultados ($r = 0,995$; $s = 0,13$). Esta abordagem baseada na superfície do soluto foi utilizada em conjunto com técnicas de PCA por Dunn e colaboradores. Estes averiguaram que cerca de 60% da variância para um conjunto de seis sistemas de solventes aquoso/apolar (apolar: octanol, éter, clorofórmio, benzeno, tetracloreto de carbono ou hexano) podem ser relacionadas às propriedades do soluto em água. Esta conclusão foi baseada na observação de que a componente principal que continha esta variância era praticamente estável para os seis sistemas. Uma segunda componente principal foi considerada responsável pela contribuição da área total acessível ao solvente. O método de 'solvatação' proposto é empírico, e os resultados dependem fortemente da maneira como isto é feito.

Moriguchi et al. tentaram também a utilização de descritores de tipo atômico conjuntamente com fatores para efeitos de proximidade, insaturação, pontes de hidrogênio intramoleculares, estruturas cíclicas e propriedades anfotéricas. Utilizando um grande conjunto de estruturas da base de dados *Pomona Masterfile* estes autores apresentaram uma equação de regressão com quatorze parâmetros capaz de prever o valor de $logP_{o/w}$ de mil duzentos e trinta solutos com coeficiente de regressão $r = 0,952$ e desvio padrão $s = 0,411$.

Klopman e Wang usaram um método com dez parâmetros atômicos e desenvolveram setenta e seis fragmentos para levar em conta o ambiente de ligação próximo. Duas variáveis indicadoras para hidrocarbonetos alifáticos e amino ácidos. Para um conjunto de novecentos e trinta e cinco estruturas no conjunto de teste apenas trinta e nove dos oitenta e oito foram considerados significativos. Os resultados são bons, $r = 0,965$ e $s = 0,385$. Klopman admite que a variável indicadora hidrocarbonetos saturados resulta em grandes desvios negativos (-0,6 para etano) para as cadeias pequenas, e também em desvios positivos para cadeias longas (+0,9 para octano).

A abordagem por meio de propriedades moleculares (5) é baseada na premissa de que as moléculas são mais que a soma das suas partes [89], como são consideradas normalmente pelas técnicas de fragmentação. Rogers e Camarata desenvolveram um método de cálculo de $logP_{o/w}$ para solutos aromáticos usando um termo de densidade, $Q_s^T$, conjuntamente com um termo de polarização induzida, $S_s^E$. Postularam que a partição para fase aquosa é controlada por carga, e para fase apolar controlada por polarizabilidade. O método não evoluiu, e não tem sido amplamente utilizado.

Um método conhecido como 'análise conformacional dependente do solvente' foi desenvolvido por Hopfinger e Battershell usando procedimentos semi-empíricos. É muito rápido se comparado a outras técnicas que utilizam orbitais moleculares. Para hidrocarbonetos simples, aromáticos ou alifáticos, e para solutos monofuncionalizados o erro é ligeiramente menor que os obtidos com o método de Fujita. O método não parece adequado para análise de solutos com grupos muito polares, ou para soluto com grupos polares próximos.

Uma equação proposta por Bodor et al. usando quinze parâmetros derivados de estruturas obtidas com método AM1: Uma variável indicadora para alcanos, a massa molecular, a área da superfície molecular, e o seu quadrado, a soma do valor absoluto das cargas sobre oxigênios e nitrogênios, a raiz quadrada da soma dos quadrados dos átomos de oxigênio conjuntamente com seu quadrado e a sua quarta potência, o mesmo para os nitrogênios, e o momento de dipolo calculado. Este método tem correlação $r = 0,938$ com erro padrão $s = 0,296$. Na tentativa de melhorar a performance foram adicionados os parâmetros número de átomos de carbono, quarta potência da ovalidade, e a soma dos valores absolutos da carga atômica em cada átomo. Este conjunto de parametrização dificilmente poderia ser classificado como 'não empírico'. Ovalidade, por exemplo, é definida como o quociente entre a superfície molecular e a superfície mínima calculada.

Sasaki et al. propuseram outra metodologia derivada da estrutura molecular, baseada em mecânica molecular e métodos com orbitais. Eles recomendam o uso de cálculos *ab initio* para determinação do potencial eletrostático da superfície. São calculados parâmetros para tensão superficial (S), para interação eletrostática (ES) e para transferência de carga (CT). Análise por regressão múltipla é usada para relacionar estes parâmetros à $logP_{o/w}$. Um coeficiente de correlação $r = 0,983$ com desvio $s = 0,260$ foi obtido para sessenta e três solutos. Este método tem um grande erro na previsão do ácido benzóico, e de muitos heterociclos.

Os métodos conhecidos como 'solvatocrômicos' (6) propostos por Kamlet, Taft, Abraham et al. são essencialmente baseados em propriedades moleculares. Trata-se de uma equação de regressão multivariada que utiliza como parâmetros um termo de volume molar, um termo para polarizabilidade e polaridade do soluto, um termo para incluir uma medida da capacidade de realizar ponte de hidrogênio, e um termo para definir o ponto de intersecção da regressão. Por que o valor do momento de dipolo calculado por métodos quânticos não é uma boa aproximação, se considerado um conjunto grande de moléculas, foram propostas metodologias alternativas para calcular as polarizabilidades, no método denominado Polarizabilidade de Excesso.

Para estabelecer quais descritores podem ser úteis na obtenção de correlação entre modelos quânticos e coeficientes de partição entre solventes deve-se compreender as interações que governam os efeitos entre soluto e solvente. Todos os efeitos entálpicos intramoleculares ou intermoleculares são de natureza eletrostática porque envolvem interações eletrônicas. De várias formas, são descritos por diferentes termos: Ligações de hidrogênio, interações tipo polar, interações eletrostáticos e de van der Waals. Os efeitos entrópicos não são eletrostáticos. Estes termos descrevem aspectos diferentes de um fenômeno eletrônico, e deve-se tomar cuidado para não sobrestimar algum efeito, sobrepondo sua descrição em dois termos diferentes.

As interações entre dois corpos carregados são alteradas pela presença de moléculas de solvente entre elas. Este efeito de blindagem é descrito pela constante dielétrica do solvente. A variação da constante dielétrica do solvente se dá por causa da polarizabilidade do solvente e pela sua orientação em resposta à polarização. Ambos os efeitos diminuem o campo elétrico, e portanto diminuem a energia das interações entre cargas. Estes efeitos podem incluir muitas moléculas de solvente, e não apenas aquelas entre as cargas, de modo que a representação acurada do solvente por meios implícitos é extremamente difícil por que o espaço ao redor do soluto pode ser preenchido por diversos tipos de moléculas, heterogeneamente distribuídos. As contribuições eletrostáticas podem ser classificadas nos seguintes grupos [91]:

- As ligações de hidrogênio têm papel fundamental quando o solvente é água. A solubilidade de um soluto em água depende da sua capacidade de substituir as pontes de hidrogênio da água. Quando o soluto realiza pontes de hidrogênio com a água com menor energia que entre as moléculas de soluto, as moléculas de água se reorganizam e aumentam o número de pontes de hidrogênio para minimizar a perda de energia. Este ganho na entalpia vem acompanhado com uma diminuição da entropia porque a mobilidade da água diminui. Este efeito entrópico é conhecido como hidrofobicidade, e o processo como um todo é energeticamente desfavorável.

- Interações tipo monopolo–monopolo e dipolo–monopolo entre o soluto e a água têm diversas origens. Átomos carregados têm interações fortes (monopolo–monopolo) e interagem com moléculas de água (dipolo–monopolo). O potencial eletrostático gerado pelo sistema soluto–solvente polariza tanto o solvente como o soluto.

- Interações tipo dipolo–dipolo são freqüentes na água. A água tem expressivo momento de dipolo, e os solutos freqüentemente também têm. Proteínas por exemplo, têm muitas regiões com momento de dipolo, a começar pela ligação peptídica.

- Interações tipo dipolo induzido–dipolo induzido (van der Waals) entre água e solutos são de pequena magnitude, já que a água é uma molécula pequena e móvel, e se reorganiza para cobrir toda a superfície acessível do soluto da melhor maneira possível. A energia de van der Waals não varia muito por causa da falta de impedimento estérico.

Os efeitos entrópicos dependem da organização do sistema soluto–solvente, em escala macroscópica. Como os movimentos do soluto dentro da água são fáceis, porque a viscosidade da rede de água é baixa, a representação dos efeitos devidos à variação da entropia em sistemas água-soluto é muito complicada. Isso é verdade para todos os sistemas com efeito implícito de solvatação. A correta representação da entropia, a despeito da acuracidade ou não dos cálculos de entalpia, compreende um entendimento profundo dos sistemas sob estudo. Simulações de estruturas ou processos biomacromoleculares envolvendo água dependem do conhecimento da estrutura cristalográfica do sistema como ponto de partida. Podem ser realizadas longas simulações por métodos de dinâmica molecular, que produzem resultados que podem ser interpretados como resposta ao problema da entropia. Estas simulações são bastante 'caras' do ponto de vista computacional, e não raro

a estrutura obtida após a simulação perde muito sua conformação secundária e terciária original obtida na cristalografia [92].

### 2.1.1 Os compostos escolhidos para o cálculo

Os compostos com toxicidade para peixes têm sua atividade muito ligada à solubilidade em fase aquosa. A classificação destes poluentes com mecanismo de ação semelhante em quatro níveis de toxicidade, leva este fator em consideração diretamente. Classe 1: Compostos de toxicidade mais baixa, também chamados de toxicidade de basal, são geralmente compostos apolares. Classe 2: Os compostos de toxicidade intermediária, também chamados de toxicidade aguda, são geralmente compostos polares. Classe 3: Os grupos de toxicidade mais alta são classificados como quimicamente reativos. Classe 4: Os compostos de ação específica, como os pesticidas e defensivos agrícolas [93].

Os métodos de classificação levam em conta a presença ou ausência de grupos funcionais, tradicionalmente. Esta classificação fica restrita aos compostos que enquadram-se nas regras estabelecidas. Contudo, levando-se em conta apenas os coeficientes de partição e mais alguns descritores mecânico-quânticos, Hermens [94] e colaboradores mostraram que pode-se realizar uma boa classificação para os poluentes das Classes 1 e 2.

Este trabalho mostra que mesmo o coeficiente de partição ($logP_{o/w}$) pode ser obtido por modelagem molecular, usando os descritores mecânico-quânticos usados por Hermens e mais alguns, aqui propostos [95]. Os valores de $logP_{o/w}$ foram modelados com o método dos Mínimos Quadrados Parciais (*Partial Least Squares*: PLS), usando descritores mecânico-quânticos obtidos de cálculos *ab initio*, com aplicação de efeitos de solvatação. As estruturas das cento e oitenta e oito moléculas mostradas na Tabela 2.1 foram utilizadas na elaboração dos modelos QSPR.

Foram escolhidos três solventes: (1) a água, que além de ser uma das fases usadas para obter o valor experimental de $logP_{o/w}$, é o solvente biológico; (2) o cicloexano que é um solvente apolar, e foi usado para modelar as interações soluto/solvente presentes na parte apolar do octanol; (3) a acetona, que é um solvente polar não prótico, deve ser capaz de modelar bem as interações entre a parte polar do *n*-octanol e o soluto [96]. O *n*-octanol não é um solvente adequado (molécula grande e não esférica) para utilização com o método PCM de solvatação, principalmente por causa da formulação do termo de cavitação.

A correlação entre valores obtidos em cálculos teóricos e coeficientes de partição já foi conseguida por diversos autores, porém no caso aqui apresentado são utilizados um número menor de descritores que nos exemplos já publicados [90], com resultados de reprodução dos valores experimentais comparáveis. Os métodos utilizados para obtenção dos modelos de regressão são baseados em regressão PLS e redes neurais. Estes métodos foram escolhidos pela sua boa aplicabilidade ao caso. Hermens utiliza-se de regressão PLS, e Análise Discriminante (também baseada em método PLS) para obter regressão entre valores de toxicidade e classificação do conjunto de poluentes.

## 2.2 Objetivos

- Obter um método para calcular o coeficiente de partição octanol/água ($logP_{o/w}$) com descritores mecânico-quânticos.
- Identificar grupos funcionais para os quais a previsão com o método é ruim.
- Determinar como os modelos de solvatação podem incluir informação útil na previsão de $logP_{o/w}$.

## 2.3 Métodos de trabalho e procedimentos adotados

Inicialmente são calculados os descritores quânticos, para então obter uma matriz de dados. Esta matriz de dados será analisada e utilizada para modelar os valores dos coeficientes de partição. Cada uma destas etapas será melhor descrita adiante.

### 2.3.1 Procedimentos mecânico-quânticos

As estruturas das 188 moléculas mostradas na Tabela 2.1 foram submetidas a uma busca das conformações de menor energia. Esta busca foi feita com método sistemático para os compostos com substituição direta no anel, usando Hamiltoniano AM1. Este método é implementado no programa `MOPAC` [97]. Para os demais casos o método de geração das estruturas para otimização foi o da matriz de distâncias geométricas implementado no programa `TINKER` [98] denominado `distgeom`. O método de otimização geométrica das estruturas geradas com o `TINKER` também foi o AM1. Para descartar as conformações que tornam-se idênticas após a minimização de energia foi utilizado o programa `superpose` que também faz parte do `TINKER`. Um programa desenvolvido por Oliveira [99] faz a comunicação entre os diversos módulos e programas para a busca conformacional e descarte das estruturas repetidas. Os programas que automatizam o procedimento de busca conformacional estão transcritos nos apêndices. Para geração de estruturas foi criado o programa transcrito no Apêndice A.1, denominado `cs.pl`, e para eliminação das estruturas repetidas o programa transcrito no Apêndice A.1.1, denominado `uniquefy.pl`.

Todas as estruturas das moléculas obtidas nos mínimos de energia, pré otimizadas com AM1, foram completamente otimizadas usando o método *ab initio* com base HF/3-21G, com o programa `GAMESS` [100]. Propriedades moleculares foram calculadas com método *ab initio*. Para comparar o resultado do aumento do conjunto de base no sistema foram usadas as bases HF/3-21G e HF/6-311G. O efeito de solvatação foi estudado com aplicação de modelos progressivamente mais complexos. Para cada nível de complexidade de modelo foram realizados cálculos para três tipos diferentes de solvente: Água, acetona e cicloexano. Para automatizar os cálculos de otimização de geometria e de inclusão dos efeitos de solvatação foi escrito o programa denominado `runall_solv.csh`, transcrito no Apêndice A.2.2.

Foram calculadas quinze propriedades para o Modelo Um (M1) e trinta para os Modelos Dois (M2) e Três (M3). Na Tabela 2.1 são mostrados os valores previstos em validação

cruzada tipo *Leave One Out* (LO) para os compostos. O modelo M1 usa base HF/3-21G e aplica efeito de cavitação segundo método de Pierotti [101] e Claverie [102]. Este tipo de cavidade será denominado S1. O modelo M2 também usa base HF/3-21G e aplica efeito de cavitação, efeito da energia livre de repulsão, e das forças de dispersão; segundo método de Amovilli e Mennucci [103], Este tipo de cavidade será denominado S2. Usando HF/6-321G//3-21G foram calculadas propriedades usando apenas o efeito de cavitação (cavidade S1) para criar o modelo M3.

Para os modelos M1 e M3 foram calculadas as seguintes propriedade: Os Momentos de Dipolo (três descritores: $\mu_{aquoso}$, $\mu_{acetona}$, e $\mu_{cicloexano}$). As cargas calculadas sobre o átomo mais negativamente carregado e as cargas calculadas sobre o hidrogênio mais positivamente carregado [104] (ou átomo de carbono ou nitrogênio mais positivamente carregado, nas moléculas sem hidrogênio) somam mais seis descritores.

O cálculo realizado calcula a energia de cavitação ao final usando o método de Pierotti e Claverie, utilizando a temperatura de 310 K. O método proposto por Pierotti e Claverie é adequado para cálculo utilizando solventes de tamanho pequeno (o nosso caso), mesmo que sejam polares, próticos ou que façam pontes de hidrogênio. Este procedimento não é o mais adequado no caso de solventes grandes, com moléculas não esféricas (ex.*n*-alcanos).

A adequação deste método para o cálculo da energia livre de Gibbs de solvatação depende simultaneamente das interações eletrostáticas consideradas, como do processo de obtenção da cavidade. A utilização do processo de Ben–Naim [105] de solvatação do soluto M no solvente S é a transferência de M da fase gasosa ideal para um posição fixa dentro do solvente S, à temperatura constante. A energia livre de Gibbs pode ser relacionada ao trabalho necessário para construir M no solvente S. O trabalho pode ser dividido em quatro componentes: (**ES**) eletrostática, (**CV**) de cavitação, (**DP**) de dispersão e (**RP**) de repulsão. A inclusão sucessiva dos termos (**ES**,**CV**) ou (**ES**, **CV**, **DP**, **RP**) no resultado é uma tentativa de separar empiricamente como cada termo pode representar os efeitos da solvatação real. Os resultados de correlação entre os descritores e os valores experimentais dos coeficientes de partição são úteis para realizar esta avaliação.

As cargas foram calculadas usando o método CHELPG para todos os casos. Para o cálculo foi utilizado um arranjo com alta densidade de pontos para determinação das cargas. Foi determinado que a distância entre os pontos no arranjo não fosse maior que 0,3 Å. Foi utilizado um arranjo com alta densidade de pontos de amostragem no ajuste de cargas, para minimizar os efeitos conformacionais na magnitude das cargas calculadas [106], uma vez que a conformação geometria das moléculas é ligeiramente alterada em função das substituições, para cada composto da série. O método CHELPG calcula cargas pontuais, nas posições atômicas, que são capazes de ajustar o valor da carga total molecular. Pode-se definir que as cargas também sejam restritas às capazes de ajustar os vetores do momento de dipolo e de quadrupolo, obtidos diretamente da aplicação do operador na função de onda Hartree-Fock. Em todos os casos foram aplicadas todas as restrições possíveis. O número total de restrições não pode ser maior que o número de átomos presentes na molécula.

Para todos os modelos (M1, M2 e M3) foram definidos três valores de partição ($P$) relativos aos três pares formados com os três solventes utilizados. Os coeficientes de partição ($P_{(a,b)}$) para cada par de solventes $(a, b)$ são calculados utilizando a energia total

de solvatação para cada solvente ($\varepsilon_S$), obtida como resultado dos cálculos quânticos de solvatação, como aproximação para a variação da energia Livre de Gibbs ($\Delta G$). Assim podemos relacionar os valores de $P$ aos de $\Delta G$ pelas Equações 2.3.

$$\begin{aligned} \Delta G_a \equiv \varepsilon_S^{(a)} \quad &, \quad \Delta G_b \equiv \varepsilon_S^{(b)} \\ -RT \ln \tfrac{k_a}{k_b} &= \Delta G_a - \Delta G_b \\ P_{(a,b)} &= \tfrac{k_a}{k_b} \\ P_{(a,b)} &= \exp\left(-\tfrac{\varepsilon_S^{(a)} - \varepsilon_S^{(b)}}{RT}\right) \end{aligned} \tag{2.3}$$

Os volumes da cavidade formada por cada solvente também foram utilizados como descritores. Estas propriedades podem ser calculadas para o cicloexano utilizando o programa GAMESS, adicionando-se as seguintes diretivas ao bloco de comandos:

```
$pcm icomp=3 icav=1 tabs=310 idisp=0 ifield=1 solvnt=cychex $end
$elpot iepot=1 where=pdc $end
$elmom iemom=3 where=nuclei $end
$eldens ieden=1 $end
$elfldg iefld=2 $end
$pdc ptsel=chelpg constr=dipole delr=0.3 maxpdc=100000 $end
```

Para o Modelo Dois (M2) foram calculadas mais cinco propriedades: (i) a variação da energia interna do solvente; (ii) a interação eletrostática soluto–solvente; (iii) a energia de cavitação de Pierotti; (iv) a energia livre de dispersão e (v) a energia livre de repulsão. Estes efeitos são incluídos no Hamiltoniano. Para cada uma das cinco propriedades são calculados descritores nos três solventes, totalizando mais quinze descritores ao bloco das variáveis independentes. Os descritores são as grandezas obtidas a partir das propriedades calculadas, e que são utilizados efetivamente na construção dos modelos. Cada conjunto de propriedades pode ser calculado para o cicloexano alterando o bloco de comandos $pcm do arquivo de entrada para o GAMESS para o seguinte:

```
$pcm icomp=3 icav=1 irep=1 idp=1 eta2=2.035 wb=0.4225
     tabs=310 idisp=0 ifield=1 solvnt=cychex $end
$newcav rhow=0.750 pm=84.16 $end
```

Para criação de uma cavidade de solvente tipo S2 são necessários o índice de refratividade elevado ao quadrado ($\eta^2$), o potencial de ionização do solvente ($PI$), a densidade do solvente relativa a água($\rho$) e a massa molar do solvente (MM). Para a acetona foram usados $\eta = 1,359$; $PI = 0,413$; $\rho = 0,788$ e $MM = 58,08$. Para o cicloexano foram usados os valores de $\eta = 1,426$; $PI = 0,4225$; $\rho = 0,750$ e $MM = 84,16$. Os valores de $\eta$ e $\rho$ foram obtidos na literatura [107], o PI foi calculado por método *ab initio* no nível HF/6-311G, e a MM através da fórmula química. Os valores para a água constam no programa GAMESS.

### 2.3.2 Métodos estatísticos

Os valores de $logP_{o/w}$ para os cento e oitenta e oito compostos orgânicos mostrados na Tabela 2.1 foram modelados utilizando a matriz de dados obtida com os cálculos quânticos, autoescalada. O conjunto de moléculas utilizado para treinar os modelos de regressão inclui álcoois, halobenzenos, anilinas, fenóis, nitrobenzenos, ésteres, aminas e piridinas. Os valores de $logP_{o/w}$ para os compostos foram obtidos na literatura, publicados por Hermens e colaboradores [94]. Também foram obtidos valores de $logP_{o/w}$ utilizando o programa `XLOGP` [108] para os compostos derivados de piridina. Os resultados de validação 'fora um' deram origem as estimativas dos três modelos (LO1, LO2 e LO3) da Tabela 2.1.

Foram obtidos modelos de regressão com métodos PLS e Redes Neurais de Retropropagação (*Back Propagation Neural Network*: BPN). Na regressão com método PLS foram implementadas funções [109] no programa `OCTAVE` [110] para a obtenção dos modelos de regressão. Os métodos para a validação cruzada tipo 'fora um' e para a validação cruzada em bloco também foram implementados. Para estes modelos obtidos com método PLS foram utilizadas entre cinco e dez Variáveis Latentes (VL) para regressão. Este número de variável latente foi selecionado baseado no valor do coeficiente de correlação em validação cruzada ($q^2$) obtido dos valores previstos com LO comparados com os valores publicados. Os valores das variâncias acumuladas pela regressão PLS dos conjuntos de dados experimentais e dos conjuntos de descritores mecânico-quânticos são mostrados na Tabela 2.2.

Foi treinada uma BPN utilizando os *scores* de uma Análise por Componentes Principais (*Principal Component Analysis*: PCA) realizada para o conjunto de trinta descritores obtidos para o Modelo dois (M2). Os dados de acumulação de variância pela PCA são mostrados na Tabela 2.2, na sexta coluna. O variância acumulada de 99,19% permite concluir que toda a informação existente no conjunto de trinta descritores foi capturada pelas dez Componentes Principais (*Principal Components*: PC). Esta rede foi treinada com onze neurônios na primeira camada: dez para os dados das PCs da PCA, mais um de *bias* . Na segunda camada foram usados trinta e seis neurônios. O número de neurônios da segunda camada foi determinado por tentativa e erro. Na terceira camada foi usado um neurônio para saída do sinal obtido, que corresponde a resposta do modelo para o valor de $logP_{o/w}$. O valor limite para o erro de treinamento da rede para os compostos foi de $10^{-4}$. O limite de iterações para atingir convergência usado foi de cinqüenta mil. Esta rede foi utilizada para previsão em validação cruzada tipo LO dos compostos desta série. O programa utilizado para treinamento de redes neurais foi o `PSDD` [111].

Tabela 2.1: Código atribuído, nomes dos compostos utilizados para teste da metodologia de cálculo do valor de $logP_{o/w}$ aqui proposto, valores experimentais publicados e valores estimados pelos Modelos 1, 2 e 3 citados no texto.

| Código | Nome | Exp. | LO1 | LO2 | LO3 |
|---|---|---|---|---|---|
| **logP1** | guanidina | -1,510 | -1,319 | -1,561 | -1,728 |
| **logP2** | 1,2-etanodiol | -1,360 | -0,761 | -0,927 | -0,605 |
| **logP3** | 2-aminoetanol | -1,310 | -0,350 | -0,462 | -0,275 |
| **logP4** | dietilenoglicol | -1,305 | -0,121 | -0,384 | -0,591 |
| **logP5** | trietilenoglicol | -1,240 | 0,575 | 0,519 | 0,111 |
| **logP6** | 1-amino-2-propanol | -0,960 | 0,556 | -0,217 | 0,340 |
| **logP7** | 2-metoxietanol | -0,770 | 0,997 | 0,689 | 0,374 |
| **logP8** | 2-metoxietilamina | -0,672 | 0,830 | 0,465 | 0,177 |
| **logP9** | etanol | -0,310 | -0,233 | -0,485 | -0,202 |
| **logP10** | acetona | -0,240 | -1,029 | -1,109 | -1,637 |
| **logP11** | etilamina | -0,130 | 0,254 | -0,056 | -0,016 |
| **logP12** | 2-etoxietanol | -0,100 | 0,898 | 0,841 | 0,945 |
| **logP13** | 4-hidroxianilina | 0,040 | 0,518 | 0,384 | 0,535 |
| **logP14** | 2-isopropoxietanol | 0,050 | 2,576 | 1,734 | 1,394 |
| **logP15** | 2-propanol | 0,050 | 0,460 | 0,034 | 0,413 |
| **logP16** | 4-hidroximetilfenol | 0,250 | 1,084 | 0,932 | 1,013 |
| **logP17** | *t*-butanol | 0,350 | 0,960 | 0,236 | 0,711 |
| **logP18** | 4-(*n*-metoximetil)aminofenol | 0,475 | 1,886 | 2,097 | 2,215 |
| **logP19** | hidroquinona | 0,590 | -0,168 | -0,002 | -0,140 |
| **logP20** | 3-nitropiridina | 0,660 | 1,131 | 0,899 | 0,651 |
| **logP21** | 1,3-diidroxibenzeno | 0,800 | 0,313 | 0,145 | 0,355 |
| **logP22** | 2-butoxietanol | 0,830 | 1,854 | 1,170 | 0,895 |
| **logP23** | dietiléter | 0,870 | 1,913 | 1,632 | 0,980 |
| **logP24** | anilina | 0,900 | 0,905 | 1,091 | 0,924 |
| **logP25** | 4-amino-2-nitrofenol | 0,960 | 1,672 | 1,257 | 1,461 |
| **logP26** | butilamina | 0,970 | 1,393 | 0,833 | 0,942 |
| **logP27** | 4-metilamoniofenol | 0,974 | 1,086 | 0,854 | 1,458 |
| **logP28** | benzilamina | 1,090 | 1,314 | 1,540 | 1,245 |
| **logP29** | 1,2-dimetilpropilamina | 1,102 | 2,297 | 1,669 | 2,057 |
| **logP30** | 2,2-dimetilpropilamina | 1,192 | 2,003 | 1,014 | 1,117 |
| **logP31** | 3-pentanol | 1,210 | 1,615 | 1,002 | 1,490 |
| **logP32** | diclorometano | 1,250 | 0,684 | 0,953 | 1,482 |
| **logP33** | 2-metilanilina | 1,320 | 1,528 | 1,657 | 1,465 |
| **logP34** | 4-metoxifenol | 1,340 | 1,420 | 1,604 | 1,085 |
| **logP35** | tiazol | 1,350 | 0,608 | 1,046 | 0,873 |
| **logP36** | 3-nitroanilina | 1,370 | 1,475 | 1,388 | 1,362 |

Continua na próxima página

Tabela 2.1: Continuação

| Código | Nome | Exp. | LO1 | LO2 | LO3 |
|---|---|---|---|---|---|
| **logP37** | 3-cloropiridina | 1,390 | 1,869 | 1,985 | 1,681 |
| **logP38** | 4-cloropiridina | 1,390 | 1,270 | 1,761 | 1,259 |
| **logP39** | 4-metilanilina | 1,390 | 1,513 | 1,784 | 1,566 |
| **logP40** | 4-nitroanilina | 1,390 | 1,584 | 1,448 | 1,222 |
| **logP41** | 3-metilanilina | 1,400 | 1,798 | 1,390 | 1,522 |
| **logP42** | 1-adamantanoamina | 1,436 | 2,266 | 1,729 | 1,564 |
| **logP43** | 3-etilpiridina | 1,460 | 2,304 | 2,180 | 1,974 |
| **logP44** | fenol | 1,460 | 1,039 | 0,825 | 1,067 |
| **logP45** | 1,2-dicloroetano | 1,480 | 1,024 | 1,361 | 1,270 |
| **logP46** | 2-cloropiridina | 1,480 | 1,893 | 2,264 | 1,713 |
| **logP47** | 4-bromopiridina | 1,570 | 1,943 | 1,828 | — |
| **logP48** | 3-metoxifenol | 1,580 | 1,546 | 1,340 | 1,413 |
| **logP49** | 4-cianofenol | 1,600 | 1,099 | 1,391 | 1,333 |
| **logP50** | 3,3-dimetilbutilamina | 1,721 | 2,258 | 1,354 | 1,681 |
| **logP51** | 1,1-dicloroetano | 1,790 | 1,378 | 1,622 | 2,128 |
| **logP52** | 2-nitroanilina | 1,850 | 1,817 | 1,564 | 1,252 |
| **logP53** | nitrobenzeno | 1,850 | 1,604 | 1,356 | 1,593 |
| **logP54** | 3-etilanilina | 1,855 | 2,085 | 2,150 | 2,022 |
| **logP55** | 4-etilanilina | 1,855 | 2,046 | 2,254 | 2,072 |
| **logP56** | 3-cloroanilina | 1,880 | 1,956 | 2,217 | 2,055 |
| **logP57** | 4-cloroanilina | 1,880 | 1,961 | 2,258 | 2,052 |
| **logP58** | 1,1,2-tricloroetano | 1,890 | 1,309 | 1,654 | 1,737 |
| **logP59** | 2-cloroanilina | 1,900 | 1,920 | 2,149 | 1,862 |
| **logP60** | 4-nitrofenol | 1,910 | 0,988 | 0,834 | 1,125 |
| **logP61** | 4-metilfenol | 1,940 | 1,661 | 1,508 | 1,771 |
| **logP62** | 2-metilfenol | 1,950 | 1,294 | 1,754 | 1,578 |
| **logP63** | 3-metilfenol | 1,960 | 1,993 | 1,490 | 1,800 |
| **logP64** | clorofórmio | 1,970 | 1,658 | 1,819 | 2,783 |
| **logP65** | 1,2,3-tricloropropano | 1,980 | 1,695 | 1,647 | 1,918 |
| **logP66** | 1,2-dicloropropano | 1,987 | 1,621 | 1,844 | 1,721 |
| **logP67** | 1,3-dicloropropano | 2,000 | 1,560 | 1,852 | 1,738 |
| **logP68** | 3-nitrofenol | 2,000 | 0,876 | 0,830 | 1,048 |
| **logP69** | 3,4-dicloropiridina | 2,010 | 2,491 | 2,612 | 2,230 |
| **logP70** | 3,5-dicloropiridina | 2,010 | 2,509 | 2,696 | 2,416 |
| **logP71** | 2-cloro-4-nitroanilina | 2,055 | 2,378 | 2,210 | 2,023 |
| **logP72** | 2,4-dicloropiridina | 2,100 | 2,643 | 2,918 | 2,506 |
| **logP73** | 2,5-dicloropiridina | 2,100 | 2,672 | 2,996 | 2,397 |
| **logP74** | 2-clorofenol | 2,150 | 1,878 | 1,952 | 2,312 |
| **logP75** | benzeno | 2,186 | 1,041 | 1,372 | 1,732 |
| **logP76** | 3,4-dimetilfenol | 2,230 | 2,108 | 2,065 | 2,306 |

Continua na próxima página

Tabela 2.1: Continuação

| Código | Nome | Exp. | LO1 | LO2 | LO3 |
|---|---|---|---|---|---|
| **logP77** | 2-cloronitrobenzeno | 2,240 | 2,451 | 2,556 | 2,212 |
| **logP78** | 3-benziloxipiridina | 2,250 | 3,159 | 3,876 | 2,698 |
| **logP79** | 4-bromoanilina | 2,260 | 2,262 | 2,219 | — |
| **logP80** | 2,4-dimetilfenol | 2,300 | 2,077 | 2,021 | 2,351 |
| **logP81** | 2-nitrotolueno | 2,300 | 2,303 | 2,098 | 2,094 |
| **logP82** | 2-cloro-6-nitrofenol | 2,326 | 2,376 | 2,099 | 2,040 |
| **logP83** | 2,6-dimetilfenol | 2,360 | 1,954 | 1,934 | 2,307 |
| **logP84** | 4-nitrotolueno | 2,370 | 2,285 | 2,255 | 2,279 |
| **logP85** | 4-etóxi-2-nitroanilina | 2,387 | 2,916 | 2,631 | 2,398 |
| **logP86** | 1,1,2,2-tetracloroetano | 2,390 | 2,395 | 2,393 | 3,186 |
| **logP87** | 4-clorofenol | 2,390 | 1,555 | 1,814 | 1,909 |
| **logP88** | 4-cloronitrobenzeno | 2,390 | 2,019 | 2,166 | 2,013 |
| **logP89** | 3-nitrotolueno | 2,420 | 2,392 | 2,103 | 2,362 |
| **logP90** | tricloroeteno | 2,420 | 2,417 | 2,352 | 3,313 |
| **logP91** | 3-cloronitrobenzeno | 2,460 | 2,382 | 2,381 | 2,314 |
| **logP92** | 1,1,1-tricloroetano | 2,490 | 2,295 | 2,541 | 2,894 |
| **logP93** | 3-clorofenol | 2,500 | 1,878 | 1,926 | 2,189 |
| **logP94** | 2-alilfenol | 2,548 | 2,533 | 2,468 | 2,712 |
| **logP95** | 4-etilfenol | 2,580 | 2,084 | 1,903 | 2,249 |
| **logP96** | 1-clorobutano | 2,640 | 2,062 | 1,852 | 1,958 |
| **logP97** | 2-cloro-4-metilfenol | 2,654 | 2,599 | 2,719 | 3,020 |
| **logP98** | 3,4-dicloroanilina | 2,690 | 2,783 | 3,008 | 2,785 |
| **logP99** | 2,3,4-tricloropiridina | 2,720 | 3,418 | 3,434 | 3,088 |
| **logP100** | 2,4,5-tricloropiridina | 2,720 | 3,161 | 3,516 | 3,098 |
| **logP101** | 2,6-diclorofenol | 2,750 | 2,725 | 2,801 | 2,686 |
| **logP102** | 3-benziloxianilina | 2,772 | 3,739 | 4,097 | 3,501 |
| **logP103** | tolueno | 2,786 | 2,416 | 2,776 | 1,912 |
| **logP104** | 2,3,6-tricloropiridina | 2,810 | 3,594 | 3,903 | 3,340 |
| **logP105** | 4-butilpiridina | 2,810 | 3,480 | 3,372 | 2,949 |
| **logP106** | 1-metileptilamina | 2,819 | 3,540 | 2,748 | 3,153 |
| **logP107** | 2,3-dimetilnitrobenzeno | 2,830 | 2,779 | 2,619 | 2,556 |
| **logP108** | tetraclorometano | 2,830 | 2,798 | 2,829 | 2,285 |
| **logP109** | 1-naftol | 2,840 | 1,860 | 2,375 | 2,350 |
| **logP110** | 2,3-diclorofenol | 2,840 | 2,380 | 2,565 | 2,548 |
| **logP111** | clorobenzeno | 2,898 | 2,167 | 2,578 | 2,827 |
| **logP112** | 2,5-dicloronitrobenzeno | 2,900 | 3,059 | 3,310 | 2,913 |
| **logP113** | 3,5-dicloroanilina | 2,900 | 2,903 | 3,235 | 2,924 |
| **logP114** | 2,4-dicloroanilina | 2,910 | 2,862 | — | 2,844 |
| **logP115** | 3,4-dimetilnitrobenzeno | 2,910 | 2,439 | 2,766 | 2,787 |
| **logP116** | 4-butilanilina | 2,913 | 2,010 | 2,166 | 2,041 |

Continua na próxima página

Tabela 2.1: Continuação

| Código | Nome | Exp. | LO1 | LO2 | LO3 |
|---|---|---|---|---|---|
| **logP117** | 2,5-dicloroanilina | 2,920 | 2,597 | 3,000 | 2,580 |
| **logP118** | 2,3,6-trimetilfenol | 2,922 | 2,511 | 2,438 | 2,796 |
| **logP119** | 2,3-dicloronitrobenzeno | 3,050 | 2,880 | 2,988 | 2,766 |
| **logP120** | 4-cloro-2-nitrotolueno | 3,050 | 2,914 | 2,925 | 2,892 |
| **logP121** | 2,4-diclorofenol | 3,060 | 2,230 | 2,613 | 2,379 |
| **logP122** | 2,5-diclorofenol | 3,060 | 2,006 | 2,371 | 2,036 |
| **logP123** | 2,4-dicloro-6-nitrofenol | 3,066 | 2,883 | 2,841 | 2,635 |
| **logP124** | 2,4-dicloronitrobenzeno | 3,090 | 3,111 | 3,158 | 2,865 |
| **logP125** | 2-cloro-6-nitrotolueno | 3,090 | 2,480 | 3,108 | 2,800 |
| **logP126** | 2-fenilfenol | 3,090 | 2,515 | 3,333 | 3,156 |
| **logP127** | 4-cloro-3-metilfenol | 3,100 | 2,106 | 2,125 | 2,340 |
| **logP128** | *o*-xileno | 3,120 | 2,375 | 2,528 | 2,411 |
| **logP129** | 3,5-dicloronitrobenzeno | 3,130 | 3,017 | 3,088 | 3,034 |
| **logP130** | *p*-xileno | 3,150 | 2,391 | 2,750 | 2,549 |
| **logP131** | 4-propilfenol | 3,200 | 2,713 | 2,508 | 2,769 |
| **logP132** | *m*-xileno | 3,200 | 2,402 | 2,716 | 1,698 |
| **logP133** | 2,3,4,5-tetracloropiridina | 3,250 | 4,081 | 4,054 | 4,196 |
| **logP134** | 4,5-dicloro-2-metoxifenol | 3,260 | 3,466 | 3,552 | 3,830 |
| **logP135** | 3-clorotolueno | 3,280 | 2,912 | 3,178 | 3,181 |
| **logP136** | 4-*t*-butilfenol | 3,310 | 2,728 | 2,484 | 2,923 |
| **logP137** | 2,3,6-tricloroanilina | 3,323 | 3,567 | 3,809 | 3,391 |
| **logP138** | 3,4-diclorofenol | 3,330 | 2,307 | 2,495 | 2,588 |
| **logP139** | 4-clorotolueno | 3,330 | 2,840 | 3,222 | 3,178 |
| **logP140** | 4-fenoxifenol | 3,350 | 3,376 | 3,728 | 3,515 |
| **logP141** | tetracloroeteno | 3,400 | 3,127 | 3,100 | 2,959 |
| **logP142** | 1,2-diclorobenzeno | 3,433 | 3,061 | 3,235 | 3,724 |
| **logP143** | 4-cloro-2,3-dimetilfenol | 3,433 | 2,766 | 2,800 | 3,012 |
| **logP144** | 2,3,5,6-tetracloropiridina | 3,440 | 4,287 | 4,474 | 4,068 |
| **logP145** | 1,4-diclorobenzeno | 3,444 | 2,583 | 3,214 | 3,238 |
| **logP146** | 4-cloro-3,5-dimetilfenol | 3,483 | 3,111 | 2,685 | 3,004 |
| **logP147** | 1,3-diclorobenzeno | 3,525 | 3,031 | 3,429 | 3,669 |
| **logP148** | 1,3-diclorobenzeno | 3,525 | 3,032 | 3,333 | 3,669 |
| **logP149** | 4-cloro-2-alilfenol | 3,558 | 3,461 | 3,524 | 3,658 |
| **logP150** | 4-*n*-butilfenol | 3,561 | 3,501 | 3,047 | 3,586 |
| **logP151** | 4-*n*-butilfenol | 3,561 | 3,413 | 3,057 | 3,501 |
| **logP152** | 2,3,5-triclorofenol | 3,577 | 3,088 | 3,303 | 3,195 |
| **logP153** | 3,5-diclorofenol | 3,620 | 2,573 | 2,821 | 2,908 |
| **logP154** | pentacloroetano | 3,627 | 3,242 | 3,058 | 4,201 |
| **logP155** | 2,3,4-tricloroanilina | 3,680 | 3,607 | 3,758 | 3,374 |
| **logP156** | 2,4,5-tricloroanilina | 3,690 | 3,649 | 3,960 | 3,574 |

Continua na próxima página

Tabela 2.1: Continuação

| Código | Nome | Exp. | LO1 | LO2 | LO3 |
|---|---|---|---|---|---|
| **logP157** | 2,4,6-triclorofenol | 3,690 | 3,444 | 3,679 | 3,619 |
| **logP158** | 4-hexiloxianilina | 3,694 | 5,016 | 4,620 | 4,340 |
| **logP159** | 2,4,5-triclorofenol | 3,720 | 2,783 | 3,079 | 2,990 |
| **logP160** | 3,4,5-tricloro-2,6-dimetoxifenol | 3,740 | 3,460 | 3,844 | 3,937 |
| **logP161** | 2,3,6-triclorofenol | 3,770 | 3,076 | 3,370 | 3,300 |
| **logP162** | 3,4,5-tricloro-2-metoxifenol | 3,770 | 4,089 | 4,038 | 4,501 |
| **logP163** | 2-*t*-butil-4-metilfenol | 3,800 | 3,659 | 3,522 | 3,720 |
| **logP164** | 4-*t*-pentilfenol | 3,830 | 3,234 | 2,925 | 3,373 |
| **logP165** | 2,3,5,6-tetraclorofenol | 3,880 | 3,968 | 4,068 | 3,728 |
| **logP166** | 2,4,6-tribromofenol | 3,917 | 4,471 | 3,873 | — |
| **logP167** | 4-cloro-2-isopropil-5-metilfenol | 3,920 | 4,884 | 3,288 | 3,991 |
| **logP168** | 4-fenilazofenol | 3,957 | 3,949 | 4,333 | 4,477 |
| **logP169** | 2,3,4,5-tetracloroanilina | 4,040 | 4,371 | 4,442 | 4,062 |
| **logP170** | 1,2,4-triclorobenzeno | 4,050 | 3,700 | 3,999 | 4,250 |
| **logP171** | 3,4-diclorotolueno | 4,067 | 4,056 | 3,986 | 3,940 |
| **logP172** | 4-*n*-pentilfenol | 4,090 | 3,936 | 3,626 | 4,104 |
| **logP173** | 1,2,3-triclorobenzeno | 4,139 | 3,818 | 3,906 | 4,368 |
| **logP174** | 1,3,5-triclorobenzeno | 4,189 | 3,627 | 4,174 | 4,206 |
| **logP175** | 2,3,4,5-tetraclorofenol | 4,210 | 4,284 | 4,091 | 4,353 |
| **logP176** | 2,4-diclorotolueno | 4,240 | 4,013 | 4,101 | 3,966 |
| **logP177** | 3,4,5-triclorofenol | 4,280 | 3,274 | 3,280 | 3,429 |
| **logP178** | 3,4,5,6-tetracloro-2-hidroxifenol | 4,290 | 3,181 | 3,201 | 2,857 |
| **logP179** | 2,3,4,6-tetraclorofenol | 4,450 | 3,696 | 3,959 | 3,755 |
| **logP180** | 2,3,5,6-tetracloroanilina | 4,460 | 4,212 | 4,548 | 4,131 |
| **logP181** | 1,2,4,5-tetraclorobenzeno | 4,604 | 4,005 | 4,578 | 4,802 |
| **logP182** | 1,2,3,4-tetraclorobenzeno | 4,635 | 4,885 | 4,465 | 4,996 |
| **logP183** | 1,2,3,5-tetraclorobenzeno | 4,658 | 4,569 | 4,718 | 4,948 |
| **logP184** | 2,4,5-triclorotolueno | 4,780 | 4,332 | 4,691 | 4,553 |
| **logP185** | pentaclorobenzeno | 5,183 | 5,110 | 5,056 | 5,575 |
| **logP186** | 4-decilanilina | 6,087 | 6,308 | 5,654 | 6,172 |
| **logP187** | 4-nonilfenol | 6,206 | 6,558 | 5,965 | — |
| **logP188** | 4-dodecilanilina | 7,145 | 7,053 | — | — |

Final

## 2.4 Resultados e discussão

Foram testados modelos com propriedades calculadas utilizando HF/6-311G//3-21G com a cavidade S1 (M3) e com HF/3-21G utilizando as cavidades S1 (M1) e S2 (M2). Os resultados obtidos com a metodologia aqui proposta para o cálculo de $logP_{o/w}$ estão listados na Tabela 2.3. Sob a coluna LO1 estão mostrados os resultados obtidos com M1 para validação cruzada LO. O valor de $q^2 = 0,79$ indica uma boa capacidade de previsão deste modelo. O valor da estimativa do desvio padrão da previsão (*Standard Deviation of Prediction* ) $SDEP = 0,68$ unidades de $logP_{o/w}$, também mostra que este modelo é válido. Este valor é da mesma ordem de grandeza que o esperado em outros métodos publicados para estimativa do valor de $logP_{o/w}$ [18]. Também se aproxima da maior exatidão obtida com método experimental tipo *Shake Flask* [112], utilizado como um padrão analítico.

Os valores sob a coluna 'EV1' da tabela são os obtidos na previsão por validação cruzada em bloco. É usado um conjunto com cento e sessenta e nove compostos para treinamento e dezenove compostos para previsão ($\approx 10\%$), escolhidos aleatoriamente. Os valores dos índices de acerto obtidos para esta previsão são os melhores obtidos para uma série de testes de validação cruzada com blocos. Este é o motivo de apontarem um modelo com maior capacidade de previsão que a validação LO.

Os resultados obtidos em M2 por validação cruzada LO são mostrados na Tabela 2.3 sob a coluna 'LO2' (este modelo usa a cavidade S2 no nível de cálculo HF/3-21G). Esta tabela mostra um valor $q^2 = 0,85$ para M2 e de $SDEP = 0,56$ para cento e oitenta e seis compostos. Os valores obtidos para estimativa do erro de previsão na validação cruzada em bloco, 'EV2', também mostram um modelo mais robusto que M1. Para este modelo não foi possível realizar cálculos para os compostos **logP114** e **logP188** da Tabela 2.1, por causa de problemas com a obtenção das cavidades de sovente em acetona.

O modelo M2 foi obtido com PLS usando dez variáveis latentes. A acumulação de variância das trinta variáveis no bloco dos descritores é alta, como pode ser visto na terceira coluna da Tabela 2.2. O grande número de variáveis latentes necessárias para acumular uma variância expressiva, mostra que os descritores usados para modelagem de $logP_{o/w}$ contém grande quantidade de informação não correlacionada, que pôde ser utilizada pelo modelo PLS somente se utilizando um número alto de variáveis latentes. Um aumento do número de variáveis latentes não seria conveniente, porque toda a variância dos dados experimentais possível de ser modelada já foi acumulada pelo modelo PLS em dez variáveis latentes. Isto é mostrado na Tabela 2.2 pelas pequenos valores de variância acumulada nas últimas variáveis latentes das atividades, na quarta coluna de dados.

Os resultados obtidos em validação cruzada LO com M3 são mostrados na Tabela 2.3 sob a coluna 'LO3'. Este modelo usa a cavidade S1 no nível de cálculo HF/6-311G//3-21G. Para M3 não foram realizados os cálculos para cinco compostos: **logP47**, **logP79**, **logP166**, **logP187** e **logP188**. Cálculos com cavidade S2 não foram realizados por causa do aumento no tempo de processamento. Para cada modelo é colocado um traço no valor da atividade biológica dos compostos na Tabela 2.1 nos casos em que os cálculos quânticos para obtenção dos descritores não foram possíveis. Os maiores problemas deste modelo são: (1) a não disponibilidade da base HF/6-311G para elementos químicos presentes

Tabela 2.2: Variâncias acumuladas nas variáveis latentes utilizadas no modelo PLS, e que foram utilizadas para a construção do modelo M2 proposto para prever o valor de $logP_{o/w}$. Os *loadings* da PCA para o mesmo número de variáveis latentes foram utilizados na camada de entrada de um modelo com BPN.

| | PLS | | | | PCA | |
|---|---|---|---|---|---|---|
| | Descritores | | Atividades | | Descritores | |
| variável latente # | Esta | Total | Esta | Total | Esta | Total |
| 1 | 40,55 | 40,55 | 54,12 | 54,12 | 41,41 | 41,41 |
| 2 | 14,84 | 55,38 | 12,30 | 66,42 | 24,69 | 66,10 |
| 3 | 16,20 | 71,58 | 5,83 | 72,25 | 15,86 | 81,96 |
| 4 | 4,28 | 75,86 | 5,81 | 78,05 | 5,94 | 87,90 |
| 5 | 13,48 | 89,35 | 1,35 | 79,40 | 3,48 | 91,38 |
| 6 | 1,86 | 91,21 | 4,01 | 83,41 | 2,89 | 94,27 |
| 7 | 2,04 | 93,25 | 1,03 | 84,44 | 1,92 | 96,19 |
| 8 | 2,17 | 95,42 | 0,16 | 84,60 | 1,57 | 97,76 |
| 9 | 2,05 | 97,47 | 0,09 | 84,69 | 1,00 | 98,76 |
| 10 | 0,50 | 97,97 | 0,14 | 84,83 | 0,43 | 99,19 |

Tabela 2.3: Resultados obtidos para validação cruzada tipo 'fora um' (LO) e para conjunto de validação cruzada em bloco (EV) usando 10% do conjunto total.

| | Modelo M1 | | Modelo M2 | | Modelo M3 | |
|---|---|---|---|---|---|---|
| | LO1 | EV1 | LO2 | EV2 | LO3 | EV3 |
| PRESS | 84,8 | 1,77 | 58,4 | 1,62 | 49 | 2,22 |
| $q^2$ | 0,79 | 0,87 | 0,85 | 0,92 | 0,87 | 0,92 |
| SDEP | 0,68 | 0,10 | 0,56 | 0,10 | 0,52 | 0,11 |
| m | 188 | 19 | 186 | 19 | 183 | 19 |
| #VLs | 10 | 10 | 10 | 10 | 10 | 10 |
| #descritores | 15 | 15 | 30 | 30 | 15 | 15 |

na molécula e (2) a dificuldade de obtenção das cavidades de solvente em acetona e/ou cicloexano.

O modelo M3 foi o que permitiu os melhores resultados na previsão de compostos em validação cruzada LO, com $q^2 = 0,87$. Comparando o aumento da capacidade de previsão de M3 com os correspondentes de M2 e M1 na Tabela 2.3 pode-se ter uma idéia da variação na capacidade de previsão em função da base Hartree–Fock e do tipo de cavidade de solvente utilizado. A maior sofisticação na aplicação do efeito do solvente pode ser relacionada a uma melhora significativa no valor de $q^2$ para os dois tipos de validação. O aumento da base também permite um aumento na capacidade de previsão da mesma ordem. O aumento da base de HF/3-21G para HF/6-311G, com cavidade S1, corresponde ao aumento por um fator de 1,7 do tempo de processamento do composto **logP180** da Tabela 2.1. A utilização da cavidade S2 torna os cálculos 4,5 vezes mais lentos que com cavidade S1, mantendo a mesma base HF/3-21G.

O modelo considerado mais indicado para uso geral foi M2, pela sua maior disponibilidade em relação ao conjunto de base, e pelos resultados obtidos. Os valores obtidos em validação cruzada tipo LO para $logP_{o/w}$ com M2, em gráfico contra os valores publicados na literatura são mostrados na Figura 2.2. Nas Tabelas 2.4 e 2.5 podemos ver listados os valores dos coeficientes de regressão do modelo M2, para cada variável latente e a soma total, que é o vetor de regressão linear resultante do modelo, conforme as Equações 1.86 de página 42. Observando a figura podemos concluir que os valores de previsão para compostos com baixo valor no coeficiente de partição são piores previstos. Na Figura 2.3 são mostrados os erros de previsão (*studentizados*) de cada composto em validação cruzada tipo LO em gráfico contra os valores de *leverage*. Os sete compostos com erro maior que dois desvios padrão são distribuídos aleatoriamente. Estes compostos são os numerados como **logP5**, **logP7**, **logP14**, **logP18**, **logP68**, **logP78** e **logP102** na Tabela 2.1 e nas Figuras 2.2 e 2.3.

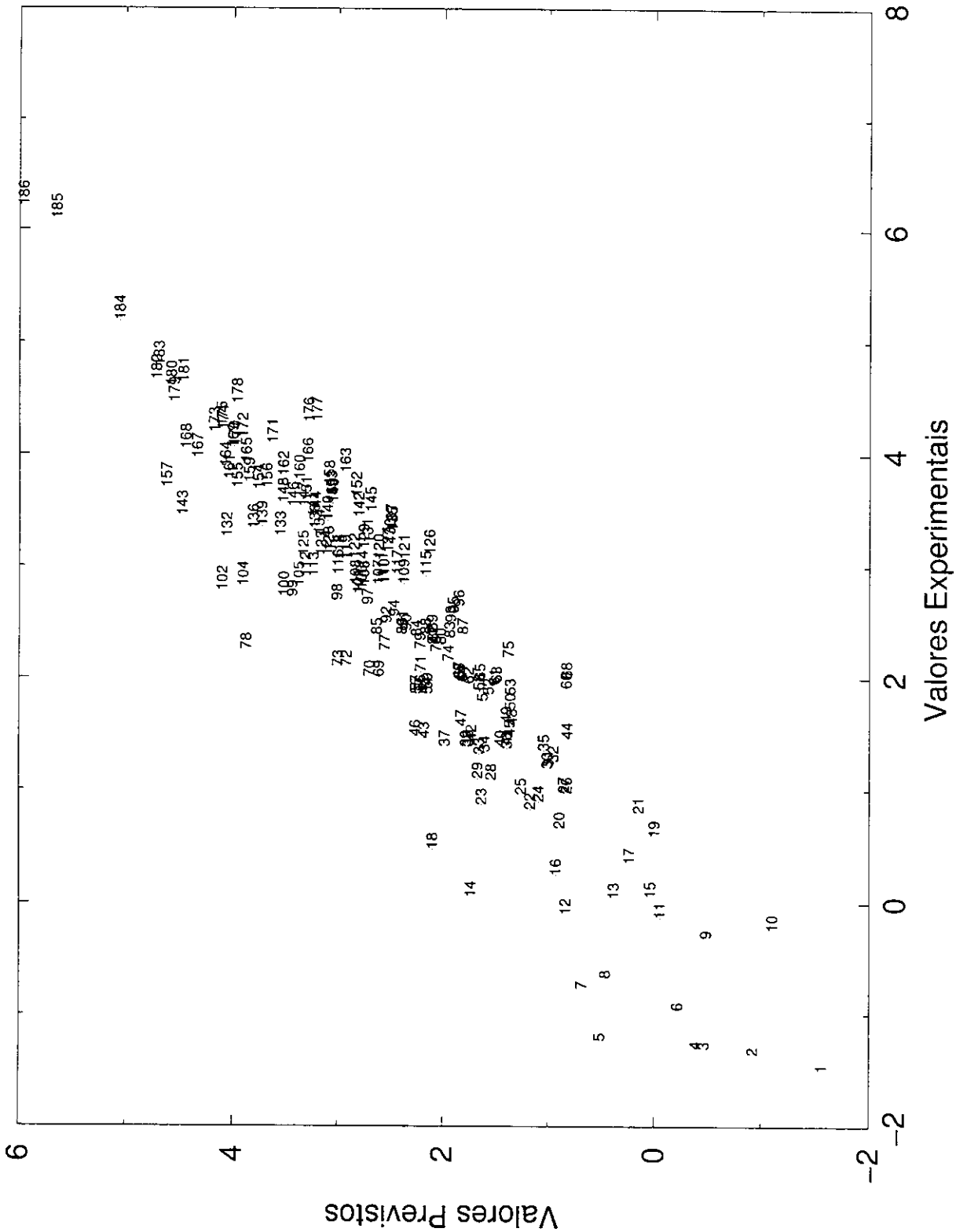
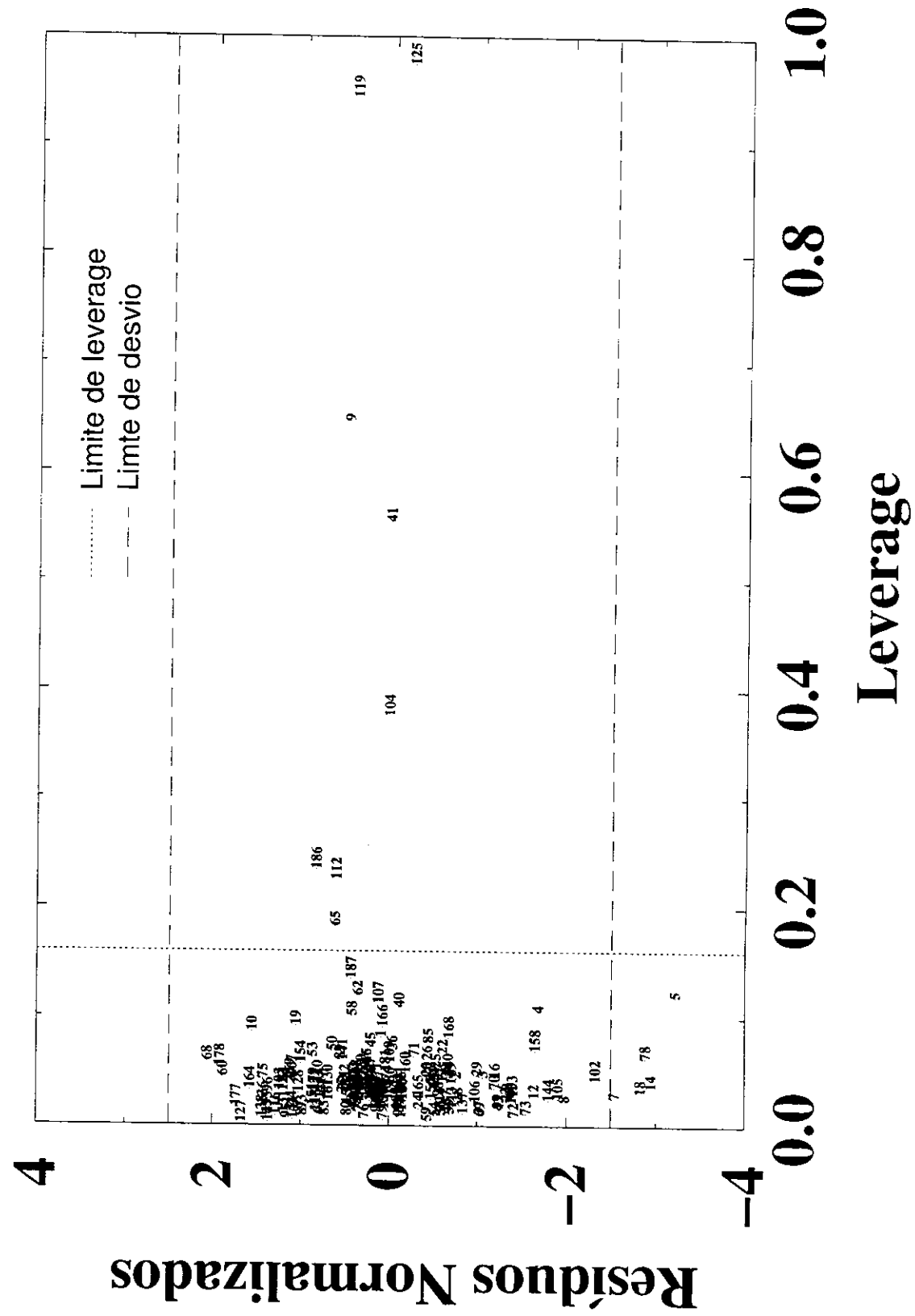



Figura 2.2: Valores experimentais de $logP_{o/w}$ versus valores calculados com o modelo M2 da Tabela 2.3 para para os compostos listados na Tabela 2.1.




A

Figura 2.3: Resíduos em unidades de variância e *leverage*, obtidos com modelo M2 da Tabela 2.3 para os compostos listados na Tabela 2.1.

Tabela 2.4: Coeficientes de regressão obtidos das quinze primeiras variáveis descritoras calculadas para todos os modelos em cada uma das dez variáveis latentes utilizadas no modelo M2, que resulta os coeficientes que podem ser aplicados em uma equação de regressão com os valores autoescalados das variáveis descritoras.

| Variável | Variável Descritora [a] | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Latente | $logP_{aq-ac}$ | $logP_{aq-ch}$ | $logP_{ac-ch}$ | $\mu_{aq}$ | $\mu_{ac}$ | $\mu_{ch}$ | $VC_{aq}$ | $VC_{ac}$ | $VC_{ch}$ | $Q^-_{aq}$ | $Q^-_{ac}$ | $Q^-_{ch}$ | $Q^{H^+}_{aq}$ | $Q^{H^+}_{ac}$ | $Q^{H^+}_{ch}$ |
| #1 | 0,0624 | 0,0644 | 0,0274 | 0,0005 | 0,0004 | -0,0005 | 0,0642 | 0,0643 | 0,0643 | 0,0098 | 0,0028 | 0,0069 | -0,0083 | -0,0026 | -0,0059 |
| #2 | 0,0184 | 0,0598 | 0,0852 | 0,0464 | 0,0464 | 0,0425 | 0,0354 | 0,0358 | 0,0358 | -0,0011 | -0,0119 | -0,0007 | 0,0043 | 0,0130 | 0,0044 |
| #3 | 0,0117 | 0,0175 | 0,0155 | 0,0577 | 0,0564 | 0,0534 | 0,0404 | 0,0408 | 0,0408 | -0,0014 | -0,0070 | -0,0010 | 0,0046 | 0,0088 | 0,0057 |
| #4 | 0,0251 | 0,0171 | -0,0055 | -0,0526 | -0,0606 | -0,0658 | 0,1167 | 0,1185 | 0,1185 | -0,0268 | 0,0220 | -0,0240 | 0,0364 | -0,0082 | 0,0518 |
| #5 | 0,0065 | 0,0076 | 0,0045 | -0,0109 | -0,0117 | -0,0125 | 0,0337 | 0,0343 | 0,0343 | 0,0213 | 0,0233 | 0,0226 | -0,0171 | -0,0197 | -0,0139 |
| #6 | 0,0184 | 0,0584 | 0,0828 | -0,0270 | -0,0184 | -0,0202 | 0,1308 | 0,1340 | 0,1340 | 0,0225 | -0,0320 | 0,0462 | 0,0100 | 0,0498 | 0,0036 |
| #7 | -0,0201 | 0,0136 | 0,0561 | 0,0058 | 0,0007 | -0,0023 | 0,0212 | 0,0221 | 0,0221 | -0,0004 | 0,0303 | -0,0009 | 0,0128 | -0,0189 | 0,0244 |
| #8 | -0,0005 | 0,0028 | 0,0061 | -0,0057 | -0,0021 | -0,0090 | -0,0082 | -0,0083 | -0,0083 | 0,0014 | 0,0046 | 0,0216 | 0,0105 | 0,0038 | -0,0014 |
| #9 | 0,0095 | 0,0026 | -0,0094 | 0,0038 | 0,0084 | 0,0028 | -0,0071 | -0,0074 | -0,0074 | -0,0030 | 0,0189 | 0,0076 | 0,0153 | -0,0088 | 0,0140 |
| #10 | 0,0190 | -0,0070 | -0,0420 | 0,0199 | 0,0257 | 0,0068 | -0,0043 | -0,0060 | -0,0060 | 0,0313 | 0,0000 | 0,0357 | 0,0198 | 0,0262 | 0,0294 |
| m [b] | 0,1504 | 0,2369 | 0,2207 | 0,0378 | 0,0451 | -0,0048 | 0,4227 | 0,4280 | 0,4280 | 0,0535 | 0,0511 | 0,1140 | 0,0884 | 0,0434 | 0,1122 |

[a] As variáveis descritoras $logP_{xx-yy}$ referem-se aos coeficientes de partição calculados pelas expressões das Equações 2.3, para cada par de solventes definidos pela seguinte abreviação: *aq* – água; *ac* – acetona; *ch* – cicloexano. A variávies descritoras $\mu_{xx}$ são os momenta de dipolo calculados em cada solvente, de acordo com a mesma abreviação. As variáveis $VC_{xx}$ são os volumes das cavidades formadas no *continuum* para a introdução do soluto, em cada solvente. As variáveis $Q^-_{xx}$ são as cargas líquidas calculadas sobre o átomo mais negativamente carregado da molécula, em cada tipo de solvente. As variáveis $Q^{H^+}_{xx}$ são as cargas líquidas calculadas sobre o hidrogênio mais positivamente carregado, ou outro átomo nas moléculas sem hidrogênio, em cada solvente.

[b] Os valores listados nesta linha correspondem aos coeficientes $C_a$ que devem ser aplicados aos $n$ valores autoescalado das variáveis descritoras $D_a$ em uma equação de regressão $\sum_{a=1}^{n} C_a D_a$, conforme definidos nas Equações 1.86 de página 42.

Tabela 2.5: Coeficientes de regressão obtidos para as quinze últimas variáveis descritoras calculadas para todos os modelos em cada uma das dez variáveis latentes utilizadas no modelo M2, que resulta os coeficientes que podem ser aplicados em uma equação de regressão com os valores autoescalados das variáveis descritoras.

| Variável Latente | Variável Descritora [a] | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Delta E^{i}_{aq}$ | $\Delta E^{i}_{ac}$ | $\Delta E^{i}_{ch}$ | $W_{MS^{aq}}$ | $W_{MS^{ac}}$ | $W_{MS^{ch}}$ | $E^{cav}_{aq}$ | $E^{cav}_{ac}$ | $E^{cav}_{ch}$ | $\Delta G^{\circ}_{D^{aq}}$ | $\Delta G^{\circ}_{D^{ac}}$ | $\Delta G^{\circ}_{D^{ch}}$ | $\Delta G^{\circ}_{R^{aq}}$ | $\Delta G^{\circ}_{R^{ac}}$ | $\Delta G^{\circ}_{R^{ch}}$ |
| #1 | -0,0027 | 0,0145 | 0,0517 | -0,0387 | 0,0640 | -0,0071 | 0,0180 | 0,0486 | -0,0387 | 0,0645 | -0,0082 | 0,0189 | 0,0480 | -0,0387 | 0,0640 |
| #2 | 0,0088 | 0,0578 | -0,0288 | 0,0599 | 0,0282 | 0,0045 | 0,0714 | -0,0417 | 0,0588 | 0,0308 | 0,0060 | 0,0717 | -0,0439 | 0,0583 | 0,0284 |
| #3 | 0,0393 | -0,0083 | -0,0161 | 0,0316 | 0,0134 | 0,0334 | 0,0020 | -0,0273 | 0,0302 | 0,0163 | 0,0486 | 0,0024 | -0,0292 | 0,0295 | 0,0150 |
| #4 | 0,0732 | -0,0509 | -0,0723 | 0,0705 | 0,0159 | -0,0100 | -0,0272 | -0,1078 | 0,0636 | 0,0326 | 0,0591 | -0,0315 | -0,1139 | 0,0606 | 0,0313 |
| #5 | 0,0045 | -0,0087 | -0,0268 | 0,0103 | -0,0064 | -0,0102 | -0,0080 | -0,0380 | 0,0077 | -0,0001 | 0,0173 | -0,0131 | -0,0400 | 0,0066 | -0,0003 |
| #6 | -0,0792 | 0,0054 | -0,1264 | -0,0578 | -0,0860 | -0,0234 | -0,0301 | -0,1738 | -0,0723 | -0,0522 | 0,0619 | -0,0683 | -0,1818 | -0,0784 | -0,0497 |
| #7 | 0,0562 | -0,0189 | -0,0478 | -0,0502 | -0,0212 | -0,0238 | 0,0135 | -0,0616 | -0,0556 | -0,0079 | 0,0090 | 0,0093 | -0,0639 | -0,0577 | -0,0058 |
| #8 | 0,0037 | -0,0042 | -0,0178 | -0,0126 | 0,0134 | 0,0324 | -0,0066 | -0,0225 | -0,0137 | 0,0179 | 0,0026 | 0,0085 | -0,0233 | -0,0141 | 0,0184 |
| #9 | -0,0179 | -0,0006 | -0,0134 | -0,0081 | 0,0123 | 0,0062 | -0,0103 | -0,0188 | -0,0086 | 0,0159 | -0,0033 | 0,0044 | -0,0197 | -0,0087 | 0,0148 |
| #10 | -0,0072 | -0,0329 | -0,0176 | -0,0114 | 0,0237 | -0,0188 | -0,0236 | -0,0391 | -0,0112 | 0,0291 | -0,0736 | 0,0267 | -0,0429 | -0,0107 | 0,0204 |
| m [b] | 0,0786 | -0,0467 | -0,3152 | -0,0064 | 0,0574 | -0,0168 | -0,0009 | -0,4821 | -0,0398 | 0,1470 | 0,1195 | 0,0289 | -0,5106 | -0,0534 | 0,1364 |

[a] As variáveis descritoras $\Delta E^{i}_{xx}$ referem-se às variações da energia interna nos solventes, para cada solvente definido pela seguinte abreviação: $aq$ – água; $ac$ – acetona; $ch$ – cicloexano. A variávies descritoras $W_{MS^{xx}}$ são as interações eletrostáticas do soluto com o solvente, de acordo com a mesma abreviação. As variáveis $E^{cav}_{xx}$ são as energias de cavitação calculadas pelo método de Pierotti, em cada solvente. As variáveis $\Delta G^{\circ}_{D^{xx}}$ são as energias livre de dispersão, em cada tipo de solvente. As variáveis $\Delta G^{\circ}_{R^{xx}}$ são as energias livre de repulsão, em cada solvente.

[b] Os valores listados nesta linha correspondem aos coeficientes $C_a$ que devem ser aplicados aos $n$ valores autoescalado das variáveis descritoras $D_a$ em uma equação de regressão $\sum_{a=1}^{n} C_a D_a$, conforme as Equações 1.86 da página 1.86.

Comparando a estrutura de compostos mal previstos (**logP5**, **logP7**, **logP14**, **logP18** e **logP68** ) vemos que têm em comum o fato de serem todos funcionalizados com hidroxila. São bastante eficientes como doadores de pontes de hidrogênio, e este fator pode ser o responsável pela má previsão de sua atividade. Os modelos utilizados para simulação de efeitos de solvente não têm como representar perfeitamente a formação de pontes. Uma simulação explícita do solvente seria necessária para este tipo de representação, com grande aumento do custo computacional. Os compostos **logP78** e **logP102** têm grupo benzilóxi ligados à piridina e anilina. Estes compostos tiveram o valor de $logP_{o/w}$ superestimado, provavelmente por causa grande do volume das cavidades de solvente necessárias para acomodá-los, que é uma característica associada a compostos mais lipofílicos de maneira geral.

Os resultados do modelo obtido com método de treinamento de BPN com os *loadings* da PCA mostrados na Tabela 2.2, indicam um modelo não válido. Há uma grande capacidade de ajuste de valores por este método, porém as redes treinadas não são capazes de gerar estimativas razoáveis para os valores de $logP_{o/w}$ dos compostos em testes de validação cruzada. O valor baixo de $q^2 = 0,55$ obtido com a rede, com $SDEP = 1,09$ e $PRESS = 170$. Além dos resultados serem estatisticamente ruins, o treinamento de uma rede neural como esta é muito demorado, se comparado ao necessário para obter um modelo análogo com PLS. O tempo é da ordem de 24h para BPN, em um computador Pentium II 350Mhz, contra alguns minutos utilizando modelos tipo PLS.

## 2.5 Conclusões

Este trabalho mostra como podem ser calculados valores de $logP_{o/w}$ para compostos orgânicos simples, de tamanho moderado, sem aplicação de parâmetros arbitrários. A aplicação de alguns poucos parâmetros necessários para simulação do efeito de solvatação mais sofisticado (nas cavidade do S2), usa dados experimentais ou calculados com HF/6-311G. Portanto não podem ser considerados arbitrários. Este tipo de simulação de efeito de solvatação (cavidade S2) aumenta bastante o tempo de processamento.

O modelo M2 foi considerado o mais promissor, pelos resultados obtidos e pela aplicação possível para um número maior de casos. Entretanto, utilização da cavidade S2 com base HF/3-21G é justificada somente se há necessidade de utilização de elementos químicos que não têm base HF/6-311G definida. Se não há elementos com base não definida para HF/6-311G este é o modelo mais indicado. A utilização de bases grandes, com cavidade S2, provavelmente levaria a resultados melhores, com custo computacional alto.

As piores previsões são para os compostos com capacidade de doar elétrons para pontes de hidrogênio, e de compostos com grande volume que são hidrofílicos. Portanto este método tem melhor aplicação aos compostos sem estas características. Estas limitações devem-se a má estimativa dos efeitos de ligação por pontes de hidrogênio, e pela grande importância do volume da cavidade do solvente nas regressões.

# Capítulo 3

# Compostos antiúlcera das séries dos guanidinotiazóis

## 3.1 A inibição da secreção ácida e o controle da doença gástrica

Atualmente podemos considerar a existência de dois grandes grupos de úlcera péptica: Úlcera relacionada á infecção por *H. pylori* e úlcera associada ao uso de Drogas Anti-Inflamatória Não Esteróides (NSAIDs). Em ambos os casos os agentes anti-secreção gástrica têm papel muito importante no tratamento da úlcera. Há também um tipo de desordem gástrica denominada Doença de Refluxo Gastro-Esofágico (GERD) que está tornando-se comum, e que não está relacionada nem à contaminação por *H. pylori* nem ao uso de NSAIDs, e sim ao extravasamento de suco gástrico para a região do esôfago. Também neste caso a redução da secreção ácida diminui o desconforto do paciente [113].

A proteína $H^+$,$K^+$-ATPase, a bomba de prótons responsável pela secreção ácida no estômago, está localizada nas vesículas da membrana gástrica e promove a troca iônica entre $H^+$ intracelular e $K^+$ extracelular acoplada com a hidrólise de ATP no citoplasma. Portanto esta proteína é um alvo farmacológico para a inibição da secreção ácida no estômago. A Figura 3.1 mostra um esquema dos principais caminhos metabólicos envolvidos na secreção ácida do estômago. Drogas como o Omeprazol, um composto da classe dos benzimidazóis substituídos, têm sido utilizadas no controle da acidez gástrica pela via metabólica da inibição da enzima $H^+$,$K^+$-ATPase [114]. Na Figura 3.1 a ligação entre uma droga e um sítio de ação é indicada pela letra X.

No entanto drogas que inibem irreversivelmente a ATPase (Omeprazol, Lanzoprazol) exigem nova síntese protéica para retomar a secreção ácida. Também é sabido que drogas desta classe são capazes de induzir o aumento da contagem de células tumorais em certos tipos de cobaias (ainda que este resultado não seja verificado em humanos) [115]. O risco relativo do desenvolvimento de desordens vasculares no olhos para usuários de Omeprazol comparado com não usuários é quase duas vezes maior. O uso de outras drogas anti-úlcera desta classe foi associado com riscos similares de desordem vascular [116].

Drogas capazes de inibir reversivelmente a $H^+$,$K^+$-ATPase poderiam permitir um

maior controle sobre a duração da inibição da secreção [117]. Uma possibilidade explorada é a modulação da secreção por meio de antagonistas do receptor H2 de histamina, que também participa de um dos primeiros passos do ciclo metabólico de secreção ácida nas células parietais do estômago. Os antagonistas do sítio H2 de histamina bloqueiam este caminho, relacionado ao aumento da secreção ácida, reduzindo a estimulação da $H^+$,$K^+$-ATPase por esta via e a secreção ácida. Diversos compostos, como a Ranitidina, têm sido utilizados para este fim.

A supressão da secreção ácida obtida com os antagonistas do receptor H2, todavia, não é suficiente para um controle efetivo das disfunções gástricas, durante a cura de feridas ulcerativas ou para o alívio dos sintomas de refluxo. Especialmente no caso da úlcera, estes antagonistas não são eficazes para controlar a estimulação da secreção ácida causada pela alimentação. Além disso, o rápido surgimento de tolerância às drogas antagonistas do sítio H2, aliada ao processo de hipersecreção de rebote que ocorre na abstinência, limitam ainda mais sua aplicação no tratamento clínico [113].

Os compostos da série dos guanidinotiazóis têm ação na secreção ácida no estômago através da inibição reversível da enzima $H^+$,$K^+$-ATPase. O seu mecanismo de ação, no entanto, é bem pouco conhecido. Estes compostos já são utilizados como inibidores do receptor H2 de histamina [118].

Os compostos da classe dos guanidinotiazóis estão sendo desenvolvidos como inibidores reversíveis da $H^+$,$K^+$-ATPase, havendo compostos com potência comparável à do Omeprazol. Esta nova geração de drogas para o combate da secreção gástrica pode ser capaz de reunir a alta inibição da secreção gástrica, oferecida pelos benzimidazóis, com as vantagens da inibição reversível oferecida por outras classes de medicamentos.

Os inibidores do tipo reversível, da classe das imidazopiridinas (SCH28080), e os do tipo irreversível, da classe dos benzimidazóis (Omeprazol), têm sua forma ativa como um cátion permanente, que pode ser uma sulfenamida nos benzimidazóis ou a forma quaternária protonada do nitrogênio nas imidazopiridinas. Estas classes de moléculas ligam-se à enzima na face luminal, e têm moléculas com dimensão da ordem de 12 Å a 16 Å. Estudos mostram que que a distância entre os resíduos de um sítio putativo de ligação de drogas do tipo bezimidazol a $H^+$,$K^+$-ATPase é da ordem de 15 Å na conformação ativada da enzima [119–122].

## 3.2 Objetivos

- Obter modelos de regressão entre os valores das atividades biológicas medidas experimentalmente e os conjuntos de descritores calculados com métodos mecânico-quânticos.

- Estudar as possibilidades de criação de modelos QSAR utilizando exclusivamente descritores de origem quântica.

- Avaliar a qualidade da descrição dos sistemas químicos com os métodos quânticos utilizados.

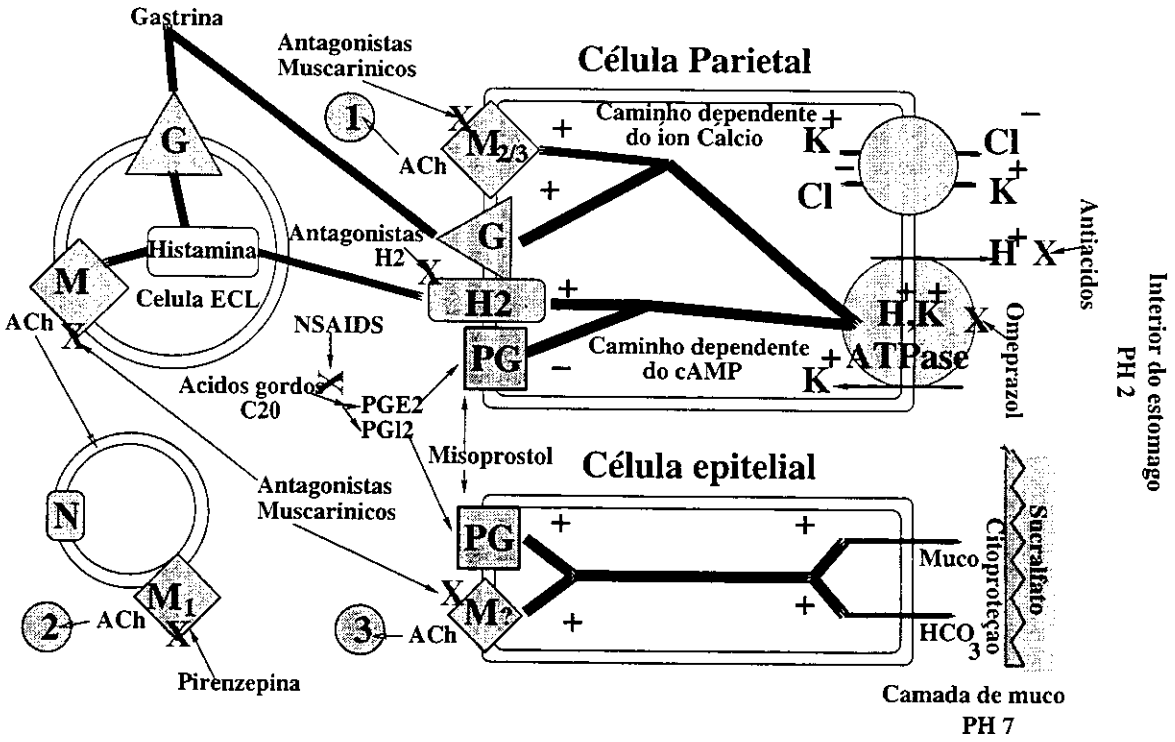



Figura 3.1: Diagrama esquemático das interações entre as células endócrinas que secretam histamina (células enterocromafínicas – ECL), as células secretoras de ácido (células parietais) e as células que secretam os fatores citoprotetores como o muco e o bicarbonato (células epiteliais). Os caminhos fisiológicos são mostrados pelas linhas mais grossas e podem ser estimulados (+) ou inibidos (-). Agonistas fisiológicos agem nos receptores transmembrana dos tipos Muscarínico (M) e nicotínico (N) para acetilcolina (ACh); receptores de Gastrina (G), receptores H2 de histamina e receptores de Prostaglandinas (PG). A ação das drogas mais utilizadas é indicada pela letra X no ponto de antagonismo. As drogas bezimidazólicas agem diretamente na face luminal (a parede interna do estômago) inibindo a $H^+$,$K^+$-ATPase. Os compostos como a Ranitidina bloqueiam o caminho dependente do cAMP cíclico, ligando-se ao receptor H2 de histamina. NSAIDS são os anti-inflamatórios não esteróides (aspirina, diclofenaco) [123, 124], que são ulcerogênicos porquê interferem na síntese de prostaglandinas a partir de ácidos-graxos [125], reduzindo a formação da camada de proteção de estômago pelas células epiteliais. Os antagonistas muscarínicos têm ação diminuidora da secreção ácida pelo caminho dependente do íon cálcio na células parietais e pela modulação da secreção de histamina nas células ECL, mas também reduzem a secreção de fatores de citoproteção (muco e carbonato) nas células epiteliais. O composto denominado pirenzepina é utilizado no combate à secreção ácida por este caminho [126, 127]. Os números '1' e '3' dentro dos círculos indicam possíveis entradas para para fibras prostaganglionares colinérgicas. O círculo com o número '2' indica a entrada para estímulo neural do nervo vago [128].

- Relacionar os descritores mecânico-quânticos com fragmentos das moléculas.
- Sugerir modificações baseadas na análise do conjunto descritor.

## 3.3 A sub-classe dos fenil-guanidinotiazóis

### 3.3.1 Metodologias de trabalho e procedimentos adotados

#### Busca conformacional por método sistemático

Para os fenil-guanidinotiazóis mostrados na Tabela 3.1 foram realizados cálculos no MOPAC com método semi-empírico AM1 para determinar a conformação de menor energia. Este método foi escolhido para busca conformacional pelo baixo custo computacional, pela razoável reprodução das geometrias moleculares e pela boa reprodução da escala de basicidade de hidrogênios [58].

Foram variados os ângulos de diedro de ligações distintas nos compostos **fgt8** e **fgt9** da Tabela 3.1, que foram escolhidos por terem substituintes representativos do grupo, em posições que pudessem criar impedimento estérico e influenciar a geometria de menor energia. O menor calor de formação para as conformações garante que estão presentes em muito maior proporção entre as conformações possíveis em fase gasosa.

Se consideramos o equilíbrio entre as conformações de menor e maior energia

$$C(E_1) \rightleftharpoons C(E_2)$$

pode-se verificar pelas Equações 3.1 qual a proporção de cada conformação em fase gasosa. A equação de Boltzman na sua forma não normalizada foi escolhida porque a utilização da equação normalizada requer o conhecimento de todas as conformações dentro de uma faixa de energia, por exemplo com $\Delta E \leq 2,5\,\mathrm{kcal.mol^{-1}}$ em relação ao mínimo de energia. Utilizamos a forma não normalizada porque desconhecemos todas as conformações que atendem este requisito para $\Delta E$. A faixa de $2,5\,\mathrm{kcal.mol^{-1}}$ corresponde a cerca de 99% de toda a população calculada com a Equação 3.1 à uma temperatura de 298 K.

$$\begin{aligned} P^{1,2} &= \exp\left(-\frac{E_1 - E_2}{RT}\right) \\ P^{1,2} &= \frac{N_{E_1}}{N_{E_2}} \end{aligned} \tag{3.1}$$

No caso do composto número **fgt9** , mostrado na Figura 3.2, o ângulo diedro $\Phi 1$ foi estudado sistematicamente com método AM1. O mínimo global tem calor de formação calculado para a estrutura de $115,17\,\mathrm{kcal.mol^{-1}}$, com um ângulo diedro de $-97,3°$ para a ligação $\Phi 1$. Em relação ao segundo mínimo local (calor de formação $121,77\,\mathrm{kcal.mol^{-1}}$ para ângulo diedro de $60,2°$) pode-se estabelecer uma população de 99,9% na conformação de menor energia, com as Equações 3.1. A conformação de menor energia assim escolhida foi considerada adequada para construção de moléculas da série com o grupo R1 benzila nas etapas posteriores de cálculo. O composto **fgt8** mostrado na Figura 3.3 teve o ângulo diedro $\Phi 2$ variado em passos de 5°. O ângulo diedro da ligação $\Phi 3$ foi fixado em $-9,35°$, para minimizar o impedimento estérico sobre os átomos da ligação $\Phi 2$. O mínimo de

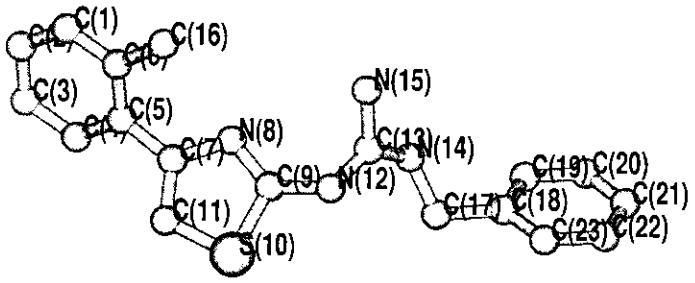



Figura 3.2: Parte do composto **fgt9** da Tabela 3.1 (os átomos de hidrogênio não estão mostrados). O diedro C(13)–N(14)–C(17)–C(18) corresponde à ligação $\Phi 1$, entre o grupo fenil e a metila no substituinte R1, em seu ponto de menor energia conformacional.

energia rotacional encontrado para a ligação $\Phi 2$ foi 89, 93 kcal.mol$^{-1}$ com ângulo diedro de $-175, 17°$. A outro mínimo de energia conformacional resulta em $\Delta H_f$ =92, 0 kcal.mol$^{-1}$ com $\Phi 2 = -95, 17°$. O percentual de moléculas na conformação mais estável calculado com a Equação 3.1 é de 96,97%. A conformação de menor energia foi considerada como base para construção para moléculas da série em que o grupo R1 é hidrogênio.

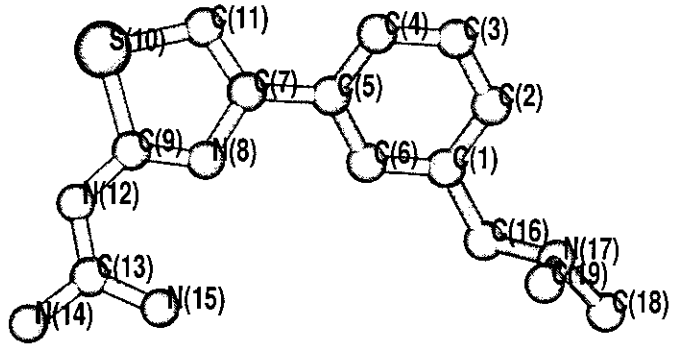



Figura 3.3: Parte do composto **fgt8** da Tabela 3.1 (os átomos de hidrogênio não estão mostrados). O diedro C(19)–N(17)–C(16)–C(1) corresponde à ligação $\Phi 2$ entre o grupo amina e a metila (substituinte R3) ligada ao anel fenílico, no ponto de menor energia conformacional. O diedro N(15)–C(13)–N(12)–C(9), denominado $\Phi 3$, foi fixado na sua posição de menor energia conformacional.

As geometrias propostas para os fenil-guanidinotiazóis foram novamente otimizadas com método *ab initio* no nível Hartree–Fock usando base 3-21G. Para todos os compostos foram realizados cálculos de propriedades físico-químicas com método *ab initio* combinando as bases (6-311G//3-21G).

Também foram calculadas propriedades usando simulação de solvatação com o Método do *Continuum* Polarizável (D-PCM) usando água como solvente, (6-311G//3-21G)AQ. As geometrias utilizadas para simulação do efeito de solvatação foram as mesmas utilizadas para cálculos no vácuo, porque a otimização de geometria com efeito de solvatação não se mostrou viável computacionalmente.

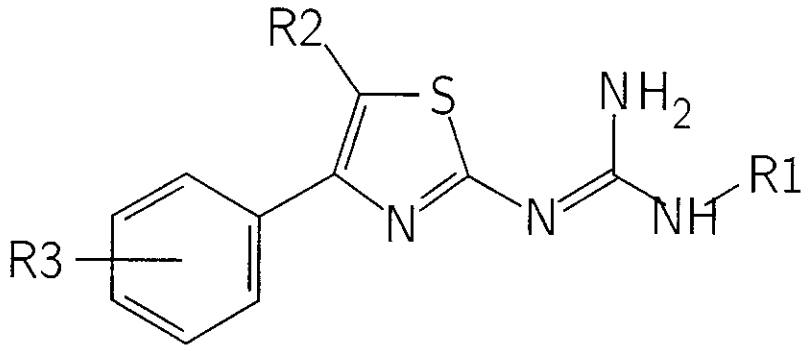



Figura 3.4: Esqueleto principal dos compostos da série dos fenil-guanidinotiazóis. A série toda pode ser obtida com as substituições da Tabela 3.1 nas posições R1, R2 e R3.

### Propriedades calculadas para os fenil-guanidinotiazóis

Para esta série de compostos foram obtidas propriedades calculadas com método *ab initio* para vinte e quatro compostos. Foram calculadas cinqüenta e seis propriedades para a conformação de menor energia de cada composto com (6-311G//3-21G) no vácuo. As propriedades calculadas [18] foram momentos de dipolo ($\mu$), polarizabilidades ($\alpha$), campos elétricos nas posições atômicas ($\nabla(n)$, n=1→15), cargas nos átomos com método CHELPG (Q($n$), n=1→15) [106] e densidades eletrônicas em HOMO ($F^H(n)$, n=1→15). Utilizando o programa GAMESS estas propriedades podem ser calculadas com a inclusão do seguinte conjunto de linhas de comando:

```
$elpot iepot=1 where=pdc $end
$elmom iemom=3 where=nuclei $end
$eldens ieden=1 $end
$elfldg iefld=2 $end
$pdc ptsel=chelpg constr=qupole delr=0.3 maxpdc=100000 $end
```

Usando solvatação em água com método D-PCM foram calculadas as mesmas propriedades já enumeradas no cálculo em vácuo e os volumes das cavidades de solvente criadas, resultando um total de cinqüenta e sete propriedades. Para a utilização como descritor nos métodos de regressão multivariada, todas as propriedades foram autoescaladas.

## 3.3.2 Resultados e discussão

### Modelo de regressão PLS baseado em propriedades calculadas com HF(6-311G//3-21G) em vácuo

Com os descritores calculados com método *ab initio* HF(6-311G//3-21G) em vácuo, foi obtido um modelo de regressão PLS com quatorze descritores no bloco das variáveis independentes [129]. Este modelo usa três variáveis latentes (VLs) e prevê os valores experimentais por eliminação sucessiva de cada composto. A Soma dos Quadrados dos Erros de Previsão (PRESS) obtida assim é PRESS=9,62. O Erro Padrão de Previsão,

Tabela 3.1: A estrutura dos compostos da série do fenil-guanidinotiazóis pode ser ontida substituindo os grupos R1, R2 e R3 nas posições determinadas no esqueleto da Figura 3.4. Os valores experimentais da inibição $IC_{50}$ da enzima $H^+K^+$-ATPase pelos compostos da série e os valores de $\log IC_{50}$ experimentais e obtidos em validação cruzada para os modelos QSAR com descritores obtidos com (6-311G//3-21G) e (6-311G//3-21G)AQ são mostrados nas últimas quatro colunas, respectivamente.

| Composto | Substituinte | | | $IC_{50}$ | $\log IC_{50}$ | | |
|---|---|---|---|---|---|---|---|
| No. | R1 | R2 | R3 | ($\mu$mol.L$^{-1}$) | Exp. | (6-311G//321G) | Aquoso |
| **fgt1** | H | H | H | 40 | 1,6 | 1,6 | 1,5 |
| **fgt2** | H | H | *o*-Cl | 18 | 1,2 | 1,9 | 1,6 |
| **fgt3** | H | H | *m*-Cl | 53 | 1,7 | 1,7 | 1,0 |
| **fgt4** | H | H | *p*-F | 33 | 1,5 | 1,3 | 0,9 |
| **fgt5** | $CH_2C_6H_5$ | H | *o*-Cl | 6 | 0,78 | 1,0 | 1,0 |
| **fgt6** | H | H | *o*-$CH_3$ | 20 | 1,3 | 1,2 | 1,6 |
| **fgt7** | H | H | *m*-$CH_3$ | 32 | 1,5 | 1,2 | 1,3 |
| **fgt8** | H | H | *m*-$CH_2N(CH_3)_2$ | 37 | 1,6 | 1,6 | — |
| **fgt9** | $CH_2C_6H_5$ | H | *o*-$CH_3$ | 10 | 1,0 | 0,88 | — |
| **fgt10** | H | H | *p*-$NH_2$ | 32 | 1,5 | 1,7 | 1,5 |
| **fgt11** | H | H | *p*-$CH(CH_3)_2$ | 14 | 1,1 | 1,2 | 1,2 |
| **fgt12** | $CH_2C_6H_5$ | H | *p*-$NH_2$ | 5,2 | 0,72 | 0,97 | 0,79 |
| **fgt13** | H | H | *o*-$OCH_3$ | 27 | 1,4 | 1,2 | 1,3 |
| **fgt14** | H | H | *m*-$OCH_3$ | 60 | 1,8 | 1,7 | 1,0 |
| **fgt15** | H | H | *p*-$OCH_3$ | 35 | 1,5 | 1,0 | 0,66 |
| **fgt16** | $CH_2C_6H_5$ | H | *o*-$OCH_3$ | 2,3 | 0,36 | 0,70 | 1,1 |
| **fgt17** | $CH_2C_6H_5$ | H | *m*-$OCH_3$ | 10 | 1,0 | 0,74 | 1,2 |
| **fgt18** | H | H | *m*-OH | 23 | 1,4 | 0,97 | 0,71 |
| **fgt19** | H | H | *p*-OH | 15 | 1,2 | 0,38 | 1,0 |
| **fgt20** | $CH_2C_6H_5$ | H | *o*-OH | 11 | 1,0 | 0,66 | 1,4 |
| **fgt21** | H | H | *m,p*-$(OH)_2$ | 1,5 | 0,18 | 0,59 | 0,59 |
| **fgt22** | $CH_2C_6H_5$ | H | *m,p*-$(OH)_2$ | 0,3 | -0,52 | -0,095 | 0,26 |
| **fgt23** | H | $CH_3$ | *m,p*-$(OH)_2$ | 1,6 | 0,20 | 0,27 | -0,011 |
| **fgt24** | $CH_2C_6H_5$ | $CH_3$ | *m,p*-$(OH)_2$ | 1,3 | 0,11 | 1,1 | 0,69 |

Tabela 3.2: Parâmetros para estimativa da confiabilidade da previsão dos valores de $\log IC_{50}$ pelos melhores modelos QSAR obtidos, em validação cruzada. Os valores calculados estão listados na 3.1 para modelos de regressão PLS com quatorze descritores quânticos calculados com HF(6-311G//3-21G) representados em três variáveis latentes e também para o modelo com dez descritores calculados com HF(6-311G//3-21G)AQ representados em duas variáveis latentes.

| Parâmetros da Regressão | HF(6-311G//3-21G) | HF(6-311G//3-21G)AQ |
|---|---|---|
| Valor de $q^2$ | 0,77 | 0,64 |
| Valor de PRESS | 9,62 | 12,6 |
| Valor de SDEP | 3,10 | 3,56 |
| Análise de variância | | |
| Teste F | 31,2 | 13,4 |

SDEP=3,10. O teste $q^2$ apresenta valor $q^2$=0,77. Os valores dos indicadores da acuracidade da regressão e da capacidade de previsão estão listados na Tabela 3.1.

O teste $q^2$ relaciona-se á validação cruzada, e seu valor situa-se entre zero e um. Valores acima de $0,3$ unidades podem ser considerados aceitaveis para modelos obtidos com a técnica denominada CoMFA (*Comparative Molecular Field Analysis*) [130], uma técnica para desenvolvimento de modelos QSAR que utiliza regressão multivariada entre conjuntos numerosos de descritores calculados e valores experimentais de atividades. O valor de PRESS deve ser tão baixo quanto possível, não havendo um limite definido, assim como o valor de SDEP. O teste $F$ deve ter o valor situado acima de um valor tabelado, que depende do número de amostras utilizadas no modelo, do número de variáveis utilizadas para construção do modelo e do intervalo de confiança adotado. No caso o valor tabelado do limite $F^{0,995}_{[24;20]} = 3,22$, o que significa que com os valores listados na Tabela 3.2 os modelos de regressão apresentados são considerados válidos com 99,5% de chance de acerto. O número de amostras utilizadas no modelo é vinte e quatro. O número de amostras menos o número de variáveis latentes menos um, totalizndo vinte, é necessário para determinar qual valor tabulado deve ser utilizado para comparação no teste F, conforme a Equação 1.100.

O gráfico do ajuste entre valores previstos e experimentais pode ser visto na Figura 3.5. O gráfico mostra resultados para o conjunto de quatorze variáveis independentes selecionadas: $\mu^{CHELPG}$, $\mu_y^{CHELPG}$, $\alpha_{xx}$, $\alpha_{yy}$, $\alpha_{xy}$, $\nabla(2)$, Q(1), Q(3), Q(4), Q(10), $Q(11)$, Q(14), $F^H(10)$ e $F^H(14)$, com as posições atômicas numeradas de acordo com os desenhos das Figuras 3.2 e 3.3. O teste de dispersão das amostras no espaço vetorial das variáveis latentes com 95% de confiança (Q) têm valor de Q = 16,14 e valor de $T^2$ = 10,10. Os compostos **fgt17** e **fgt19** da Tabela 3.1 têm valores de Q além deste limite. Nenhum composto da Tabela 3.1 tem valor de $T^2$ além do limite.

Na Figura 3.6 podem ser vistos os valores de Q e $T^2$ para cada composto da Tabela 3.1, e o valor do limite de Q e $T^2$ com 95% de confiança. Valores acima do limite para os parâmetros de dispersão freqüentemente indicam a presença de *outliers*, que são compostos com valores anômalos no conjunto de variáveis independentes e/ou nas relações entre

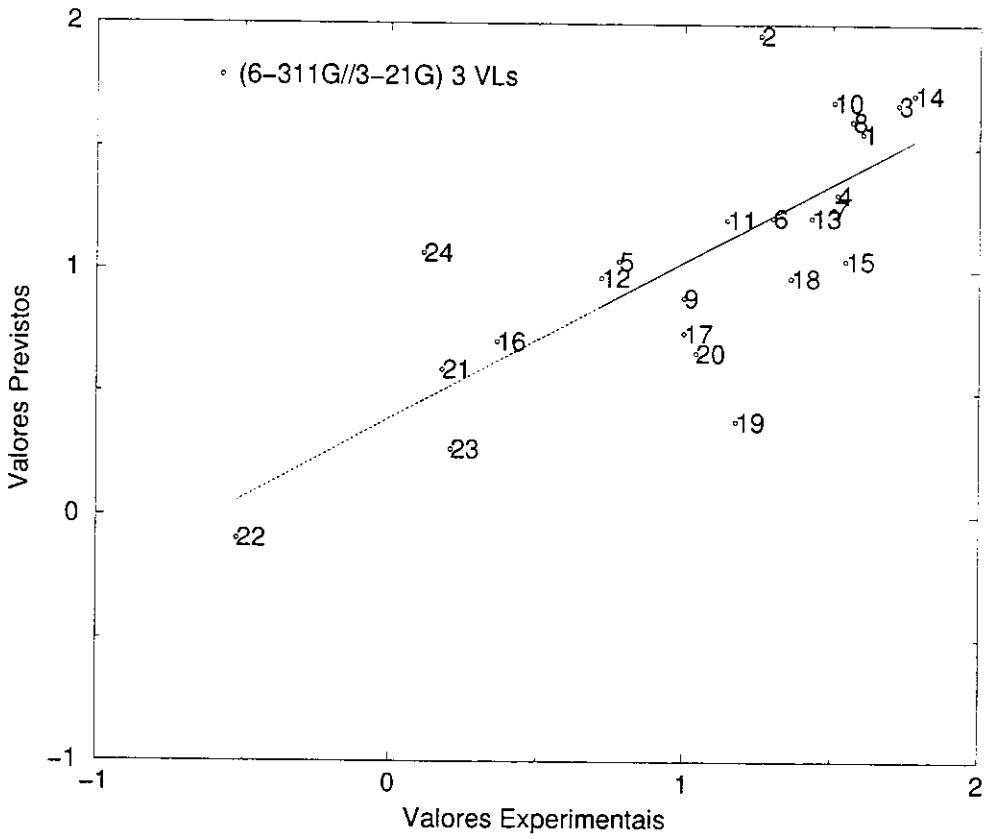



Figura 3.5: Valores das atividades experimentais versus valores previstos por validação cruzada, por eliminação sucessiva para cada composto listado na Tabela 3.1, com os descritores obtidos em cálculo *ab initio* HF(6-311G//3-21G).

variáveis independentes. Os valores de $Q$ e $T^2$ mostram como os compostos estão distri-

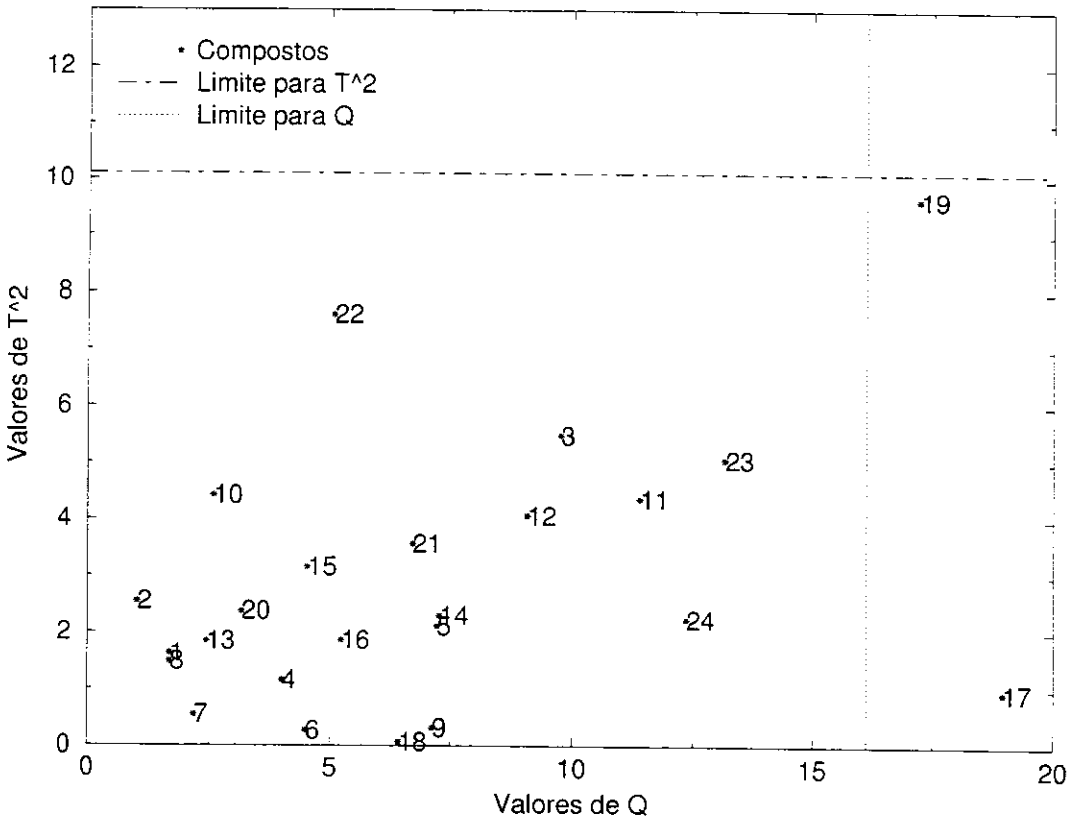



Figura 3.6: Valores de Q e $T^2$ obtidos no modelo de regressão PLS com três variáveis latentes, para os compostos listados na Tabela 3.1, utilizando-se quatorze descritores quânticos calculados com HF(6-311G//3-21G) no bloco das variáveis independentes e $\log IC_{50}$ no bloco das dependentes. Este gráfico mostra uma dispersão aceitável para os compostos, exceto **fgt17** e **fgt19** que têm valores elevados para estes indicadores, indicando que os resultados não são confiáveis para estes compostos neste modelo.

buídas no espaço vetorial das três variáveis latentes em função dos seus *scores*. *Scores* são as novas coordenadas geradas pelo modelo PLS para cada compostos em substituição ao conjunto original de coordenadas representadas pelas variáveis independentes. Os *scores* das variáveis independentes serão utilizadas para regressão dos valores de *scores* das variáveis dependentes. O método NIPALS para regressão PLS cria matrizes de coeficientes chamadas 'matriz de pesos' e 'matriz de relações internas', que relacionam as coordenadas de cada amostra com as coordenadas das suas respectivas variáveis dependentes levando em conta as rotações nos vetores das coordenadas que o sistema sofre após a obtenção de cada variável latente. Pode haver mais de uma variável dependente, ex. $\log IC_{50}$ e toxicidade, em um único modelo de regressão PLS.

Uma variável latente é um vetor que tem o compromisso de acumular a melhor variância possível dos blocos aliada ao melhor ajuste entre os vetores obtidos para cada bloco. Obtida uma variável latente faz-se a subtração da variância dos dados acumulados nesta variável latente e nova rotação de coordenadas para obtenção da próxima variável

Tabela 3.3: Valores de variância acumulada para cada variável latente do modelo PLS com quatorze descritores obtidos com método ab-initio (6-311G//3-21G) para os fenil-guanidinotiazóis

| | Bloco X | | Bloco Y | |
|---|---|---|---|---|
| VL# | Esta VL | Total | Esta VL | Total |
| #1 | 17,19 | 17,19 | 72,53 | 72,53 |
| #2 | 20,94 | 38,13 | 9,72 | 82,25 |
| #3 | 10,70 | 48,83 | 3,14 | 85,38 |

latente, que acomodará novamente a maior variância possível, levando em conta também o ajuste entre os *scores* das variáveis independentes e dependentes [131].

Na Tabela 3.3 estão listadas as variâncias acumuladas em cada variável latente utilzada no modelo final. Na Tabela 3.4 estão listadas as contribuições de cada variável latente para o modelo final de regressão, e os coeficientes do vetor de regressão para um modelo de regressão da forma $\sum_{k=1}^{n} C_k D_k$, onde os valores de $C$ são os coeficientes do vetor de regressão para as $n$ variáveis descritoras $D$, autoescaladas. Estes valores da contribuição de cada variável são utilizados para seleção de variáveis independentes. A idéia é que se uma variável independente tem grande contribuição em variáveis latentes com grande quantidade de variância acumulada, ela será importante para obtenção de um bom modelo de regressão. Usando este raciocínio foram selecionadas quatorze variáveis independentes do conjunto inicial de cinqüenta e seis. O vetor de regressão linear $\mathbf{m}$ é definido de acordo com as Equações 1.86 da página 42.

A menor variância acumulada no bloco das variáveis independentes reflete uma grande quantidade de informação não utilizada deste bloco para a construção do modelo final. Não significa que o modelo não será capaz de previsões boas, e sim que no bloco das variáveis independentes há bastante variação que não pode ser associada ao bloco das variáveis dependentes, que são as atividades biológicas que deseja-se modelar.

### Modelo de regressão PLS baseado em parâmetros (6-311G//3-21G) com solvatação PCM em água

Os dados obtidos com o modelo de solvatação PCM para as moléculas foram utilizados na construção do modelo PLS de regressão com dez variáveis independentes. Este modelo obtido com método PLS com duas variáveis latentes tem resultados piores que o obtido com os dados sem solvatação se consideramos a validação cruzada *leave one out*, conforme pode ser observado na listagem das Tabelas 3.1 e 3.2. O gráfico da Figura 3.7 mostra ajuste entre os valores experimentais e os valores obtidos no teste de validação cruzada para a regressão.

Estes valores mostram uma regressão com algum poder de previsão. Os valores para os parâmetros de dispersão das amostras nos *scores* são de Q=14,05 e $T^2$=7,33. Para os compostos **fgt3** e **fgt17** os valores de Q são bem mais altos que o limite mostrado na Figura 3.8. Apesar de estarem fora da média dos valores dos demais compostos no

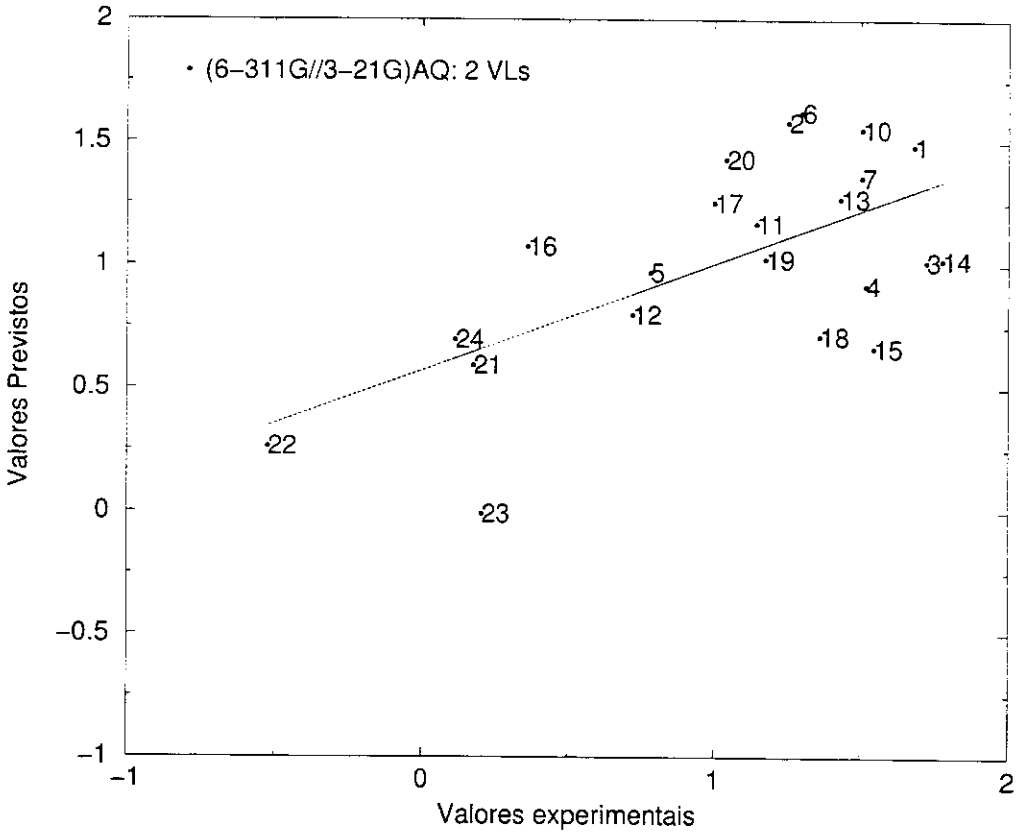



Figura 3.7: Valores experimentais versus valores previstos com validação cruzada por eliminação sucessiva, com dez descritores obtidos em cálculo ab-initio (6-311G//3-21G), com efeito de solvatação em água obtido com método PCM. Os descritores utilizados são $\mu^{CHELPG}$, $\alpha_{xx}$, $\alpha_{yy}$, $\nabla(2)$, $\nabla(9)$, Q(1), Q(6), $F^H(7)$, $F^H(11)$, $F^H(14)$

Tabela 3.4: Coeficientes de regressão para cada variável latente do modelo PLS com quatorze descritores obtidos com método ab-initio (6-311G//3-21G) para os fenil-guanidinotiazóis e coeficiente de regressão **m** (ver Eq. 1.86 da página 42) para equação de regressão simples utilizando os descritores autoescalados

| Variável | Variável Latente # | | | Coeficiente |
|---|---|---|---|---|
| Descritora | #1 | #2 | #3 | de Regressão |
| $\mu^{CHELPG}$ | -0,1391 | 0,0482 | 0,0193 | -0,0715 |
| $\mu_y^{CHELPG}$ | -0,1400 | 0,0469 | 0,0173 | -0,0758 |
| $\alpha_{xx}$ | 0,1604 | -0,0401 | -0,0328 | 0,0875 |
| $\alpha_{yy}$ | -0,1983 | 0,0301 | 0,0506 | -0,1176 |
| $\alpha_{xy}$ | -0,0745 | 0,0025 | 0,0171 | -0,0549 |
| $\nabla(2)$ | -0,0530 | 0,0660 | 0,0289 | 0,0419 |
| Q(1) | -0,1749 | -0,0353 | -0,0610 | -0,2713 |
| Q(3) | 0,1806 | 0,0417 | 0,0032 | 0,2255 |
| Q(4) | -0,0874 | -0,0665 | 0,0387 | -0,1151 |
| Q(10) | 0,1368 | 0,0951 | -0,0116 | 0,2204 |
| Q(11) | -0,1727 | -0,0525 | -0,1350 | -0,3603 |
| Q(14) | -0,1533 | 0,0112 | -0,0156 | -0,1577 |
| $F^H(10)$ | -0,2128 | -0,0852 | -0,0050 | -0,3031 |
| $F^H(14)$ | -0,2240 | -0,0818 | -0,0236 | -0,3294 |

Tabela 3.5: Valores de variância acumulada para cada variável latente do modelo PLS com dez descritores obtidos com método ab-initio (6-311G//3-21G) utilizando efeito de solvatação em água com método D-PCM para os fenil-guanidinotiazóis

| | Bloco X | | Bloco Y | |
|---|---|---|---|---|
| VL# | Esta VL | Total | Esta VL | Total |
| # 1 | 25,98 | 25,98 | 55,60 | 55,60 |
| # 2 | 18,24 | 44,22 | 9,79 | 65,38 |

sub-espaço vetorial dos *scores*, estes compostos têm boa previsão como pode ser visto no gráfico da Figura 3.8, e por isso foram mantidos no modelo. As variâncias de cada variável latente do modelo estão listadas na Tabela 3.5. O número de variáveis latentes escolhido é o que retorna o menor PRESS para o conjunto. As variáveis independentes foram escolhidas adotando os mesmos critérios adotados para seleção de variáveis no modelo sem solvatação, em dois passos. Foi testado um modelo com cinqüenta e sete variáveis, selecionadas trinta e uma variáveis e finalmente as dez variáveis descritoras que resultaram o modelo apresentado. Os descritores utilizados são $\mu^{CHELPG}$, $\alpha_{xx}$, $\alpha_{yy}$, $\nabla(2)$, $\nabla(9)$, Q(1), Q(6), $F^H(7)$, $F^H(11)$, $F^H(14)$. Utilizando estes descritores o número de variáveis latentes foi reduzido para duas no modelo final, levando em consideração o número de compostos presentes no conjunto de treinamento, a variância acumulada nas variáveis latentes desprezadas e o PRESS.

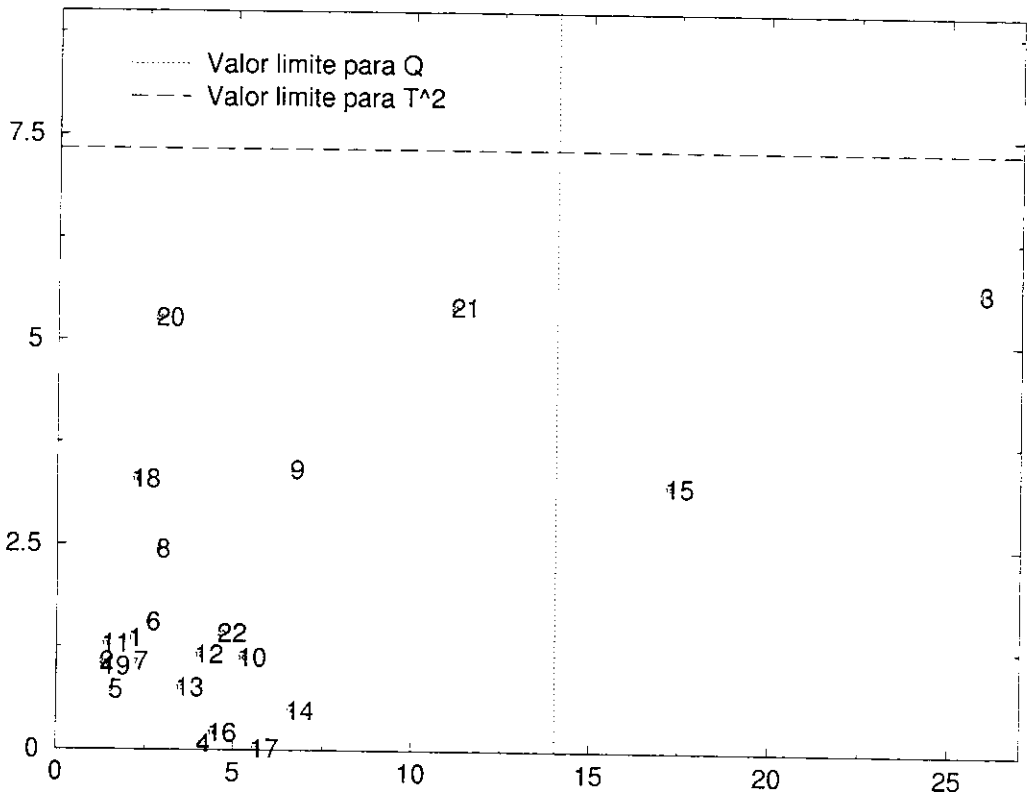



Figura 3.8: Valores de Q e $T^2$ obtidos para o modelo de regressão PLS com duas variáveis latentes para os compostos listados na Tabela 3.1 utilizando-se dez descritores quânticos no bloco das variáveis independentes e $\log IC_{50}$ no bloco das dependentes. Este gráfico mostra uma dispersão bastante pequena para a maioria dos compostos, e para os compostos **fgt3** e **fgt17** uma dispersão acima do limite proposto nas metodologias de análise de dados.

### 3.3.3 Conclusões da comparação entre os modelos em vácuo e com solvatação

- A aplicação dos efeitos de solvatação permitu a criação de um modelo mais compacto, com menor número de variáveis descritoras e de variáveis latentes, para uma qualidade de previsão comparável.
- O aumento no tempo de computação faz do procedimento de solvatação uma etapa a ser considerada a cada caso.
- Os descritores selecionados são bastante semelhantes, havendo grande coerência na indicação das regiões de interesse na série de compostos pelos dois tipos de modelos obtidos.

  **Os momentos de dipolo e as polarizabilidades** são utilizados de maneira consistente nos dois modelos.

  **As cargas** sobre as posições atômicas Q(1) e Q(3) correspondem igualmente à posição *meta* de substituição no anel, assim como Q(6) e Q(4) correspondem à substituição *orto*.

  **As densidades eletrônicas em HOMO** $F^H(7)$ e $F^H(11)$ são relevantes no modelo com solvatação, e correspondem á mesma região do orbital molecular que a densidade $F^H(10)$, que foi indicada como relevante no modelo sem solvatação.

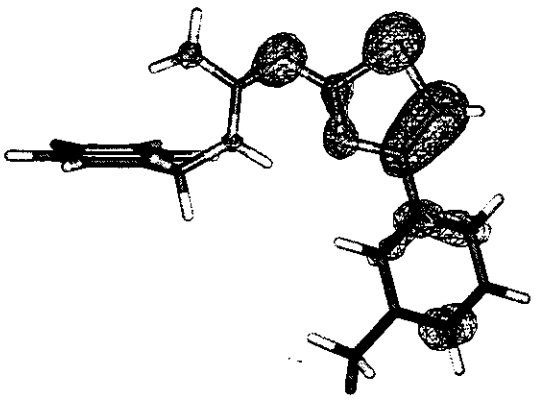


Figura 3.9: A superfície de densidade eletrônica do HOMO mostra as regiões onde a densidade é mais intensa, resultando uma maior capacidade para doação de elétrons para orbitais de alta energia nestas regiões. Na região do anel de tiazol a densidade sobre os átomos C(7), C(11) e S(10) são apontadas nos modelos obtidos como relevantes, e provavelmente caracterizam uma região doadora de elétrons. A região do átomo N(14) foi apontada como relevante, e provavelmente serve como indicadora da substituição do grupo R(1) por benzila ou hidrogênio, considerando que a densidade nesta região, em HOMO, é sempre pequena.

As cargas nas posições atômicas do anel benzênico *orto* e *para* são indicadores o tipo de substituição no anel. Variando a eletrofilicidade do substituinte temos os efeitos indutivos

e de hiperconjugação em diferentes graus, que influem nas cargas calculadas sobre os átomos do anel. O conjunto dos descritores não tem variáveis mapeadas diretamente sobre os substituintes, para contornar o problema do alinhamento e comparação entre descritores calculados sobre átomos de elementos diferentes.

As densidades eletrônicas em HOMO no anel de tiazol (posições sete, dez e onze) indicam que esta região provavelmente age como doadora de elétrons na interação com o receptor. A densidade no nitrogênio da guanidina (posição quatorze) está indicando a presença ou não do substituinte benzila ligado à posição. A suposição de que a densidade eletrônica nesta região indique um caráter doador de elétrons nesta posição pode ser um engano, já que a densidade eletrônica nesta região é bem pequena, independentemente da substituição ou não pelo grupo benzila, como pode ser observado na Figura 3.9.

## 3.4 Os compostos anti-úlcera da classe dos indolil-guanidinotiazóis

### 3.4.1 Métodologia de trabalho e procedimentos adotados

O seleção da conformação bioativa foi realizada selecioando-se o tipo de conformação em que os descritores calculados permitiram a construção de um modelo de calibração melhor. Para dois dos vinte e oito compostos da série dos indolil-guanidinotiazóis foram realizados estudos detalhados sobre o espaço conformacional. Cálculos com o programa `distgeom` (faz parte do pacote `TINKER`) usando o método da Matriz de Distâncias Métricas para geração de conformações, seguido de otimização das estruturas com o método semi-empírico AM1 no `MOPAC`, comparação das geometrias após otimização e eliminação das repetidas com o programa `superpose` (`TINKER`).

Foram geradas duas mil estruturas, com imposição de quiralidade para os átomos tetraédricos, imposição de geometria planar e/ou quiral para átomos trigonais e imposição de planaridade torsional para átomos trigonais adjacentes. O programa para geração das estruturas (`distgeom`) e a otimização de geometrias de cada estrutura gerada foi automatizada em um *script* na linguagem *Perl* escrito por Godinho [99]. A comparação entre estruturas e eliminação das repetidas foi realizada com outro *script*, também desenvolvido por Godinho.

Os compostos **igt44** e **igt45** da Tabela 3.7 são representativos dos dois tipos gerais de estrutura para a série, e foram escolhidos para estudo conformacional. As estruturas tipo **I** têm substituinte hidrogênio na posição R1 da Figura 3.10. As estruturas tipo **II** têm substituinte benzila na posição R1 (Figura 3.10). A escolha do composto **igt44** para representar as estruturas tipo **I** foi por ser o mais ativo da série de compostos. O composto **igt45** foi escolhido porque tem estrutura semelhante ao mais ativo, com substituinte benzila na posição R1, representando as estruturas tipo **II** .

Para os compostos do tipo **I** foi escolhida a ligação **phi1** do composto **igt44** para caracterização da conformação, mostrada na Figura 3.11. Os resultados da busca estão listados na Tabela 3.6, onde as duas conformações evidentemente diferentes são denominadas **igt44.A** e **igt44.B**. No composto **igt45** , os ângulos diedro das ligações **phi1**,

Tabela 3.6: Valores dos ângulos de diedro para as conformações **A** e **B** dos compostos **igt44** e **igt45** (ver Figura 3.11). As energias relativas de cada conformação obtidas com método AM1 foram utilizadas para determinar as proporções relativas de cada conformação. As proporções calculadas pela expressão de Boltzman são iguais com aproximação de um digito decimal.

| Composto | **igt44** | | **igt45** | |
|---|---|---|---|---|
| Conformação | **A** | **B** | **A** | **B** |
| Ângulo **phi 1** | 48 | -152 | 49 | -152 |
| Ângulo **phi 3** | — | — | 168 | 1 |
| Ângulo **phi 5** | — | — | -154 | 68 |
| $\Delta H_f$ [a] | -126,0811 | -126,0822 | -162,0294 | -162,0312 |
| População [b] | 1 | 1 | 1 | 1 |

[a]Calor de formação em kcal.mol$^{-1}$(AM1)

[b]Distribuição relativa entre as populações de cada uma das conformações de mínimo local de energia.

**phi2**, **phi3**, **phi4**, **phi5** e **phi6** foram utilizados para caracterizar a conformação. As conformações do composto **igt43** foram classificadas pelas variações dos ângulos de diedro das ligações **phi1**, **phi3** e **phi5**. Estas são as ligações que caracterizam mais claramente as diferenças entre as conformações de mínimo encontradas, denominadas **igt45.A** e **igt45.B** . O ângulo de diedro das ligações **phi2**, **phi4** e **phi6** também muda, porém as variações nestes ângulos não caracterizam bem as diferentes conformações. Os calores de formação resultantes para as conformações de mínimo encontradas com AM1 estão listados na Tabela 3.6.

As populações relativas de cada conformação de mínimo local dos compostos **igt44** e **igt45** são praticamente iguais como pode ser observado na Tabela 3.6. O cálculo das populações considera o espaço conformacional restrito apenas às duas conformações apresentadas, que são os mínimos locais de menor energia do conjunto de conformações otimizadas a partir das duas mil geradas. As energias dos pares de conformações foram utilizadas na Equação de 3.1, para obter as proporções entre as populações das conformações **igt44.A** e **igt44.B** ou **igt45.A** e **igt45.B** . As energias calculadas com AM1 mostram uma população igual para todas as conformações listadas na Tabela 3.6.

Todos os compostos da série podem ser obtidos pelas substituições da Tabela 3.7 na estrutura da Figura 3.10. Para obter os compostos tipo **I** e **II** em dois grupos de conformações, foram usadas as estruturas da Figura 3.11. As estruturas **igt44.A** e **igt45.A** foram usadas para obter as conformações **A** dos compostos tipo **I** e **II** , respectivamente. As estruturas **igt44.B** e **igt45.B** foram usadas para obter as conformações **B** dos compostos.

Todas os compostos nas conformações **A** e **B** foram completamente otimizados com método semi-empírico AM1. Estas moléculas foram utilizadas para cálculo de propriedades com método *ab initio* no nível Hartree-Fock com base 3-21G (3-21G//AM1). As propriedades calculadas foram as energias dos últimos onze orbitais ocupados ($E_{HOMO-n}$ $n = 0 \rightarrow 11$), os momentos de dipolo e quadrupolo (dez descritores), as cargas CHELPG para

os dezoito átomos do esqueleto fundamental numerados de um até dezoito na conformação **B** da Figura 3.10, as densidades eletrônicas totais para os mesmos dezoito átomos do esqueleto e as densidades eletrônicas em HOMO e em LUMO para os dezoito átomos do esqueleto (trinta e seis descritores). No total foram calculados noventa e quatro descritores quânticos. Estes cálculos foram realizados no programa GAMESS. Programas para automatização das tarefas de cálculo e tabulação de variáveis são transcritos no Apêndice A.2.

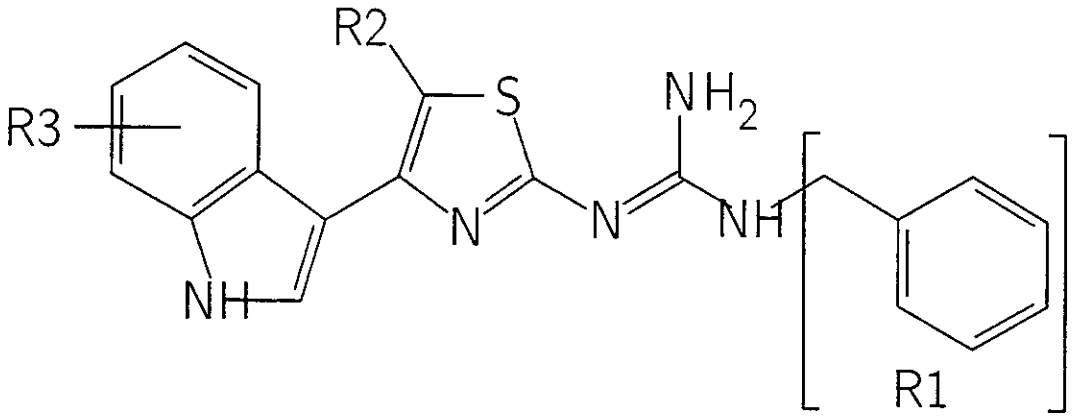


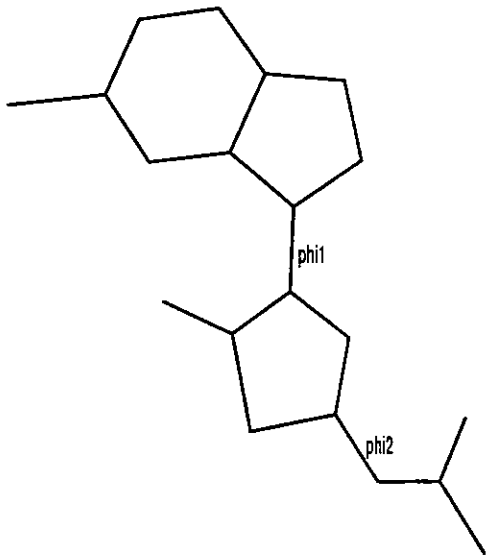

Figura 3.10: Estrutura do esqueleto principal dos compostos da série dos indolil-guanidinotiazóis. Adicinando-se os substituintes listados na Tabela 3.7 nas posições R1, R2 e R3, podemos construir toda a série de compostos estudada. O grupo R1 corresponde a todo o grupo benzila na figura. A numeração mostrada na figura é a mesma adotada para numeração das variáveis relacionadas ás posições atômicas.

Utilizando os vinte e oito compostos listados na Tabela 3.7 nas conformações **A** e **B** foram realizados dois modelos de regressão multivariada entre os noventa e quatro descritores já citados de cada conformação e as atividades biológicas. Estes modelos foram realizados com método PLS utilizando os dados autoescalados representados em por vinte variáveis latentes. A conformação extendida **A** resultou o modelo com melhor ajuste entre os valores obtidos em validação cruzada *leave one out* e valores experimentais de atividade. Este resultado foi utilizado para selecionar a conformação **A** para o prosseguimento dos cálculos. As moléculas dos compostos listados na Tabela 3.7 que pertencem ao tipo **I** criadas com base no composto **igt44.A** e aquelas do tipo **II** com no composto **igt45.A**, foram completamente otimizadas no método *ab-initio* no nível Hartree-Fock usando base 3-21G.

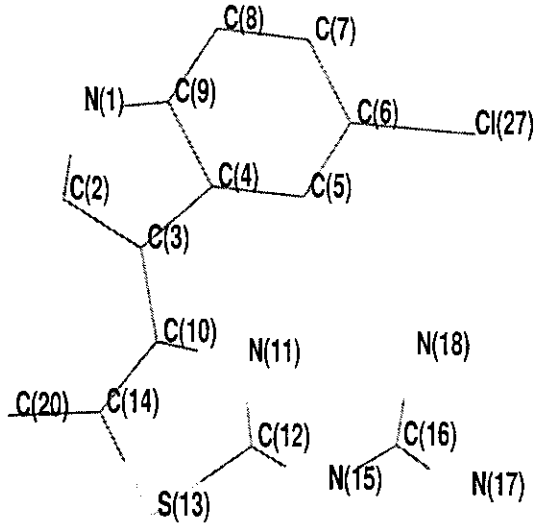
Para cada um dos vinte e oito compostos na conformação **A** foram calculadas um total de setenta e seis variáveis no nível Hartree-Fock (6-311G//3-21G). Os momentos de dipolo e quadrupolo totalizam onze variáveis. Os potenciais eletrostáticos nas posições atômicas numeradas de um até dezoito nos esqueletos da Figura 3.11, as cargas líquidas calculadas com método CHELPG nas mesmas dezoito posições e as densidades eletrônicas em HOMO totalizam mais cinqüenta e quatro variáveis. Finalmente, as energias dos orbitais HOMO, HOMO-1, ..., HOMO-10 totalizam mais onze variáveis. As densidades eletrônicas totais não foram utilizadas, nem as densidades eletrônicas do LUMO. Apesar

Tabela 3.7: Substituintes nas posições R1, R2 e R3 do esqueleto da Figura 3.10 para obtenção dos compostos da série dos compostos indolil-guanidinotiazóis, e valores de $IC_{50}$ na enzima $H^+K^+$-ATPase, experimentais e estimados com validação cruzada pelos modelos com redes neurais (LO4), PLS (LO5) e PLS para mais de um composto por vez (EV4 e EV5).

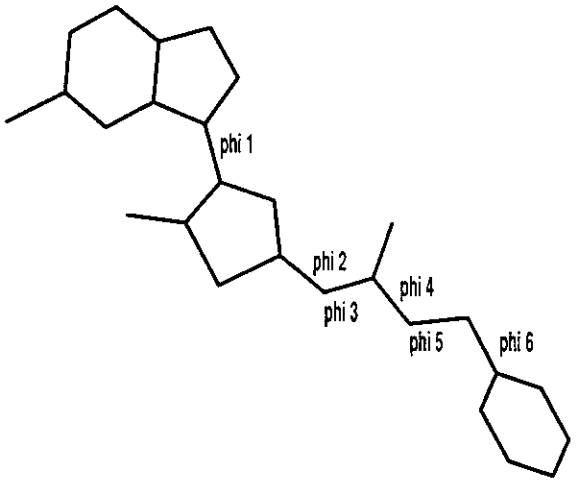
| Composto | Substituinte | | | $IC_{50}(\mu mol.l^{-1})$ | | | | |
|---|---|---|---|---|---|---|---|---|
| No. | R1 | R2 | R3 | Exp. | LO4 | LO5 | EV4 | EV5 |
| **igt25** | H | H | H | 9,0 | 15,5 | 2,41 | — | — |
| **igt26** | $CH_2C_6H_5$ | H | H | 1,5 | 1,07 | 1,16 | 1,47 | — |
| **igt27** | H | $CH_3$ | H | 7,6 | 2,92 | 4,96 | — | — |
| **igt28** | H | H | 5-$OCH_3$ | 19,0 | 9,43 | 10,6 | — | 2,13 |
| **igt29** | $CH_2C_6H_5$ | H | 5-$OCH_3$ | 4,3 | 0,955 | 1,21 | — | — |
| **igt30** | H | H | 5-$OCH_2C_6H_5$ | 1,2 | 2,16 | 0,78 | — | — |
| **igt31** | $CH_2C_6H_5$ | H | 5-$OCH_2C_6H_5$ | 1,8 | 0,215 | 1,12 | — | — |
| **igt32** | H | H | 2-$CH_3$ | 9,0 | 11,9 | 5,64 | — | — |
| **igt33** | $CH_2C_6H_5$ | H | 2-$CH_3$ | 1,7 | 1,65 | 2,04 | — | 1,49 |
| **igt34** | H | $CH_3$ | 2-$CH_3$,5-Cl | 0,6 | 0,245 | 0,39 | 0,62 | — |
| **igt35** | $CH_2C_6H_5$ | $CH_3$ | 2-$CH_3$,5-Cl | 1,1 | 0,889 | 0,59 | — | — |
| **igt36** | $CH_2C_6H_5$ | H | 4-$CH_3$ | 3,3 | 1,44 | 1,17 | — | — |
| **igt37** | H | H | 5-$CH_3$ | 1,8 | 3,81 | 3,35 | — | — |
| **igt38** | $CH_2C_6H_5$ | H | 5-$CH_3$ | 0,94 | 1,04 | 1,04 | — | 0,42 |
| **igt39** | $CH_2C_6H_5$ | H | 6-$CH_3$ | 1,6 | 1,30 | 1,22 | — | — |
| **igt40** | H | H | 7-$CH_3$ | 7,6 | 2,67 | 5,08 | 7,43 | — |
| **igt41** | $CH_2C_6H_5$ | H | 7-$CH_3$ | 1,0 | 1,94 | 0,79 | — | — |
| **igt42** | H | H | 5-Cl | 1,8 | 1,93 | 1,20 | — | — |
| **igt43** | $CH_2C_6H_5$ | H | 5-Cl | 0,7 | 0,88 | 0,38 | 0,70 | 0,05 |
| **igt44** | H | $CH_3$ | 5-Cl | 0,59 | 0,703 | 0,38 | — | — |
| **igt45** | $CH_2C_6H_5$ | $CH_3$ | 5-Cl | 0,7 | 5,47 | 0,82 | — | — |
| **igt46** | H | H | 5-Br | 0,96 | 1,16 | 1,00 | — | — |
| **igt47** | $CH_2C_6H_5$ | H | 5-Br | 0,7 | 0,804 | 0,28 | — | — |
| **igt48** | H | H | 5-F | 7,4 | 2,52 | 3,91 | — | 8,62 |
| **igt49** | $CH_2C_6H_5$ | H | 5-F | 2,7 | 1,34 | 1,82 | 2,55 | — |
| **igt50** | H | H | 5-$CO_2CH_3$ | 6,7 | 2,79 | 1,50 | — | — |
| **igt51** | $CH_2C_6H_5$ | H | 5-$CO_2CH_3$ | 3,1 | 5,58 | 2,32 | — | — |
| **igt52** | $CH_2C_6H_5$ | H | 5-NHCOEt | 23,8 | 5,04 | 10,7 | — | — |



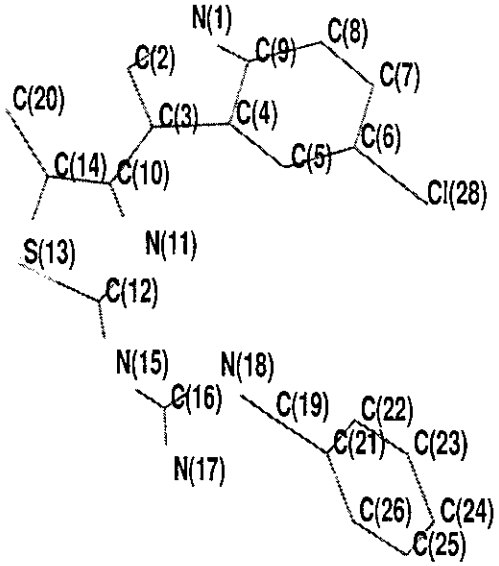

**igt44.A**




**igt44.B**




**igt45.A**




**igt45.B**

Figura 3.11: Conformações correspondentes aos mínimos de energia encontrados na busca conformacional dos compostos **igt44** e **igt45** da Tabela 3.7, mostrados nas conformações **A** e **B**. As conformações **B** mostram a numeração adotada para os átomos para as propriedades calculadas.

de terem sido calculadas no modelo preliminar para seleção da conformação, seu poder de modelagem foi considerado pequeno.

Um *script* em `csh` denominado `buscadados.csh` foi desenvolvido para realizar a busca de variáveis nos arquivos de dados de saída do programa `GAMESS`. O programa está transcrito no Programa A.3.2 do Apêndice A.3, assim como as funções em para o comando `awk`, utilizadas para converter as linhas de dados dos arquivos de saída em vetores de dados convenientes para entrada no programa `octave`, para análise de dados.

As variáveis descritoras calculadas foram todas autoescaladas antes de iniciar o tratamento estatístico. O conjunto de variáveis com média zero e variância unitária constitui as variáveis independentes para o modelo QSAR. O conjunto de dados experimentais que mostra as atividades de cada droga sobre a enzima $H^+$,$K^+$-ATPase gástrica foi tratado de outra maneira. Foram obtidos os logarítimos (base dez) de cada valor de concentração necessária para causar uma inibição de cinqüenta porcento da atividade da enzima ($IC_{50}$), e este conjunto de valores foi ainda autoescalado. Estas variáveis assim tratadas constituem o conjunto de variáveis dependentes para o tratamento estatístico.

Os conjuntos de variáveis dependentes e independentes foram utilizados em regressão tipo PLS para estabelecer a possível correlação entre eles. Os resultados das análises preliminares com todo o conjunto de variáveis independentes foram utilizados sucessivamente para selecionar variáveis com maior poder de modelagem. Melhor poder de modelagem significa que são capazes de estabelecer um modelo que estima melhores valores das variáveis dependentes em validação cruzada tipo *leave one out*. Este tipo de análise foi utilizada com todo o conjunto de compostos para tentar selecionar qual o subconjunto de variáveis independentes tem maior poder de modelagem.

O processo de seleção de variáveis utiliza os vetores de regressão obtidos de acordo com e Equação 1.86, somando os quadrados dos coeficientes de regressão para cada uma das variáveis latentes necessárias para obter o mínimo PRESS. Os valores quadrados foram utilizados para evitar que uma variável com grande contribuição em duas variáveis latentes, porém com sinais trocados, tivesse sua importância sub-estimada. O efeito de cancelamento da contribuição medida pelo soma dos coeficientes do vetor de regressão ocorre fortemente nas variáveis $\mu_x$, $q_{zz}$, $\nabla(8)$, $Q(6)$ e $Q(13)$ da Tabela 3.12, que tem sinais invertidos nos coeficientes da primeira e segunda variáveis latentes, ou sinais iguais nas duas primeiras e invertido nas demais.

Foram testados seis modelos com setenta e seis, cinqüenta, quarenta e cinco, vinte e sete, dezesseis e quatorze variáveis independentes. A seleção de variáveis foi realizada por inspeção do vetor de regressão, tendo como critério a minimização do PRESS. Nos modelos de validação cruzada *leave one out* foram selecionados os descritores com alto coeficiente de regressão nas variáveis latentes que resultavam concomitantemente o menor valor de PRESS. O subconjunto de variáveis independentes foi utilizado para obter um novo modelo, e o vetor de regressão deste novo modelo, com o número de variáveis latentes que resultava no menor PRESS, foi inspecionado para verificar quais variáveis independentes tinham coeficiente maior que o limite. Este procedimento foi realizada até que o modelo resultante fez previsões piores dos valores experimentais, e o penúltimo modelo foi aplicado.

O número ótimo de variáveis latentes foi então determinado para este subconjunto de

Tabela 3.8: Variâncias acumuladas nas variáveis dependentes e independentes pelo modelo PLS com dezesseis variáveis descritoras para os compostos indolil-guanidinotiazóis da Tabela 3.7

| VL # | Var. Independentes | | Var. Dependentes | |
|---|---|---|---|---|
| | Esta VL | Total | Esta VL | Total |
| # 1 | 19,09 | 19,09 | 58,16 | 58,16 |
| # 2 | 18,12 | 37,21 | 12,16 | 70,32 |
| # 3 | 16,03 | 53,25 | 5,32 | 75,64 |
| # 4 | 4,65 | 57,90 | 5,27 | 80,92 |
| # 5 | 3,53 | 61,43 | 3,06 | 83,98 |
| # 6 | 5,93 | 67,37 | 1,35 | 85,33 |
| # 7 | 6,32 | 73,69 | 0,70 | 86,03 |
| # 8 | 5,77 | 79,46 | 0,40 | 86,43 |
| # 9 | 6,94 | 86,40 | 0,12 | 86,54 |
| # 10 | 3,13 | 89,53 | 0,06 | 86,61 |

variáveis selecionado para resultar maior correlação entre dados experimentais e previstos. O valor máximo do coeficiente de correlação entre os valores previstos com validação cruzada *leave one out* foi novamente utilizado para determinar qual o número ótimo de variáveis latentes para o modelo. Este número de variáveis latentes determinado assim não é o mesmo determinado minimizando PRESS. Os valores das variâncias acumuladas em cada variável latente, para as variáveis dependentes e independentes são mostrados na Tabela 3.8. É importante notar que o número de variáveis latentes selecionadas acumula uma variância expressiva do conjunto de dados originais, principalmente das variáveis dependentes. As variáveis independentes não têm toda sua variância capturada. Isso não é desejável se houver informação não correlacionada as variáveis dependentes nas variáveis latentes mais altas.

As dezesseis variáveis selecionadas no modelo com melhor capacidade de previsão foram os momentos de dipolo $\mu_x$, $\mu_y$ e $\mu_z$; os momentos de quadrupolo $q_{xx}$, $q_{zz}$ e $q_{xz}$; os campos elétricos nas posições atômicas $\nabla(3)$, $\nabla(8)$ e $\nabla(11)$; as cargas líquidas calculadas pelo método CHELPG nas posições atômicas $Q(6)$, $Q(11)$ e $Q(13)$; as densidades eletrônicas em HOMO sobre os átomos $F^H(1)$, $F^H(4)$ e $F^H(17)$; a energia do orbital $\varepsilon_{HOMO-10}$.

Uma vez estabelecido o conjunto de variáveis independentes com o maior poder de modelagem e o número ideal de variáveis latentes, o modelo foi testado na sua capacidade de previsão de conjuntos externos. Estes conjuntos externos foram selecionados aleatoriamente do total de vinte e oito compostos, com 15% do número total de compostos. Para cada conjunto externo foi realizado uma modelagem dos valores das atividades dos compostos restantes (85% do número total), e este modelo assim concebido foi utilizado para estimar a atividade do conjunto de compostos externo. A capacidade de previsão varia para os conjuntos. Os valores máximo e mínimo do coeficiente de correlação entre os valores estimados para cada conjunto de validação externa foram considerados uma estimativa da capacidade de prever dados externos do modelo obtido. Estes indicadores

levam em consideração tanto os melhores quanto os piores resultados.

Para o melhor conjunto de variáveis descritoras foi realizada uma análise por componentes principais (PCA). Os *scores* das PCs que acumulavam cerca de 90% da variância total da variáveis independentes foram utilizados como valores de entrada em redes neurais de retropropagação. Isto corresponde aos *scores* das nove primeiras PCs listadas na Tabela 3.9.

Tabela 3.9: Variâncias percentuais acumuladas em cada uma das nove primeiras PCs da PCA e variância percentual total acumulada até a determinada PC. Os valores correspondem ao modelo utilizando as dezesseis variáveis descritoras selecionadas para os compostos indolil-guanidinotiazóis da Tabela 3.7

| PC # | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 |
|---|---|---|---|---|---|---|---|---|---|
| Variância | 27,93 | 18,3 | 11,26 | 8,99 | 8,55 | 5,97 | 5,05 | 3,74 | 2,97 |
| Total | 27,93 | 46,32 | 57,58 | 66,56 | 75,11 | 81,08 | 86,13 | 89,87 | 92,84 |

As redes foram construída com dez neurônios na primeira camada, o número de *scores* utilizado e mais um de *bias*. Na segunda camada foram utilizados o dobro de neurônios da primeira camada. Na terceira camada foi utilizado um neurônio, que após o treinamento respondia com o valor da atividade de validação cruzada *leave one out*, da rede treinada. Os valores dos demais parâmetros de ajuste da rede neural foram mantidos constantes para todos os modelos. Apenas os limites de tolerância para cada ciclo de retropropagação das redes foram variados. Mais alguns detalhes da construção de cada BPN estão listados na Tabela 3.10

Tabela 3.10: Detalhes das BPNs construídas com os *scores*da PCA das variáveis selecionadas com método PLS e resultados obtidos para cada modelo obtido

| | RN5 | RN6 | RN7 |
|---|---|---|---|
| 1ª Camada | 10 | 10 | 10 |
| 2ª Camada | 20 | 20 | 20 |
| Limite Erro | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ |
| PRESS | 727,1 | 633,9 | 669,0 |
| $q^2$ | 0,13 | 0,25 | 0,20 |
| SDEP | 6,7 | 6,3 | 6,5 |
| n | 28 | 28 | 28 |

## 3.4.2 Resultados e discussão

Os principais resultados que podem ser obtidos com os métodos de regressão multivariada são as correlações entre os conjuntos de variáveis dependentes e os valores previstos para estas variáveis pelos modelos obtidos [132]. As previsões são listadas na Tabela 3.11.

Os valores para o ajuste de todos os dados através de regressão PLS (sem validação cruzada) são mostrados na coluna 'AJ2'. O valor de coeficiente de correlação (r=0,73)

Tabela 3.11: Valores dos indicadores da qualidade de regressão e de capacidade de previsão dos modelos com BPN e PLS para os compostos indolil-guanidinotiazóis obtidos com as substituições da Tabela 3.7 na Figura 3.10

| | Modelos BPN | | Modelos PLS | | | |
|---|---|---|---|---|---|---|
| Modelo | AJ1 | LO4 | AJ2 | LO5 | EV3 | EV4 |
| r | 0,99 | — | 0,73 | — | — | — |
| s | 0,77 | — | 4,22 | — | — | — |
| F | 85,0 | — | 5,07 | — | — | — |
| PRESS | — | 634 | — | 369,5 | 0,05 | 0,15 |
| $q^2$ | — | 0,25 | — | 0,68 | 0,69 | 0,66 |
| SDEP | — | 6,3 | — | 3,92 | 0,23 | 0,86 |
| n | 28 | 28 | 28 | 28 | 5 | 5 |
| #VLs | — | — | 5 | 5 | 5 | 5 |
| #Descritores | 10 | 10 | 16 | 16 | 16 | 16 |

mostra que o modelo foi capaz de reproduzir o conjunto de dados. A estimativa do desvio padrão (s≈4) mostra que estes resultados não são capazes de distinguir o mais ativo, entre os compostos de alta atividade. O modelo tem capacidade para distinguir entre compostos de baixa atividade e de alta atividade. A razão entre variâncias (F=5,07) mostra há significância estatística ($F^{0,995}_{[27;30]}$=2,73; 5>2,73) para esta comparação. Os valores $q^2$ e SDEP para validação *leave one out* no modelo com dezesseis variáveis independentes estão listados na coluna 'LO5' da Tabela 3.11. O valor de $q^2 = 0,68$ para esta análise mostram que o modelo tem boa capacidade de previsão. O valor de SDEP=3,92 mostra que não tem exatidão suficiente para determinar o composto mais ativo num conjunto de amostras, já que na região de maior atividade as variações de concentração são muito menores que o desvio apontado. Os valores obtidos na correlação entre os conjuntos de validação externa (melhor e pior previsões) são mostrados na mesma tabela, nas colunas 'EV3' e 'EV4', respectivamente. Estes valores dão uma idéia da exatidão que pode ser esperada dos modelos para previsão de conjuntos externos. A utilização de mais de uma amostra em validação externa é recomendável para estimar a capacidade de previsão, uma vez que a retirada de apenas uma amostra permite que se tenha a previsão comprometida, caso esta amostra seja atípica (*outliers*). Os resultados de validação externa apresentados mostram que pode-se esperar previsões capazes de distinguir os compostos em dosi grupos facilmente.

Na Tabela 3.12 são mostrados os coeficientes de regressão do modelo final, utilizando cinco variáveis latentes para representar o conjunto de dezesseis variáveis selecionadas. Na última coluna da Tabela 3.12 é mostrado o vetor de regressão que pode ser utilizado para obter uma regressão linear, bastando multiplica-lo pela matriz de dezesseis descritores autoescalados. Os resultados obtidos com treinamento de BPNs não são bons. A rede é capaz de produzir bons resultados com treinamento, para todo o conjunto, como pode se observado na coluna 'AJ1' da Tabela 3.11. Estes resultados não se consumam na validação cruzada. O resultado da coluna 'LO4' da Tabela 3.11 e os resultados da Tabela 3.10 mostram isto. O *overfitting* das redes ocorre se o limite de erro (ver Equação 1.88)

Tabela 3.12: Variáveis selecionadas pela análise dos coeficientes de regressão do modelo PLS nas variáveis latentes. São mostradas as variáveis, os coeficientes em cada uma das cinco variáveis latentes utilizadas e a soma dos coeficientes conforme as Equações 1.86 da página 42.

| Variáveis | Coeficientes na variáveis latentes | | | | | Vetor de |
|---|---|---|---|---|---|---|
| Selecionadas | VL 1 | VL 2 | VL 3 | VL 4 | VL 5 | Regressão |
| $\mu_x$ | 6,36e-02 | -1,62e-02 | -2,87e-02 | -9,90e-02 | 3,29e-02 | -0,0473 |
| $\mu_y$ | 1,30e-01 | 4,50e-03 | 6,19e-03 | -8,99e-02 | 5,31e-02 | 0,1044 |
| $\mu_z$ | -6,65e-02 | -1,35e-01 | -1,56e-01 | -1,71e-01 | -1,14e-01 | -0,6423 |
| $q_{xx}$ | 1,74e-01 | 1,67e-02 | -2,45e-02 | 3,73e-02 | -7,23e-02 | 0,1315 |
| $q_{zz}$ | -1,98e-01 | -2,60e-02 | 2,85e-02 | 1,88e-02 | 9,18e-02 | -0,0852 |
| $q_{xz}$ | 1,45e-01 | 3,65e-02 | -9,37e-03 | 4,35e-02 | 1,07e-01 | 0,3224 |
| $\nabla(3)$ | 9,62e-04 | 7,16e-02 | -7,84e-03 | -1,84e-02 | 1,93e-01 | 0,2392 |
| $\nabla(8)$ | 8,34e-02 | -1,85e-03 | 2,56e-02 | -1,68e-02 | 1,62e-03 | 0,0919 |
| $\nabla(11)$ | 8,31e-02 | -4,93e-02 | 3,31e-02 | 5,90e-02 | 1,12e-01 | 0,2380 |
| $Q(6)$ | 9,64e-02 | 4,08e-02 | 3,19e-02 | -9,07e-02 | -4,35e-02 | 0,0350 |
| $Q(11)$ | -1,23e-01 | 4,96e-02 | 3,78e-02 | -8,83e-02 | -1,17e-01 | -0,2414 |
| $Q(13)$ | -3,07e-03 | -5,75e-02 | 4,26e-02 | 6,19e-02 | 1,73e-02 | 0,0612 |
| $F^H(1)$ | -1,69e-01 | -2,55e-02 | -5,86e-02 | -1,04e-01 | -4,57e-02 | -0,4033 |
| $F^H(4)$ | 8,39e-02 | 1,19e-01 | 2,03e-02 | -2,05e-02 | -8,94e-02 | 0,1128 |
| $F^H(17)$ | 3,95e-02 | -7,98e-02 | -3,03e-02 | -2,11e-01 | -1,73e-01 | -0,4548 |
| $\varepsilon_{HOMO-10}$ | -1,71e-01 | -1,13e-01 | -5,01e-02 | -2,69e-02 | 1,72e-02 | -0,3445 |

para convergência no processo iterativo é muito pequeno. Com a rede de dez neurônios o melhor resultado é obtido quando o limite de erro é igual $\sigma^{BPN} = 0,001$, mostrado na coluna 'RN6' na Tabela 3.10. Valores menores para o limite de erro diminuem a capacidade de generalização da rede obtida, e pioram os resultados de previsão. O método é mais lento do ponto de vista computacional e não possibilita resultados de previsão tão bons quanto os obtidos por PLS. O melhor resultado obtido com BPN, mostrado na Tabela 3.10 para o modelo denominado 'RN6', tem valor do teste $q^2 = 0,25$. Ainda que valores maiores que 0,3 para $q^2$ já podem ser considerados razoáveis [130], estes modelos não parecem promissores para estabelecer QSAR.

Os descritores que têm grande coeficiente na primeira variável latente na Tabela 3.12 são muito importantes, aqui citados por ordem decrescente de importância: $q_{zz}$, $q_{xx}$, $\varepsilon_{HOMO-10}$ $F^H(1)$, $q_{xz}$ $\mu_y$ e $Q(11)$. Esta primeira variável latente tem cerca de dezenove porcento da variância total do conjunto descritor e cerca de cinqüenta e oito porcento da variância dos dados experimentais. Podemos considerar que os descritores com maior peso nesta variável latente são os que mais influenciam as respostas dadas pelo modelo na etapa de previsão. Aqui temos que cinco dos sete descritores são diretamente relacionados à distribuição de carga das moléculas.

A energia do orbital HOMO-10 é um fator que contribui com peso expressivo, e a superfície de densidade eletrônica deste orbital está mostrada na Figura 3.13 para o composto **igt44**, o mais ativo da série. Na Figura 3.12 é mostrado o mesmo orbital no composto **igt45**, que tem o grupo benzila em R1. Este orbital tem densidade na região da ligação entre os anéis de indolil e tiazol, afetando a liberdade rotacional da ligação. Outras regiões da densidade eletrônica deste orbital não são comuns em um número significativo dos compostos da série, variando bastante em função dos substituintes em R1, R2 e R3. A densidade em HOMO no átomo um $F^H(1)$ pode constituir um ponto de doação de elétrons para a ligação dos compostos desta série na enzima $H^+$,$K^+$-ATPase.

### 3.4.3 Conclusões sobre os indolil-guanidinotiazóis

O método de seleção de conformações por estimativa do poder de modelagem da série nas diferentes conformações parece ter sido adequado. Os inibidores do tipo reversível (SCH28080) e os do tipo irreversível (Omeprazol) tem sua forma ativa como um cátion permanente, que pode ser uma sulfenamida nos benzimidazóis ou a forma protonada (quaternária) do nitrogênio nas imidazopiridinas. Ambas as classes ligam-se à enzima na face luminal, e têm comprimento entre 12 Å e 16 Å [119, 120]. Em outro estudo [121], mostrou-se que a distância entre os resíduos do sítio putativo de ligação é da ordem de 15 Å na conformação ativada de enzima, coerente com a distância entre os extremos dos compostos na sua conformação extendida.

Este método pode ser uma solução para o problema da seleção da conformação bioativa, nos casos em que a diferença de energia entre diferentes confôrmeros não é capaz de orientar a escolha. Em casos onde as diferenças de energia são grandes, a escolha recai sobre aquela de menor energia, pois se um composto liga-se ao sítio receptor numa conformação desfavorável, o processo todo torna-se desfavorável e o composto será um mau ligante.

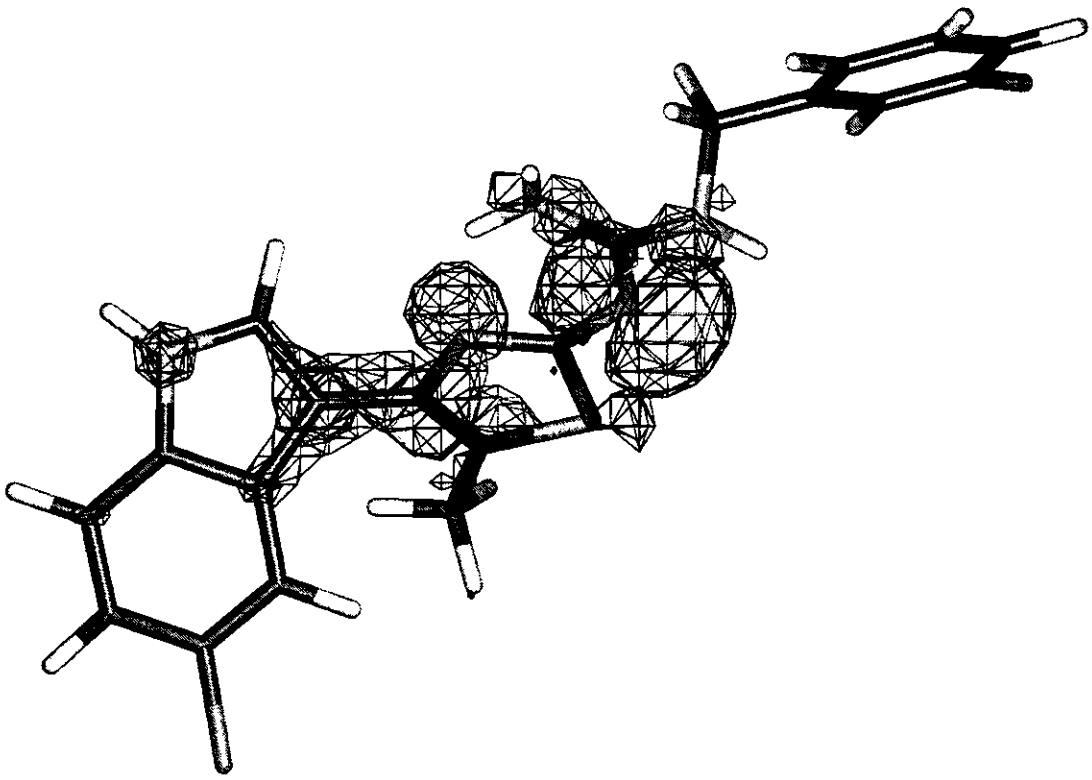


Figura 3.12: Desenho do orbital HOMO-10 do composto **igt45** da Tabela 3.7, mostrando a região do orbital com densidade entre os anéis indolil e tiazol, comum aos compostos da série.

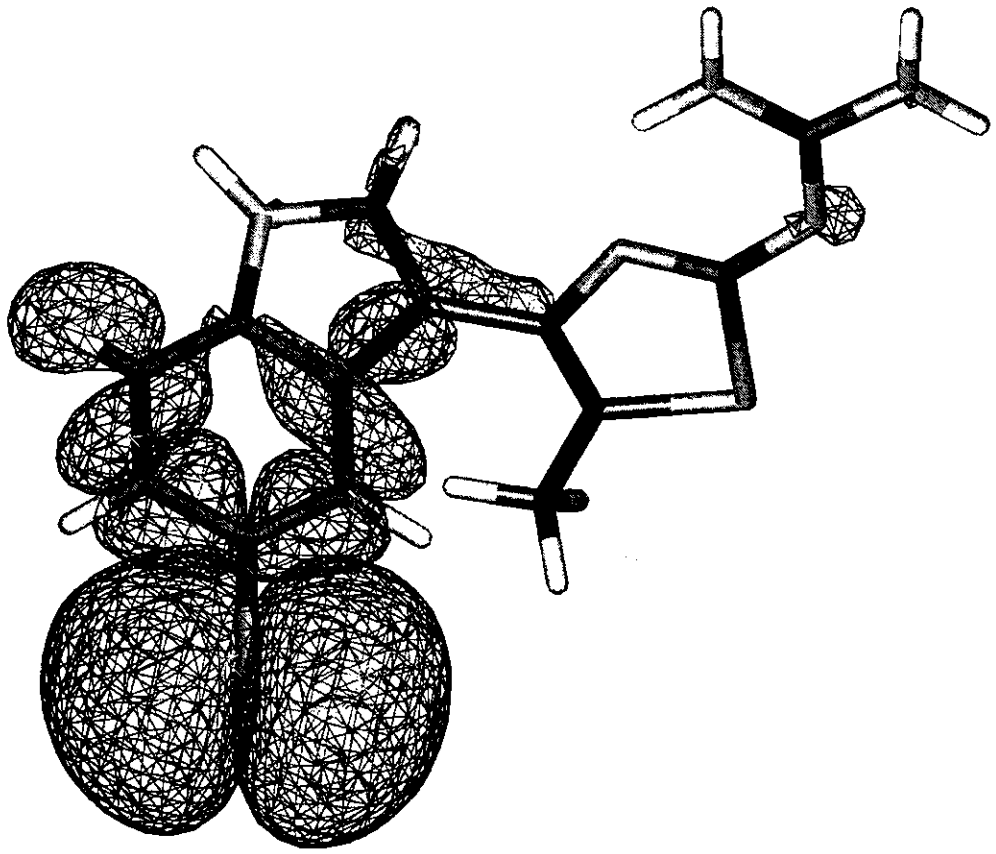


Figura 3.13: Desenho do orbital HOMO-10 do composto **igt44** da Tabela 3.7, mostrando a densidade eletrônica localizada na região da ligação entre os aneis indolil e tiazol em um composto sem o grupo benzil substituído na região de R1. A substituição em R1 pode ser hidrogênio ou benzila.

Tabela 3.13: Inibição da enzima $H^+$,$K^+$-ATPase estimada para compostos indolil-guanidinotiazóis não testados experimentalmente. Os substituintes nas posições R1, R2 e R3 devem ser adicionados ao esqueleto da Figura 3.10 para obtenção dos compostos. Os valores de $IC_{50}$ previstos pelo modelo PLS com dezesseis descritores com cinco variáveis latentes.

| No. | R1 | R2 | R3 | $IC_{50}(\mu\text{mol.l}^{-1})$ |
|---|---|---|---|---|
| **igt29** | $CH_2C_6H_5$ | $CH_3$ | H | 2,3 |
| **igt30** | H | H | 6-$CH_3$ | 2,4 |
| **igt31** | H | H | 5-CN | 2,5 |
| **igt32** | $CH_2C_6H_5$ | H | 5-CN | 2,4 |

O modelo final de regressão multivariada tem capacidade de prever as atividades de novos compostos dentro de uma expectativa de erro que torna impossível distinguir qual dos compostos da Tabela 3.13 seria o mais ativo. Este modelo criado pode ser utilizado para descartar compostos classificados como pouco ativos do processo de síntese.

## 3.5 Conclusões Gerais

Os dados de correlação apresentados, para validação cruzada especialmente, permitem que se utilize estes modelos para estimativa da atividade de compostos pertencentes à mesma série. Esta estimativa é apenas qualitativa, considerando e expectativa de erro apresentada. Ainda assim estes modelos podem ser ferramentas úteis para ajudar na decisão sobre a síntese ou não de um determinado derivado.

As regressões com método PLS ou BPN não permitem que se façam muitas considerações sobre como controlar as variáveis descritoras para melhorar a potência das drogas. Os resultados podem ser analisados em termos das regiões apontadas pelos descritores. Modificações nestas regiões são potencialmente capazes de alterar significativamente a potência das drogas, segundo o modelo. Dada a proposição de um novo derivado, devem-se realizar os cálculos quânticos para este novo composto. O conjunto de dados obtido dos cálculos deve ser utilizado como entrada para o modelo completo, como é feito em validação externa *leave one out*.

# Capítulo 4

# Os compostos inibidores reversíveis da $H^+$, $K^+$-ATPase gástrica da série das Quinolinas

## 4.1 Aspectos da bioquímica dos compostos

Esta série de compostos é relatada como sendo inibidores reversíveis do sítio do potássio da enzima $H^+$,$K^+$-ATPase, ligando-se competitivamente em relação ao potássio à face luminal da $H^+$,$K^+$-ATPase [133]. A reversibilidade na inibição é uma característica bastante importante, porque permite a flexibilidade de resposta farmacológica característica de drogas antagonistas do sítio receptor H2 de histamina, com a acloridria profunda dos inibidores irreversíveis da $H^+$,$K^+$-ATPase. Os compostos desta classe têm como inconveniente o fato de serem nefrotóxicos e metabolicamente instáveis. As modificações em curso, entretanto, permitiram que alguns destes compostos fossem candidatos a agentes anti-secretores em humanos. O composto número **1** da Tabela 4.1 é bem tolerado, e os estudos da fase I de análise farmacológica (testes com voluntários humanos, em pequenas doses) mostraram sua eficácia como agente anti-secreção gástrica em humanos.

Este trabalho busca estabelecer QSAR para uma série de cento e vinte e quatro compostos, divididos em dois subconjuntos que têm pequenas diferenças no esqueleto fundamental. Estes compostos têm os esqueletos fundamentais mostrados nas Figuras 4.1 e 4.2 e os padrões de substituição mostrados nas Tabelas 4.1 e 4.2, referindo-se a cada tipo de esqueleto fundamental. O fato do esqueleto fundamental ter pequenas variações para os subconjuntos de compostos da série introduz um grau maior de dificuldade na elaboração de modelos, porque limita escolha dos descritores ligados às posições atômicas. Para que estes descritores sejam coerentes é necessário que haja estreita correpondência entre as grandezas calculadas em cada posição. Cargas líquidas e densidades eletrônicas foram escolhidas como descritores nas posições atômicas. Para esta série de compostos também não temos informações detalhadas sobre qual o mecanismo de ação, ou resíduos a que se ligam os compostos, apenas uma indicação de que trata-se de inibição competitiva e reversível em relação ao sítio do potássio.

## 4.2 Objetivos

- Estudar as possibilidades de criação de modelos QSAR utilizando descritores de origem quântica.
- Avaliar a qualidade da descrição dos sistemas químicos com os métodos quânticos.
- Obter modelos capazes de estimar corretamente as atividades dos compostos.

## 4.3 Métodos e procedimentos adotados

Para esta série os compostos **11** e **25** foram escolhidos para busca conformacional, porque são os mais ativos de cada subgrupo da série. Os métodos sistemático e aleatório foram utilizados para percorrer o espaço conformacional.

Para o composto **11** foi utilizado método sistemático para busca conformacional. Os ângulos de diedro das ligações C(8)—N(11) e N(11)—C(12) foram variados conjuntamente (numeração mostrada na Figura 4.3), e a energia de cada uma das geometrias geradas foi calculada com o método AM1. Também foi variado o ângulo diedro da ligação C(9)—C(18), separadamente, utilizando a conformação de menor energia obtida na primeira etapa. Os valores dos ângulos diedro foram variados com passos de 5°. Nas regiões do espaço conformacional onde há inversão da configuração do nitrogênio foi utilizado um passo $0,5°$ para melhor definir o ponto de inversão.

Também foi feito um estudo conformacional baseado na geração aleatória de conformações para a molécula para o composto **11**. Foram geradas duas mil estruturas utilizando o método da matriz de distâncias métricas, em dois conjuntos de mil estruturas. Os detalhes da geração das estruturas com o programa `distgeom` são os mesmos já utilizados e mostrados na Secção 3.4.1. Foram selecionadas seiscentas e dezessete conformações em um limite de $3,0\,\mathrm{kcal.mol^{-1}}$ à partir da mais estável.

Os resultados da busca conformacional sistemática e aleatória foram coerentes, apontando para uma mesma estrutura de mínimo global. O confôrmero de menor energia encontrado para o composto **11** é mostrado na Figura 4.3, e foi encontrado com a busca aleatória.

Para o composto **25** foi realizada busca sistemática em torno da ligação N(11)—C(12). O confôrmero de menor energia é mostrado na Figura 4.4, assim como a numeração adotada para os átomos. Deve-se observar que os átomos de carbono C(20) e C(21), e de oxigênio O(22) da Figura 4.3 não têm equivalentes na Figura 4.4. As variáveis utilizadas no modelo para os compostos da série das aril-metil-pirrolo-quinolinas são as mapeadas nas seguintes posições: No carbono C(22) mostrado na Figura 4.4 para a posição equivalente ao carbono C(21) da Figura 4.3, no hidrogênio ligado ao carbono C(18) para a posição O(22) e no hidrogênio ligado ao carbono C(19) para a posição C(21). Os hidrogênios não aparecem nas figuras para evitar um maior congestionamento visual.

Para obter todos os compostos da série foram feitas as substituições listadas na Tabela 4.1 nos grupos R1, R2 e R3 da estrutura mostrada na Figura 4.1, usando a conforma-

ção mostrada na Figura 4.3. Também foram feitas as substituições listadas na Tabela 4.2 nos grupos R1, R2 e R3 na estrutura da Figura 4.2, usando a conformação mostrada na Figura 4.4. Estas geometrias assim obtidas foram todas completamente otimizadas com método semi-empírico AM1.

As geometrias resultantes da otimização das estruturas com AM1 foram submetidas a otimização de geometria com método *ab initio* usando HF(3-21G) e pseudo-potencial tipo SBKJC, para qual usaremos a notação HF(SBKJC-21G). Utilizando HF(3-21G//AM1) e HF(SBKJC-21G) foram obtidas propriedades moleculares para o estudo QSAR dos cento e vinte e quatro compostos listados nas Tabelas 4.1 e 4.2.

Foram calculados cinqüenta e cinco descritores no total. As cargas calculadas com método CHELPG para os vinte e dois átomos numerados conforme mostrado nas Figuras 4.3 e 4.4, a densidade eletrônica total para os mesmos átomos e a energia dos onze orbitais de fronteira entre HOMO-10 e LUMO. Os descritores obtidos foram todos autoescalados para obter o conjunto da variáveis independentes do modelo de regressão multivariada. As atividades foram convertidas para unidades logarítmicas e autoescaladas a seguir, para constituir o conjunto de variáveis dependentes do modelo QSAR. Os valores de atividade previstos nas Tabelas 4.1 e 4.2 foram obtidos com um modelo que utilizou vinte variáveis descritoras selecionadas do conjunto total, com seis variáveis latentes na regressão PLS. Os resultados de atividade previstos com QSAR realizado com os descritores calculados usando HF(SBKJC-21G) são mostrados nas tabelas sob a coluna 'MOD1'.

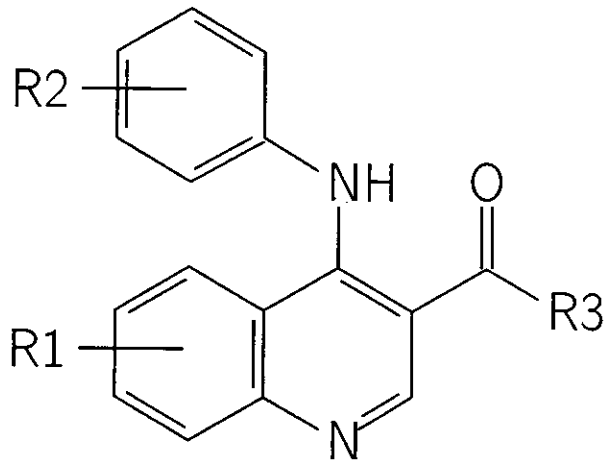



Figura 4.1: Estrutura do esqueleto principal dos compostos da série das acil-arilamino-quinolinas listadas na Tabela 4.1.

Tabela 4.1: Número atribuído para cada composto da série das acil-arilamino-quinolinas e os substituintes que devem ser incluídos no esqueleto da Figura 4.1 para formar cada estrutura. Os valores das atividades experimentais publicadas [133–135] e os valores estimados pelo modelo proposto também são mostrados, em unidades de concentração $\mu$mol.l$^{-1}$ para a inibição de 50% do processo

| | Substituinte | | | $IC_{50}(\mu\text{mol.l}^{-1})$ | |
|---|---|---|---|---|---|
| No. | R1 | R2 | R3 | Exp. | MOD1 |
| **1** | 8-OMe | *o*-Me | *n*-Pr | 1,7 | 1,41 |
| **2** | 8-$OCH_2CH_2OH$ | *o*-Me | *n*-Pr | 2,4 | 1,40 |
| **10** | 6-OH | H | *n*-Pr | 0,24 | 0,31 |
| **11** | 6-OH | *p*-OH | *n*-Pr | 0,071 | 0,16 |
| **12** | 7-OH | H | *n*-Pr | 1,2 | 3,19 |
| **13** | 8-OH | *o*-Me | *n*-Pr | 3,3 | 5,33 |
| **14** | 8-OH | *o*-Me-*p*-F | *n*-Pr | 7,6 | 8,83 |
| **15** | 8-OH | *o*-Me | *i*-Pr | 1,2 | 2,13 |
| **16** | 8-OH | *o*-Me-*p*-F | *i*-Pr | 2,7 | 4,41 |
| **36** | 8-$CH_2OH$ | *o*-Me | Et | 0,38 | 0,84 |
| **37** | 8-$CH_2OH$ | *o*-Me-*p*-OH | Et | 0,18 | 0,31 |
| **38** | 8-$CH_2OH$ | *o*-Me | *n*-Pr | 0,84 | 0,83 |
| **39** | 8-$CH_2OH$ | *o*-Me-*p*-F | *n*-Pr | 2,2 | 1,56 |
| **40** | 8-$CH_2OH$ | *o*-Me-*p*-OH | *n*-Pr | 0,18 | 0,53 |
| **41** | 8-$CH_2OH$ | *o*-Me | *i*-Pr | 0,82 | 0,74 |
| **42** | 8-$CH_2OH$ | *o*-Me-*p*-F | *i*-Pr | 2,1 | 0,81 |
| **43** | 8-CHO | *o*-Me | *n*-Pr | 1,2 | 0,70 |
| **44** | 8-oxiranil | *o*-Me | *n*-Pr | 3.0 | 2,72 |
| **45** | 8-$CH(OH)CHCH_2$ | *o*-Me | *n*-Pr | 0,66 | 1,22 |
| **46** | 8-$CHCHCH_2OH$ | *o*-Me | *n*-Pr | 2,3 | 2,19 |
| **49** | 8-OMe | H | Et | 4,0 | 5,24 |
| **50** | 8-OMe | *o,m*-di-Me | Et | 1,4 | 3,85 |
| **51** | 8-OMe | *o,o'*-di-Me | *n*-Pr | 1,6 | 1,41 |
| **52** | 8-OMe | *o,p,o'*-tri-Me | *n*-Pr | 3,1 | 0,96 |
| **53** | 8-OMe | *m,m'*-di-Me | *n*-Pr | 11,3 | 26,04 |
| **54** | 8-OMe | *o*-Et | *n*-Pr | 0,89 | 1,71 |
| **55** | 8-OMe | *o*-OEt | *n*-Pr | 2,0 | 1,11 |
| **56** | 8-OMe | *o*-OMe | *n*-Pr | 1,5 | 1,11 |
| **57** | 8-OMe | *o,p*-di-OMe | *n*-Pr | 2,1 | 1,12 |
| **58** | 8-OMe | *o*-Me-*p*-OMe | *n*-Pr | 1,27 | 0,99 |
| **59** | 8-OMe | *o*-$CH_2OMe$ | *n*-Pr | 5,5 | 4,2 |
| **60** | 8-OMe | *o*-Me-*m*-$CH_2OH$ | *n*-Pr | 6,0 | 1,8 |
| **61** | 8-OMe | *o*-Me-*m'*-$CH_2OH$ | *n*-Pr | 10,2 | 8,52 |

Continua na próxima página

Tabela 4.1: Continuação

| No. | Substituinte | | | $IC_{50}(\mu mol.l^{-1})$ | |
|---|---|---|---|---|---|
| | R1 | R2 | R3 | Exp. | MOD1 |
| **62** | 8-OMe | *o*-OH | *n*-Pr | 1,3 | 0,83 |
| **63** | 8-OMe | *p*-OH | *n*-Pr | 0,67 | 0,67 |
| **64** | 8-OMe | *o*-Me-*p*-OH | *n*-Pr | 0,21 | 0,35 |
| **65** | 8-OMe | *o,o'*-di-Me-*p*-OH | *n*-Pr | 0,5 | 0,75 |
| **66** | 8-OMe | *o*-Me-*p*-(OCOEt) | *n*-Pr | 19 | 6,17 |
| **67** | 8-OMe | *o,o'*-di-Me-*m,p*-di-OH | *n*-Pr | 2,8 | 3,93 |
| **68** | 8-OMe | *m,p*-di-metilenodioxil | *n*-Pr | 3,6 | 5,81 |
| **69** | 8-OMe | *o,o'*-di-Me-*m,p*-di-metilenodioxil | *n*-Pr | 2,7 | 3,3 |
| **70** | 8-OMe | *o*-Me-*p*-F | *n*-Pr | 2,3 | 1,57 |
| **71** | 8-Me | *o*-Me | *n*-Pr | 3,0 | 2,55 |
| **72** | 8-Me | *o,o'*-di-Me | *n*-Pr | 3,1 | 4,03 |
| **73** | 8-Me | *o*-Et | *n*-Pr | 2,9 | 3,22 |
| **74** | 8-Me | *o*-OMe | *n*-Pr | 2,2 | 1,83 |
| **75** | 8-Me | *o,p*-di-OMe | *n*-Pr | 2,4 | 1,23 |
| **76** | 8-Me | *o*-Me-*p*-OMe | *n*-Pr | 1,7 | 1,01 |
| **77** | 8-Me | *o*-Cl | *n*-Pr | 6,3 | 4,7 |
| **78** | H | *o*-Me | *n*-Pr | 0,97 | 1,95 |
| **79** | 6-$NH_2$ | *o'*-Me | *n*-Pr | 0,16 | 0,27 |
| **80** | 6-OH-8-OMe | H | *n*-Pr | 1,3 | 1,00 |
| **81** | 6-OH-8-OMe | *p*-F | *n*-Pr | 2,1 | 1,46 |
| **82** | 8-$NH_2$ | *o*-Me | *n*-Pr | 6,0 | 3,68 |
| **83** | 8-$CONH_2$ | *o*-Me | *n*-Pr | 1,9 | 1,41 |
| **84** | 8-Ac | *o*-Me | *n*-Pr | 1,5 | 1,62 |
| **85** | 8-Ac | *o*-Me-*p*-OH | *n*-Pr | 0,55 | 0,5 |
| **86** | 8-$CH_2COOH$ | *o*-Me | *n*-Pr | 7,2 | 2,77 |
| **87** | 8-$CH_2COOMe$ | *o*-Me | *n*-Pr | 5,4 | 3,94 |
| **88** | 8-$NHCH_2CH_2OH$ | *o*-Me | *n*-Pr | 7,1 | 4,68 |
| **89** | 8-$O(CH_2)_3COOEt$ | *o*-Me | *n*-Pr | 1,6 | 2,39 |
| **90** | 8-$O(CH_2)_3NMe_2$ | *o*-Me | *n*-Pr | 0,39 | 0,93 |
| **91** | 8-$O(CH_2)_3NMe_2$ | *o*-Me | *n*-Pr | 0,52 | 1,08 |
| **92** | 8-$O(CH_2)_3$-1-piperidinil | *o*-Me | *n*-Pr | 0,5 | 1,19 |
| **93** | 8-$O(CH_2)_3$-1-morfolinil | *o*-Me | *n*-Pr | 0,89 | 1,23 |
| **94** | 8-$O(CH_2)_3$-1-pirrolidinil | *o*-Me | *n*-Pr | 0,51 | 1,13 |
| **95** | 8-$O(CH_2)_3N(Me)CH_2Ph$ | *o*-Me | *n*-Pr | 0,65 | 0,5 |

Continua na próxima página

Tabela 4.1: Continuação

| No. | Substituinte R1 | R2 | R3 | $IC_{50}(\mu\text{mol.l}^{-1})$ Exp. | MOD1 |
|---|---|---|---|---|---|
| **96** | 8-$O(CH_2)_3N$-(Me)$(CH_2)_3$Ph | *o*-Me | *n*-Pr | 0,74 | 0,55 |
| **97** | 8-$NH(CH_2)_3$-1-morfolinil | *o*-Me | *n*-Pr | 2,4 | 2,56 |
| **98** | 8-$NHCH_2CH_2$-(4-imidazolil) | *o*-Me | *n*-Pr | 1,4 | 2,19 |
| **99** | 8-$CH_2$(2-imidazo-[4,5-c]piridil) | *o*-Me | *n*-Pr | 2,8 | 0,78 |
| **100** | 8-$O(CH_2)_3$-CONH(4-piridil) | *o*-Me | *n*-Pr | 2,6 | 1,32 |
| **101** | 8-$NH_2$ | *o,o'*-di-Me | *n*-Pr | 5,8 | 2,76 |
| **102** | 8-$O(CH_2)_3NH$-(2-tiazolil) | *o,o'*-di-Me | Et | 1,7 | 1,06 |
| **103** | 8-$O(CH_2)_3NH$-(2-piridil) | *o,o'*-di-Me | *n*-Pr | 0,9 | 1,75 |
| **104** | 8-$O(CH_2)_3$-(2-benzimidazolil) | *o,o'*-di-Me | *n*-Pr | 1,4 | 1,13 |
| **105** | 8-$O(CH_2)_3NHAc$ | *o,o'*-di-Me | *n*-Pr | 1,8 | 2,39 |
| **106** | 8-$O(CH_2)_3CONH_2$ | *o,o'*-di-Me | *n*-Pr | 2,2 | 2,03 |
| **107** | 8-$O(CH_2)_3CONH$-$C(Me)_2CH_2OH$ | *o,o'*-di-Me | *n*-Pr | 2,8 | 2,26 |
| **108** | 8-$OCH_2CH_2OH$ | *o*-Me | Et | 0,76 | 0,76 |
| **109** | 8-$OCH_2CH_2OH$ | *o*-Me-*p*-F | Et | 2,0 | 2,34 |
| **110** | 8-$OCH_2CH_2OH$ | *o*-Me-*p*-F | *n*-Pr | 2,7 | 2,76 |
| **111** | 8-$OCH_2CH_2OH$ | *o*-Me-*p*-OH | *n*-Pr | 0,21 | 0,54 |
| **112** | 8-$OCH_2CH_2OH$ | *o*-Me | *i*-Pr | 0,99 | 1,77 |
| **113** | 8-$OCH_2CH_2OH$ | *o*-Me-*p*-F | *i*-Pr | 1,6 | 2,48 |
| **114** | 8-$OCH_2CH_2OMe$ | *o*-Me | Et | 0,78 | 0,78 |
| **115** | 8-$OCH_2CH_2OMe$ | *o*-Me-*p*-F | Et | 2,0 | 2,06 |
| **116** | 8-$OCH_2CH_2OMe$ | *o*-Me | *n*-Pr | 2,5 | 1,37 |
| **117** | 8-$OCH_2CH_2OMe$ | *o*-Me-*p*-F | *n*-Pr | 3,5 | 2,48 |
| **118** | 8-$OCH_2CH_2OMe$ | *o*-Me | *i*-Pr | 1,4 | 1,6 |
| **119** | 8-$OCH_2CH_2OMe$ | *o*-Me-*p*-F | *i*-Pr | 1,7 | 2,53 |
| **120** | 8-$(OCH_2CH_2)_2OH$ | H | *n*-Pr | 2,1 | 1,82 |
| **121** | 8-$(OCH_2CH_2)_2OH$ | *o*-Me-*p*-F | *n*-Pr | 4,0 | 3,08 |
| **122** | 8-$(OCH_2CH_2)_3OH$ | *o*-Me | *n*-Pr | 2,4 | 2,09 |
| **123** | 8-$(OCH_2CH_2)_3OH$ | *o*-Me-*p*-F | *n*-Pr | 3,6 | 3,20 |
| **124** | 8-$O(CH_2)_3OH$ | *o,o'*-di-Me | *n*-Pr | 1,21 | 1,09 |

Final

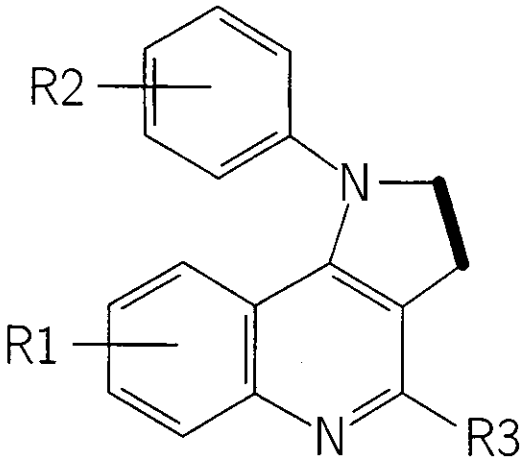



Figura 4.2: Estrutura do esqueleto principal dos compostos da Tabela 4.2. A ligação em negrito é simples ou dupla, conforme indicação da tabela.

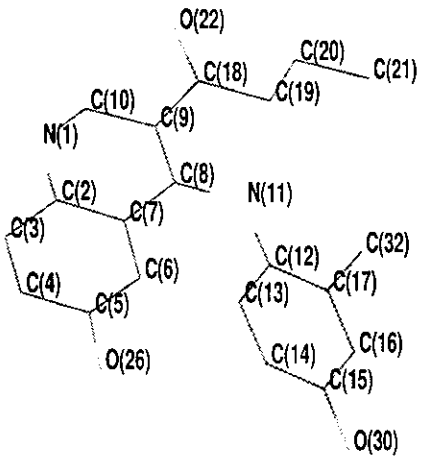



Figura 4.3: A conformação do composto **11** utilizada para construção das geometrias iniciais dos compostos série das acil-arilamino-quinolinas listadas na Tabela 4.1. A numeração corresponde ás posições atômicas das propriedades calculadas.

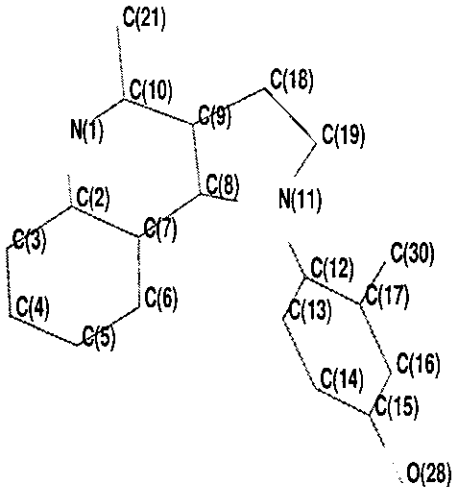



Figura 4.4: A conformação do composto **25** utilizada para construção das geometrias iniciais dos compostos da série das aril-metilpirrolo-quinolinas listadas na Tabela 4.2. A numeração corresponde ás posições atômicas das propriedades calculadas.

Tabela 4.2: Nome atribuído aos compostos da série dos aril-metilpirrolo-quinolinas, e os respectivos substituintes nas posições R1, R2 e R3 para o esqueleto da Figura 4.2. Os valores das atividades biológicas experimentais e estimados com o modelo 'MOD1' também são mostrados, em unidades de $IC_{50}$.

| Comp. | Substituinte | | | Ligação | $IC_{50}(\mu mol\ l^{-1})[\pm\sigma]$ | |
|---|---|---|---|---|---|---|
| No. | R1 | R2 | R3 | dupla | Exp. | MOD1 |
| **3** | $OCH_3$ | *o*-$OCH_3$ | $NH_2$ | não | 0,44[0,03] | 0,21 |
| **4** | $CH_3$ | *o*-$CH_3$ | $NH_2$ | não | 0,22 | 0,23 |
| **5** | $OCH_3$ | *o*-$OCH_3$ | H | não | 1,29 | 1,52 |
| **6** | $CH_3$ | *o*-$CH_3$ | H | não | 0,98 | 1,24 |
| **7** | H | *o*-$CH_3$ | H | sim | 9,8[0,3] | 26,85 |
| **8** | | Ver Figura 4.5 | | | 69 | 44,2 |
| **9** | | Ver Figura 4.3 | | | 22 | 21,3 |
| **17** | $CH_3$ | *o*-$CH_3$ | $NHCH_3$ | não | 0,18[0,02] | 0,26 |
| **18** | $CH_3$ | *o*-$CH_3$ | $N(CH_3)_2$ | não | 1,03 | 0,49 |
| **19** | $OCH_3$ | H | $CH_3$ | não | 2,4[0,1] | 3,98 |
| **20** | $OCH_3$ | *o*-$OCH_3$ | $CH_3$ | não | 1,6[0,2] | 1,00 |
| **21** | $OCH_3$ | *o*-$CH_3$ | $CH_3$ | não | 0,66[0,11] | 1,41 |
| **22** | $CH_3$ | *o*-$CH_3$ | $CH_3$ | não | 0,42[0,03] | 0,91 |
| **23** | H | *o*-$CH_3$ | $CH_3$ | não | 0,41 | 1,41 |
| **24** | H | *o*-$CH_3$,*p*-$OCH_3$ | $CH_3$ | não | 0,53 | 1,41 |
| **25** | H | *o*-$CH_3$,*p*-OH | $CH_3$ | não | 0,17 | 0,19 |
| **26** | H | *o,o'*-di-$CH_3$ | $CH_3$ | não | 1,5 | 0,91 |
| **27** | $OCH_3$ | *o*-$CH_3$,*p*-$OCH_3$ | $CH_3$ | não | 1,0 | 0,69 |
| **28** | F | *o*-$CH_3$ | $CH_3$ | não | 0,75 | 0,23 |
| **29** | OH | *o*-$CH_3$ | $CH_3$ | não | 1,1 | 1,10 |
| **30** | $CH_3$ | *o*-$CH_3$ | $NHCH_3$ | sim | 0,29 | 0,13 |
| **31** | $CH_3$ | *o*-$CH_3$ | $NH(CH_3)_2OH$ | sim | 0,17 | 0,38 |
| **32** | $CH_3$ | *o*-$CH_3$ | $NH(CH_3)_3OH$ | sim | 0,41 | 0,43 |
| **33** | $CH_3$ | *o*-$CH_3$ | $NH(CH_3)_4OH$ | sim | 0,22 | 0,39 |
| **34** | $CH_3$ | *o*-$CH_3$ | $NHCONHCH_3$ | não | 4,97 | 1,84 |
| **37** | $CH_3$ | *o*-$CH_3$ | $NHCONH_2$ | não | 1,99 | 2,69 |
| **47** | $CH_3$ | *o*-$OCH_3$,*p*-$CH_3$ | $NHCH_3$ | não | 0,28 | 0,26 |
| **48** | $OCH_3$ | *o*-$CH_3$ | $NHCH_3$ | não | 0,25 | 0,34 |

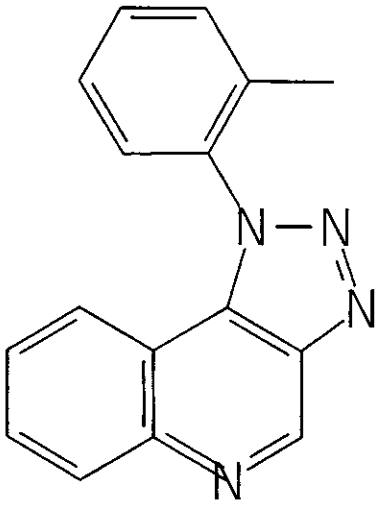



Figura 4.5: Estrutura do composto **8** da série dos aril-metil-quinolínicos listados na Tabela 4.2.

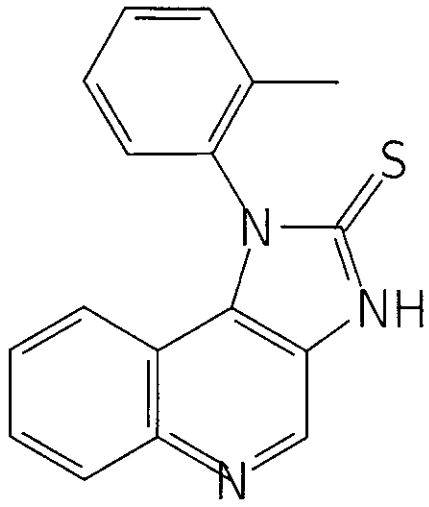



Figura 4.6: Estrutura do composto **9** da série dos metil-aril-quinolínicos listado na Tabela 4.2

## 4.4 Resultados da modelagem e relações entre as estruturas e as atividades

O método escolhido para criação do modelo de regressão foi o de Mínimos Quadrados Parciais (PLS). Para determinar o número de variáveis latentes que resultam no modelo que melhor descreve os dados experimetais foi utilizado o critério do Mínimo Erro Quadrado de Previsão (PRESS) em validação cruzada. A validação cruzada foi realizada para conjuntos aleatórios de dez compostos, que foram eliminados do treinamento em vinte vezes blocos, de maneira que todos os compostos ficaram de fora do conjunto de treinamento pelo menos uma vez. As condições que resultam no menor valor de PRESS foram utilizadas para dar prosseguimento ao trabalho.

A seleção de variáveis foi realizada por inspeção. Com o número escolhido de variáveis latentes, os coeficientes de regressão das variáveis latentes foram utilizado para selecionar as variáveis descritoras que são utilizadas no modelo. Descritores com pequeno valor absoluto dos coeficientes foram eliminados, e o procedimento de determinação do número ótimo de variáveis latentes por minimização de PRESS foi repetido.

Um novo modelo é testado, uma nova matriz de coeficientes é calculada e os descritores com coeficientes pequenos nas variáveis latentes que capturam percentuais altos de variância são eliminados. Repete-se este ciclo até que o valor de PRESS total aumente com a retirada de variáveis. Determina-se o número razoável de variáveis latentes usando o modelo final, levando-se em conta o número de amostras, variáveis independentes e capacidade de previsão. Este modelo usa seis variáveis latentes para representar as vinte variáveis descritoras finalmente selecionadas. O modelo foi usado para validação cruzada de grupos de dez compostos para obter os valores de previsão listados na coluna 'MOD1' das Tabelas 4.1 e 4.2.

A análise dos indicadores de qualidade de regressão é mostrada na Tabela 4.5. Para o modelo obtido com dados calculados com HF(SBKJC-21G) os indicadores são mostrados na coluna 'MOD1'. Nas colunas 'MOD2' e 'MOD3' temos valores de erros médios quadrados em validação cruzada em bloco (RMSECV) e *leave one out* (RMSELOCV) para modelos intermediários. Estes modelos intermediários foram utilizados na seleção de variáveis e eliminação de alguns compostos do conjunto. Os indicadores de qualidade do modelo utilizando dados obtidos com HF(3-21G//AM1) são mostrados na coluna 'MOD4'. Podemos observar que os erros em validação cruzada para os modelos 'MOD1' e 'MOD4' são muito próximos. O método de escolha foi HF(SBKJC-21G) por simplificação dos cálculos geometria. A utilização de método semi-empírico pode ser vantajosa se os recursos computacionais forem insuficientes para completar os cálculos *ab initio*.

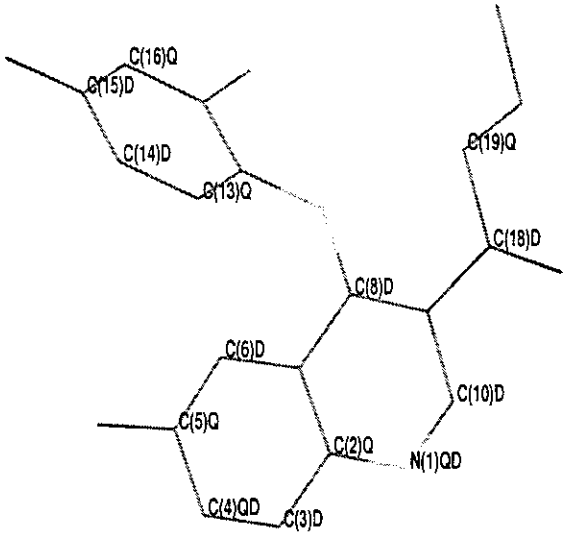



Figura 4.7: Estrutura do composto **11** , mostrando as variáveis ligadas aos átomos selecionadas para o modelo QSAR. A figura da esquerda mostra a letra 'Q' ao lado de cada átomo cuja carga líquida selecionada no modelo final, e a letra 'D' no caso em que a densidade eletrônica total foi selecionada.

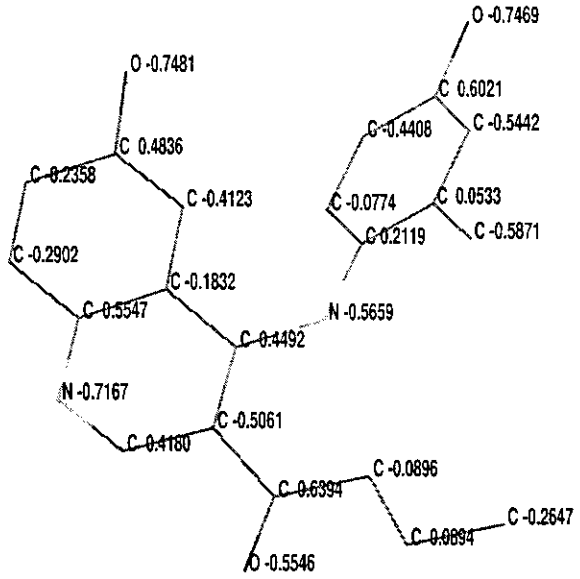



Figura 4.8: Cargas líquidas calculadas com o método CHELPG para o composto **11** que é o mais ativo da série no subgrupo das aril-acil-amino-quinolinas.

As variáveis selecionadas para as posições atômicas são mostradas na Figura 4.7 para o composto **11**, e as cargas nos átomos pesados da molécula na Figura 4.8, para o mesmo composto. As energias dos orbitais do composto **11** selecionadas no modelo final são as seguintes: $\epsilon_{HOMO-8} = -0,4581\,\text{eV}$, $\epsilon_{HOMO-6} = -0,4303\,\text{eV}$, $\epsilon_{HOMO-4} = -0,3902\,\text{eV}$ e $\epsilon_{HOMO} = -0,2968\,\text{eV}$. Estes orbitais moleculares do composto **11** estão desenhados nas Figuras 4.9, 4.10, 4.11 e 4.12.

O desenho dos lóbulos de HOMO sobre a estrutura da molécula é mostrado na Figura 4.9. Podemos ver em primeiro plano os átomos do anel benzênico. Nas posições C(14) e C(15), conforme a numeração da Figura 4.3, as densidades eletrônicas do HOMO

e as cargas nos átomos de carbono C(13), C(15) e C(16) ($Q_{C(13)} = -0,0774; Q_{C(15)} = 0,6021; Q_{C(16)} = -0,5442$ no composto **11**, ver Figura 4.8) sugerem que o potencial eletrostático pode direcionar a região da molécula para um sítio onde os lóbulos do orbital possam interagir adequadamente com o receptor. Estas variáveis também podem agir apenas como indicadoras do tipo da substituição na posição R2 da molécula. Uma interpretação inequívoca do significado destas variáveis relevantes no modelo depende do conhecimento do mecanismo detalhado da ligação da droga ao receptor. Sem conhecer estes detalhes, apenas suposições podem ser feitas [136].

Na Figura 4.10 é mostrado o orbital HOMO-4. Este orbital tem lóbulos grandes na região do nitrogênio N(1) e da carbonila C(18)=O(22). No nitrogênio N(1) temos que tanto a densidade como a carga calculada são relevantes no modelo criado, enquanto na região da carbonila temos que a carga em C(19) (carbono $\beta$) é relevante assim como a densidade eletrônica em C(18). Uma interpretação é que a região do nitrogênio participa com elétrons em algum tipo de ligação com o receptor, e que a acidez dos hidrogênios ligados ao carbono $\beta$ pode ser importante. A acidez de um próton está ligada a polarização da ligação ao átomo pesado [137]. O orbital HOMO-6 do composto **11** da Tabela 4.1 é mostrado na Figura 4.11. Este orbital tem seus principais lóbulos na mesma região que o HOMO-4, e podemos considerar válida a mesma interpretação já dada no caso do orbital HOMO-4. Nesta região da molécula foram obtidas propriedades sobre diferentes átomos para as moléculas dos dois diferentes subgrupos para compor o quadro dos descritores, conforme explicado anteriormente.

A Figura 4.12 mostra os lóbulos do orbital HOMO-8 do composto **11**. Sua importância deve-se ao fato de ter densidade na região das ligações que unem o anel de quinolina com o grupo fenil, na região do nitrogênio N(11). Esta densidade eletrônica localizada nesta região altera o perfil da barreira de energia potencial para rotacionar as ligações C(8)—N(11)—C(12). Sua interpretação está ligada à restrição da posição do grupo fenil na molécula.

No anel de quinolina há alternância entre a importância de cargas (Q) e densidades (D). Traçando um eixo imaginário entre os átomos C(5) e C(10) na Figura 4.7, pode-se observar uma alternância Q-D-Q-D no eixo ao longo do eixo imaginário. As cargas são importantes na região dos átomos C(5); C(4); C(2) e N(1). Há polarização nas ligações ($Q_{C(5)} = -0,4836 \mapsto Q_{C(4)} = 0,2358$ ; $Q_{N(1)} = -0,7167 \mapsto Q_{C(2)} = 0,5547$), bastante elevada no caso da ligação N(1)—C(2). As densidades eletrônicas em C(6) e C(10) são expressivas em HOMO, como pode ser observado na Figura 4.9. O átomo de nitrogênio N(1) tem densidade expressiva em HOMO-4 e HOMO-6. Trata-se de um possível sítio doador de elétrons, por suposição.

O modelo QSAR obtido para esta série de compostos, utilizando as vinte variáveis independentes descritas agrupadas em seis variáveis latentes, foi denominado 'MOD1'. A variância acumulada em cada uma das variáveis latentes (VL) está listada na Tabela 4.3. Mostra uma aquisição razoável da informação das variáveis descritoras, e uma boa aquisição da variância para as atividades. Os coeficientes obtidos nas variáveis latentes para cada variável descritora e o vetor de regressão estão listados na Tabela 4.4.

Os resultados de regressão para esta série de compostos mostram valores razoáveis dos indicadores de qualidade de regressão. Deve-se levar em conta que o conjunto é numeroso,

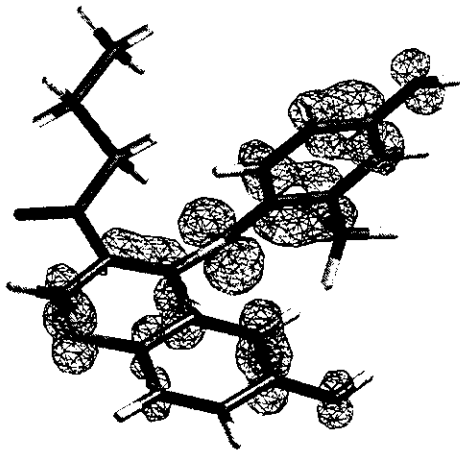


Figura 4.9: Orbital HOMO do composto **11**, mostrando a densidade eletrônica na região do anel de benzeno em primeiro plano.

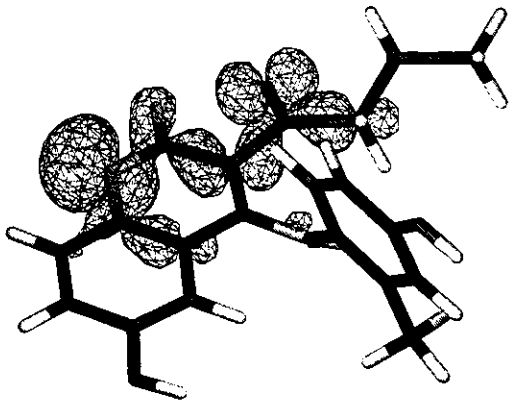


Figura 4.10: O orbital HOMO-4 do composto **11** tem densidade eletrônica expressiva na região dos átomos de nitrogênio do anel de quinolina e na região da carbonila.

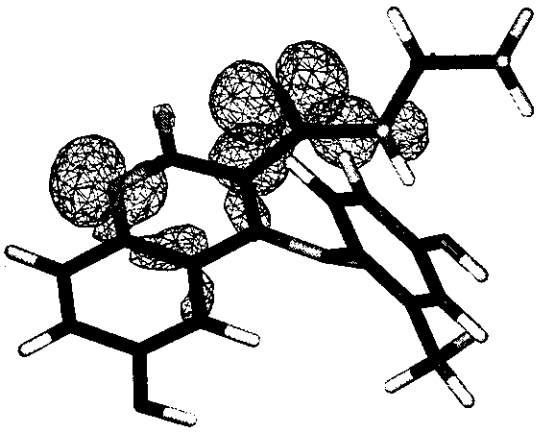


Figura 4.11: Orbital HOMO-6 do composto **11** tem lóbulos parecidos aos do orbital HOMO-4.

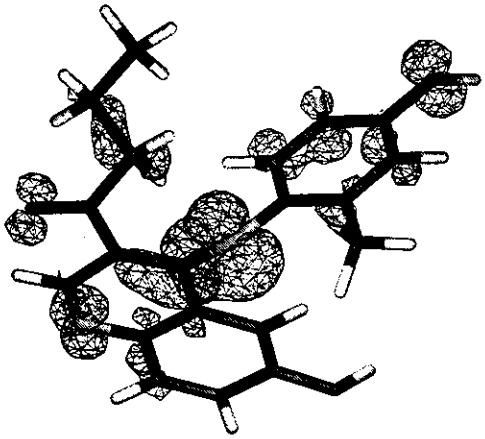


Figura 4.12: O orbital HOMO-8 do composto **11** mostra densidade eletrônica na região de ligação entre os anéis, relacionando-se a liberdade rotacional das ligações.

Tabela 4.3: Tabela com as variâncias acumuladas pelas seis variáveis latentes do modelo final com o conjunto de variáveis descritoras selecionadas autoescalados.

| | Independentes | | dependentes | |
|---|---|---|---|---|
| VL# | Esta | Total | Esta | Total |
| 1 | 11,68 | 11,68 | 46,76 | 46,76 |
| 2 | 15,23 | 26,92 | 10,89 | 57,65 |
| 3 | 10,91 | 37,83 | 9,26 | 66,91 |
| 4 | 4,32 | 42,15 | 12,27 | 79,18 |
| 5 | 9,15 | 51,30 | 2,41 | 81,59 |
| 6 | 7,88 | 59,18 | 1,33 | 82,91 |

dividido em dois subconjuntos com mudanças no esqueleto fundamental. Na Tabela 4.5 são mostrados valores da qualidade de alguns modelos preliminares e do modelo final denominado 'MOD1'. Na Figura 4.13 é mostrado um gráfico com os valores experimentais e previstos, em unidade de $\log IC_{50}$. Alguns compostos com previsão ruim e estrutura química com grupos substituintes muito volumosos e diferenciados foram eliminados do modelo e das Tabelas 4.1 e 4.2.

Na Figura 4.14 podemos ver o gráfico dos indicadores de dispersão do conjunto de compostos no subespaço das variáveis latentes. Os valores de $Q$ e $T^2$ para cada composto, assim como os limites calculados para estes parâmetros foram desenhados. Na Figura 4.15 podemos observar os valores dos desvios padrão da estimativa da atividade de cada composto em relação ao valor experimental, assim como o *leverage* de cada uma das amostras. Os gráficos das Figuras 4.14 e 4.15 foram utilizados conjuntamente para definir os compostos que poderiam ser eliminados do conjunto de treinamento do modelo. Valores maiores que os limites nos indicadores de dispersão acompanhados por desvios maiores que $2,5\,\sigma$ foram considerados indicações para eliminação dos compostos do modelo. Um aumento da dispersão de uma dada amostra pode ser compensado pela introdução de um maior número de amostras com características semelhantes, caso contrário teremos erros de previsão para o compostos com as tais características. Um alto valor de *leverage* na Figura 4.15 significa que um dado composto tem alta influência na construção do modelo. Amostras com valores nos extremos da escala geralmente têm valores altos de *leverage*. O caso dos compostos **7**, **8** e **9** é exemplar. Suas estruturas são distintas, seus valores de atividade situam-se no extremo inferior da escala de atividade, têm valores altos dos indicadores de dispersão (porquê têm estruturas diferenciadas), porém sua estimativa é boa e foram mantidos na criação do modelo final.

## 4.5 Conclusões

Para esta classe de compostos o modelo proposto é capaz de prever aceitavelmente a atividade dos compostos em validação cruzada. O conjunto de variáveis descritoras calculadas e utilizadas na previsão tem relação bastante direta com elementos estruturais, constituindo um ponto de partida para um processo de alinhamento desta classe de dro-

Tabela 4.4: Coeficientes de regressão de cada variável latente do modelo de regressão PLS 'MOD1' e vetor de regressão, obtidos segundo as Equações 1.86 da página 42.

| Variável | Variávies Latentes | | | | | | Vetor de |
|---|---|---|---|---|---|---|---|
| Descritora | #1 | #2 | #3 | #4 | #5 | #6 | Regressão |
| Q(1) | 0,1555 | 0,0624 | -0,0343 | -0,0642 | -0,0263 | -0,0321 | 0,0610 |
| Q(2) | -0,0047 | -0,0875 | -0,0432 | -0,0551 | 0,0081 | -0,0121 | -0,1944 |
| Q(4) | -0,0102 | -0,0422 | -0,0212 | -0,1511 | 0,0018 | -0,0174 | -0,2402 |
| Q(5) | -0,0814 | -0,0970 | -0,1337 | -0,1058 | -0,0362 | -0,0004 | -0,4544 |
| Q(13) | -0,0214 | 0,0483 | 0,0973 | 0,1221 | -0,0026 | 0,0055 | 0,2493 |
| Q(16) | 0,1363 | 0,0618 | 0,0654 | 0,0555 | 0,0719 | 0,0449 | 0,4359 |
| Q(19) | -0,0965 | 0,0012 | 0,0676 | -0,0484 | -0,0751 | -0,0612 | -0,2125 |
| D(1) | 0,1893 | 0,0104 | 0,0139 | 0,1693 | 0,0255 | 0,0209 | 0,4294 |
| D(3) | -0,0508 | -0,0014 | 0,1361 | 0,2624 | 0,0817 | 0,0243 | 0,4523 |
| D(4) | 0,0088 | -0,0857 | -0,0849 | -0,2112 | -0,0491 | -0,0850 | -0,5071 |
| D(6) | 0,0393 | -0,0139 | 0,0766 | 0,0628 | 0,0541 | 0,0306 | 0,2495 |
| D(8) | -0,1006 | -0,0303 | 0,0690 | 0,0870 | 0,0139 | 0,0357 | 0,0747 |
| D(10) | 0,0494 | -0,1011 | -0,0759 | -0,1183 | -0,0289 | -0,0094 | -0,2842 |
| D(14) | 0,1627 | 0,0489 | 0,1211 | 0,1840 | 0,0188 | -0,0328 | 0,5028 |
| D(15) | -0,0138 | -0,0241 | -0,0568 | -0,3099 | -0,0165 | 0,0062 | -0,4149 |
| D(18) | 0,2084 | 0,1338 | 0,1298 | 0,0104 | -0,0077 | 0,0058 | 0,4804 |
| $\varepsilon_{HOMO-8}$ | 0,0193 | 0,0484 | 0,0113 | 0,1922 | 0,0773 | 0,0097 | 0,3582 |
| $\varepsilon_{HOMO-6}$ | -0,1137 | 0,0109 | -0,0443 | -0,0780 | -0,0134 | -0,0682 | -0,3067 |
| $\varepsilon_{HOMO-4}$ | -0,1430 | 0,0145 | -0,0152 | -0,0687 | 0,0222 | 0,0128 | -0,1773 |
| $\varepsilon_{HOMO}$ | -0,1750 | 0,0051 | -0,0016 | 0,0508 | 0,0246 | 0,0070 | -0,0892 |

Tabela 4.5: Tabela com os indicadores de qualidade de previsão no modelo final (MOD1) e nos modelos correlatos. RMSECV significa a média quadrática do erro de validação cruzada em bloco, RMSELOCV significa a média quadrática dos erros de validação cruzada tipo *leave one out*. O valor de $q^2$ foi obtido com validação cruzada em blocos de dez compostos.

| Indicador | MOD1 | MOD2 | MOD3 | MOD4 |
|---|---|---|---|---|
| RMSECV | 0,24 | 0,27 | 0,30 | 0,25 |
| RMSELOCV | 0,21 | 0,24 | 0,36 | 0,21 |
| $q^2$ | 0,67 | — | — | — |
| SDEP | 0,26 | — | — | — |
| PRESS | 7,87 | — | — | — |
| Descritores | 20 | 20 | 20 | 20 |
| VLs | 6 | 6 | 6 | 6 |
| Nº Compostos | 124 | 128 | 133 | 124 |

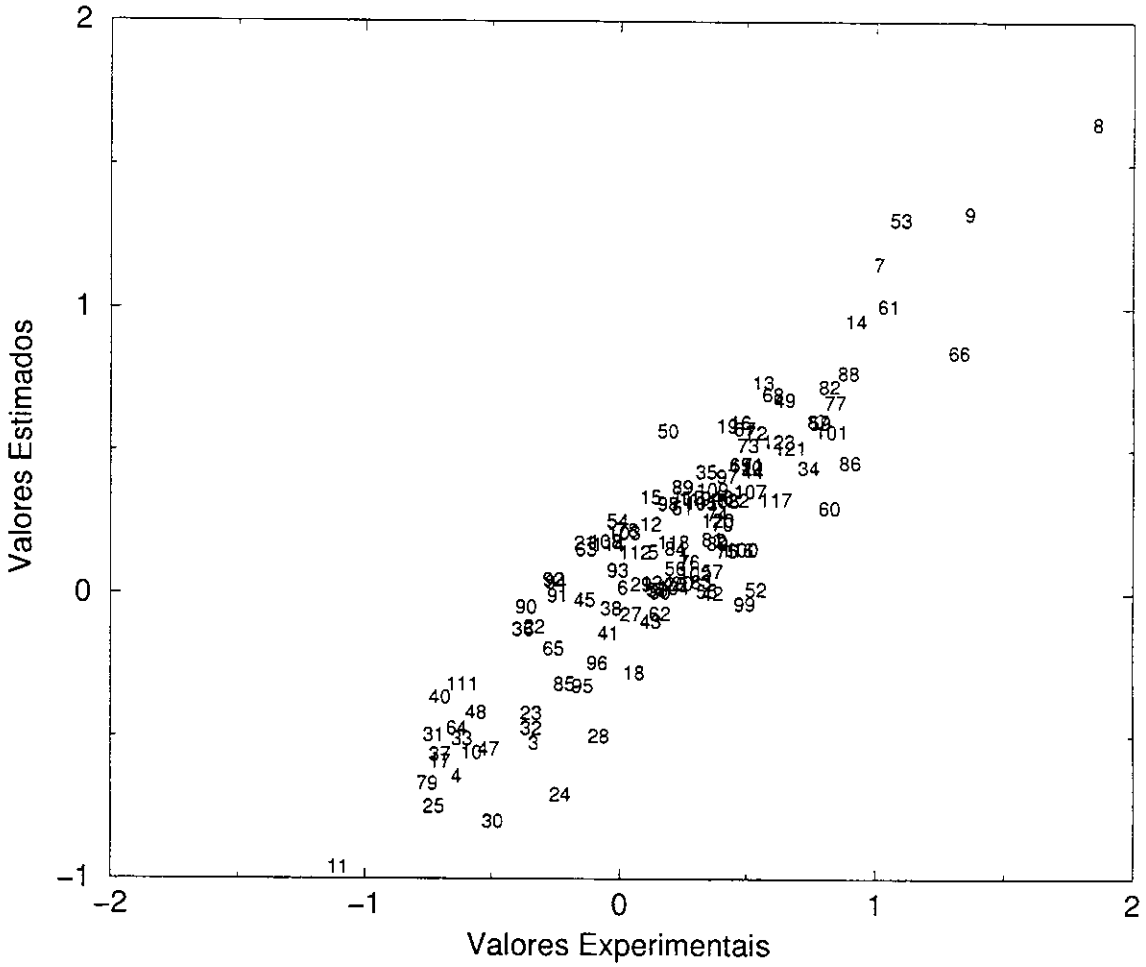



Figura 4.13: Gráfico mostrando os valores de log $IC_{50}$ das atividade biológicas medidas experimentalmente e os valores obtidos com o modelo denominado 'MOD1', obtido com seis variáveis latentes representando os vinte descritores selecionados, para os cento e vinte e quatro compostos listados nas Tabelas 4.1 e 4.2. Os valores foram previstos com o modelo em validação cruzada usando dez compostos retiraddos aleatoriamente uma única vez do conjunto de treinamento.

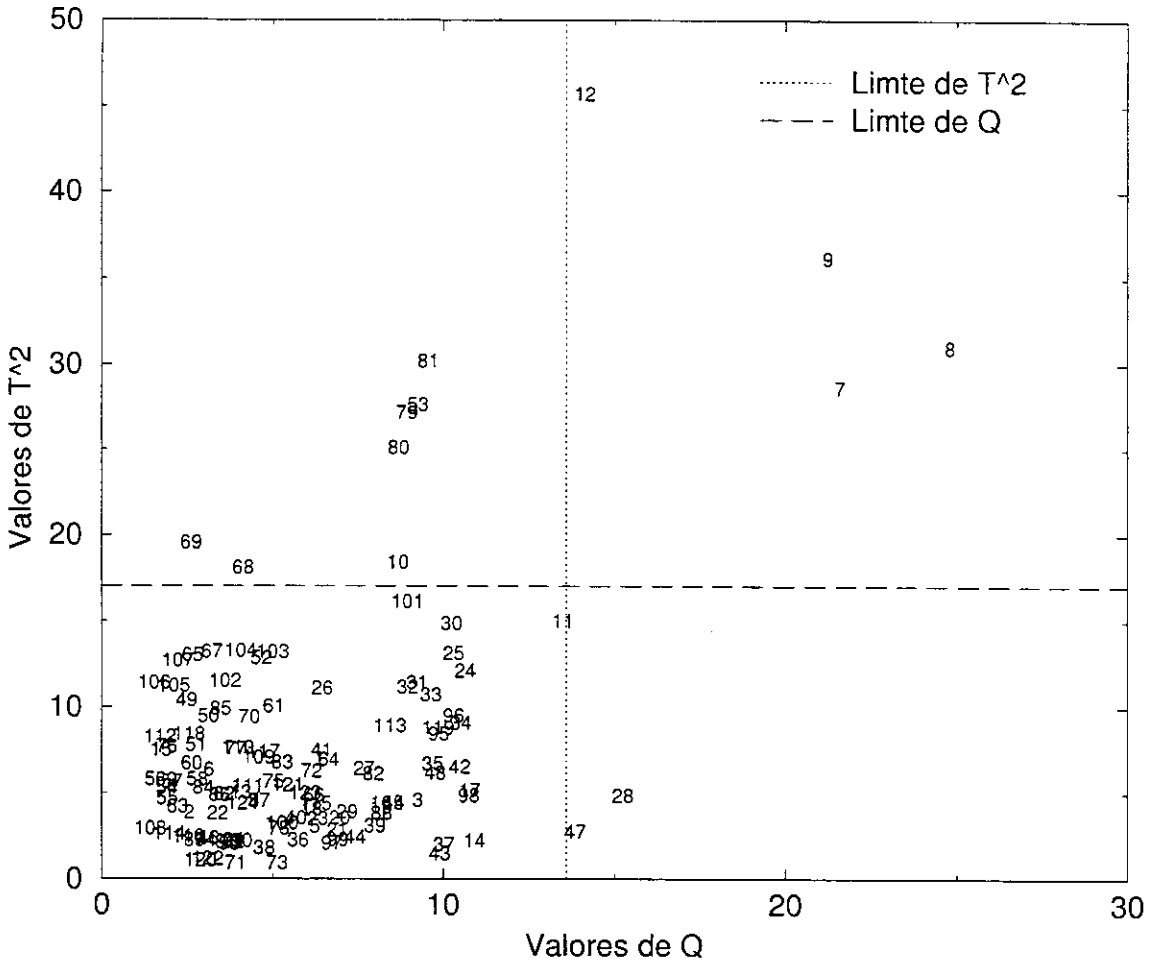



Figura 4.14: Gráfico mostrando os valores dos indicadores de dispersão $Q$ e $T^2$ dos compostos no conjunto de treinamento para o modelo 'MOD1'. Os Valores limite para cada parâmetro são mostrados pelas linhas paralelas aos eixos, e devem ser utilizados para avaliar o quanto cada um dos compostos tem de atípico no conjunto, tendo em vista o compromisso entre uma melhor capacidade previsão e uma grande capacidade de generalização do modelo obtido.

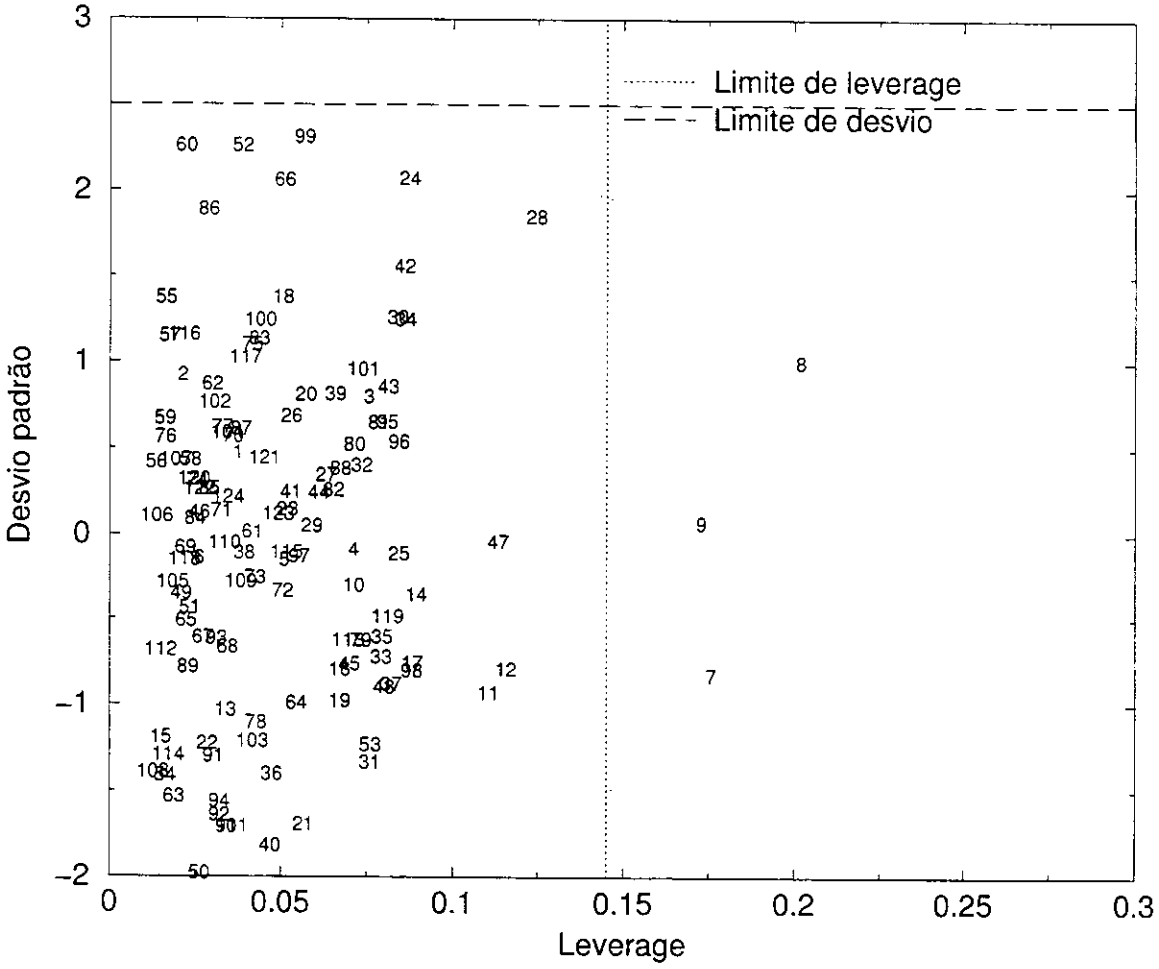



Figura 4.15: Gráfico mostrando os valores dos desvios padrão para cada estimativa de $\log IC_{50}$, e *o leverage* de cada um dos compostos no modelo 'MOD1'.

gas ao sítio receptor. Também pode-se utilizar cálculos mecânico quânticos para estimar a atividade de um novo composto ainda não sintetizado, com restrições. A estimativa de desvio, SDEP= $0,27$, mostrada na Tabela 4.5 não permite a distinção do mais ativo entre compostos com alta atividade, com valor pequeno de $\log IC_{50}$, porém aponta inequivocamente as direção de maior atividade. Na região de menor atividade há maior dispersão dos valores estimados, o que pode ser entendido se consideramos que a ligação de uma droga ao um sítio receptor se dá por motivos específicos. O impedimento da ligação da molécula ao receptor pode ser causado por inúmeros fatores. Ou seja: O conjunto dos compostos que são bons ligantes certamente compartilha propriedades físico-químicas, sobre os demais pouco pode-se afirmar.

# Capítulo 5

# Estudo do mecanismo da reação de eliminação das Nicotinamidas

## 5.1 A conversão da pró-droga na espécie bioativa

Os compostos da classe das 2-[(2,4-Dimetoxibenzil) sulfinil]-N-(4-piridinil) piridina-3-carboxamidas são derivados de Nicotinamidas e foram estudados por Terauchi e colaboradores como possíveis inibidores da secreção gástrica [138–140]. Esses compostos têm um esqueleto principal como o mostrado na Figura 5.1.

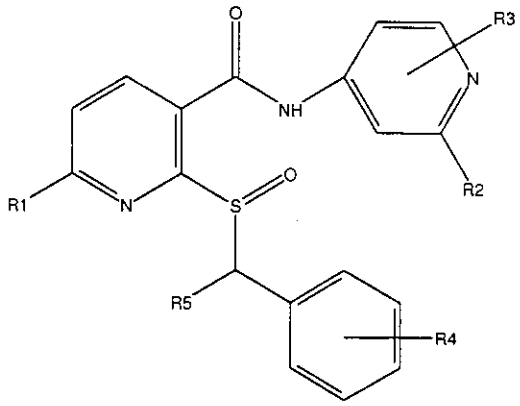



Figura 5.1: Estrutura do esqueleto principal dos compostos da série das nicotinamidas listadas na Tabela 5.1.

Estes compostos produzem inibição irreversível da secreção gástrica, provavelmente atuando na enzima $H^+$,$K^+$-ATPase. Estes compostos agem como pró-drogas, sendo metabolizados e convertidos na forma ativa no meio ácido do estômago, conforme o mecanismo da Figura 5.2. Esta característica é responsável pela alta seletividade destas drogas para a enzima presente nas face luminal da membrana das células apicais da parede estomacal [141]. Acredita-se que estes compostos podem agir no organismo através de um mecanismo semelhante ao de compostos como o Omeprazol, mostrado na Figura 5.3. Podemos deduzir que para ser ativo um composto deve ser capaz de tansforma-se facilmente em

um dos intermediários mostrados na Figura 5.3. Porém a forma ativa reage rapidamente com resíduos de metionina de qualquer proteína.

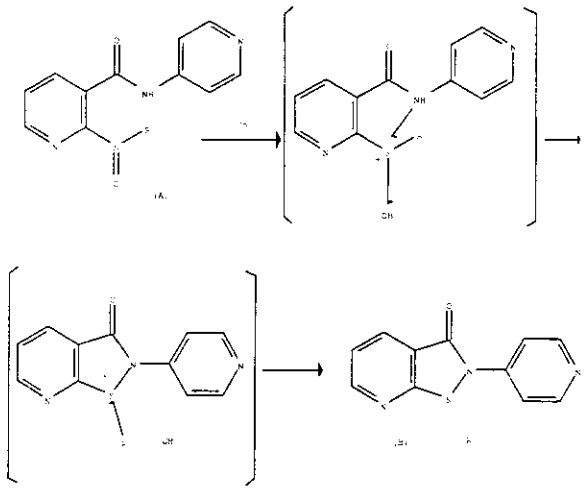


Figura 5.2: Mecanismo de geral de reação dos compostos derivados de Nicotinamida, que reulta na forma ativa para a redução da secreção gástrica destes compostos. O grupo mostrado como $R^+$ na figura corresponde ao 2,4-dimetoxibenzil nas Nicotinamidas. A forma de pró-droga dos composto é mostrada em (A) e a forma ativa em (B).

Para alguns compostos da série de moléculas, as atividades de inibição da secreção gástrica e as estabilidades para decomposição espontânea em solução ácida, foram medidos e publicados por Terauchi e colaboradores. Estes dados são mostrados na Tabela 5.1. Para modelar a inibição da secreção gástrica *in vitro*, cujos dados foram omitidos da tabela, foram realizados modelos SAR e QSAR utilizando todos os métodos propostos nos capítulos anteriores. Para estes compostos também foram realizados cálculos de Orbitais de Ligação Naturais (NBO[1]) [142–144], na tentativa de obter descritores capazes de resultar modelos SAR e QSAR aceitáveis. Nenhum modelo baseado nos descritores mecânico-quântico calculados tem capacidade razoável de previsão das atividades inibidoras publicadas em modelos SAR ou QSAR.

Considerando que a conversão dos compostos para a forma ativa deve ocorrer apenas na região dos canalículos que ligam as células excretoras de prótons ao lúmen gástrico, onde o pH$\approx$3, podemos supor que compostos capazes de suportar condições de pH intermediárias podem ser mais potentes *in vivo*, que aqueles que sáo ótimos ligantes *in vitro*, por que somente transforma-se na forma ativa nas imediações dos canais de transporte de prótons. Os percentuais de inibição de secreção gástrica são medidas *in vivo*, e por isso levam em conta todos os fatores que contribuem para a resposta biológica. Os dados do percentual de inibição mostrados aqui são a resultante entre estabilidade relativa em meio ácido e afinidade pela enzima da forma resultante da reação de ativação.

[1]A decomposição NBO representa os cálculos, realizados com qualquer base adequada para cálculos mecânico-quânticos, em termos de combinações de orbitais moleculares com referência direta às ligações químicas de 2 ou 3 centros, e dos orbitais antiligantes correspondentes. Este tipo de representação pode dar boa ancoragem para idéias sobre o mecanismo de reação, ou reatividade de compostos

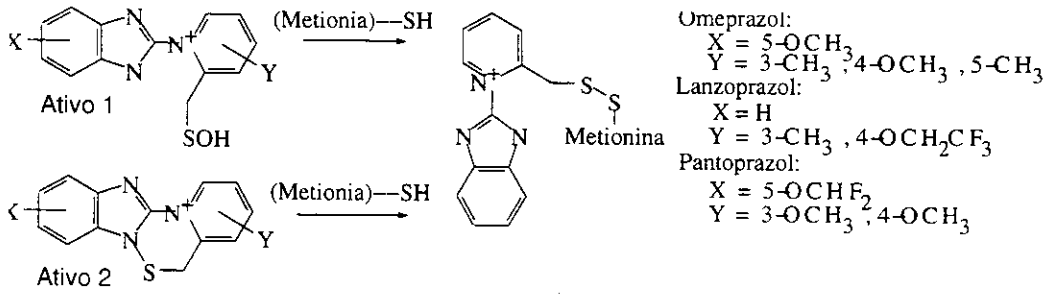



Figura 5.3: Mecanismo de geral de reação de ativação dos compostos semelhantes ao Omeprazol

Tabela 5.1: Nome atribuído aos compostos da série das Nicotinamidas; substituintes nas posições R1, R2, R3, R4 e R5 para o esqueleto da Figura 5.1; percentual de inibição da secreção gástrica em camundongos e tempos de meia vida ($t_{1/2}$) dos compostos em pH ácido.

| No. | R1 | R2 | R3 | R4 | R5 | % Inibição (dose mg.kg$^{-1}$) | $t_{1/2}$ em h | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | pH 3,0 | pH 5,0 |
| **nm-6b** | H | H | H | 2-NH-*n*-$C_3H_7$ | H | 12,9(10) | <0,02 | 0,10 |
| **nm-6c** | H | H | H | 2-NH-*iso*-$C_4H_9$ | H | 14,6(100) | <0,02 | 0,16 |
| **nm-7f** | H | H | H | 2-N($CH_3$)$C_2H_5$ | H | 51,2(10) | 0,38 | 0,62 |
| **nm-7g** | H | H | H | 2-N($CH_3$)-*n*-$C_3H_7$ | H | 22,8(10) | <0,02 | 0,29 |
| **nm-7h** | H | H | H | 2-N($CH_3$)-*iso*-$C_4H_9$ | H | 35,2(10) | 0,11 | 0,39 |
| **nm-9b** | H | H | H | 2-F,3-F,4-N($CH_3$)$_2$,5-F | H | 28,1(10) | 1,88 | >50 |
| **nm-9c** | H | H | H | 2-F,4-N($CH_3$)$_2$,5-F,6-F | H | 16,7(100) | <0,02 | 0,73 |
| **nm-22** | N($CH_3$)$_2$ | N($CH_3$)$_2$ | H | 2,4-$OCH_3$ | H | 57,5(10) | — | — |

A obtenção do percentual de inibição foi realizada por Terauchi e colaboradores segundo o procedimento (simplificado) que segue. A medição é realizado com grupos entre quatro até sete animais com o piloro (músculo que controla a passagem do estômago para o intestino) cirurgicamente obstruído. A droga ou o veículo de controle é administrada imediatamente após a obstrução cirúrgica e o abdômen do animal é suturado. Trinta minutos após a sutura, a secreção ácida é estimulada com administração de histamina (30 mg.kg$^{-1}$). Quatro horas depois o animal é sacrificado. O estômago é retirado e o volume de suco gástrico e sua acidez são medidos. O valor de produto do volume do suco gástrico pela sua acidez é utilizado para comparação com o valor obtido da mesma maneira para os animais do grupo de controle. O percentual de inibição é assim estabelecido como a relação entre as média dos produtos acidez-volume dos animais do grupo drogado e do grupo de controle. Valores da Dose Efetiva para 50% de Inibição ($ED_{50}$) são obtidos pela regressão linear de uma série de medidas do percentual de inibição.

## 5.2 Objetivos

- Modelar o mecanismo da reação de eliminação que converte estes compostos, que são pró-drogas, para a forma bioativa.

- Comparar as propostas mecanísticas encontradas nos modelos mecânico-quânticos

com as informações publicadas.

- Relacionar as energias calculadas para as estruturas de transição às taxas da decomposição destes compostos.

## 5.3 Procedimentos para busca do estado de transição e discussão

O fato dos compostos da série serem pró-drogas é considerado o principal fator de dificuldade para obtenção de modelos SAR/QSAR. Outro fator de dificuldade é a grande liberdade rotacional para estes compostos, que torna praticamente impossível isolar grupos de estruturas com maior estabilidade rotacional, em face do elevado número de conformações encontradas no intervalo de $3,0\,\mathrm{kcal.mol^{-1}}$ durante a análise conformacional.

Inicialmente foi realizado um estudo conformacional com o composto **nm-22** utilizando o método da matriz de distâncias métricas para gerar as conformações partindo de uma semente. As 2000 conformações geradas foram otimizadas com método semi-empírico AM1, e foram selecionadas somente as conformações com diferença maior que 0,4Å na soma dos valores RMS das distâncias nucleares, segundo a matriz de conectividade. Dentro deste conjunto de conformações foram selecionadas as trezentas e sessenta e três conformações dentro de um intervalo de $3,0\,\mathrm{kcal.mol^{-1}}$. Entre estas foi selecionada a conformação de mínimo local que aproximava os átomos das regiões citadas no mecanismo da Figura 5.2 como participantes de reação de eliminação. O composto **nm-22** é mostrado nesta conformação, denominada **nm-22**$\alpha$, na Figura 5.4. A conformação obtida pela inversão da configuração do átomo de enxofre, denominada **nm-22**$\beta$ é mostrada na Figura 5.5.

Modelar diretamente a reatividade de um composto requer assumir condições, já que a simulação completa do meio reacional é um problema computacional e teórico fora das possibilidades atuais. Assume-se que o composto não está protonado inicialmente, e que a protonação ocorre durante a mudança de conformação do átomo de enxofre, no átomo de oxigênio ligado a ele. O composto **nm-22** desta série foi escolhido porquê é o mais ativo nos testes *in vitro* de afinidade com a $H^+$,$K^+$-ATPase e faz parte do grupo dos mais ativos nas atividades *in vivo* mostradas na Tabela 5.1. O valor da sua meia vida não foi determinado pelos autores, entretanto.

Utilizando-se o confôrmero **nm-22**$\alpha$ como geometria inicial, manualmente fez-se a conformação do átomo de enxofre planar, alterando-se os valores de ângulos de ligação no arquivo de coordenadas atômicas da molécula. Esta geometria assim preparada foi submetida à busca de coordenado intrínseca de reação. O nível de cálculo utilizado foi Hartree-Fock com pseudo-potencial no cerne do tipo SBKJC e base 3-21G, com o programa GAMESS utilizando as seguintes diretivas no arquivo de entrada:

```
$contrl scftyp=rhf runtyp=sadpoint ecp=sbk coord=cart $end
$statpt  nstep=1000 $end
$basis gbasis=sbk  $end
$statpt hess=calc hssend=.true. $end
```

```
$scf dirscf=.true. diis=.true. soscf=.false. $end
```

A partir do ponto de sela pode-se calcular a coordenada de reação para cada um dos mínimos locais que ocorrem nas suas imediações. Os mínimos de menor energia e de segunda menor energia das configurações do átomo de enxofre foram denominados **nm-22**$\alpha$ e **mn-22**$\beta$. Utilizando-se as seguintes diretivas no *input* do GAMESS caminhamos do estado de transição para a conformação $\alpha$.

```
$contrl scftyp=rhf runtyp=irc ecp=sbk coord=cart $end
$basis gbasis=sbk  $end
$irc saddle=.true. pace=rk4 stride=0.1 npoint=1500 $end
$statpt hess=read hssend=.true. $end
$scf dirscf=.true. diis=.f. soscf=.t. maxvt=100 $end
```

Para selecionar a direção da conformação $\beta$ devemos incluir mais uma diretiva de comando no arquivo de entrada, no cartão `$irc forwrd=.t. $end`.

Partindo do ponto de sela foram obtidos os valores da barreira de inversão no sentido da conformação $\alpha$, mostrada na Figura 5.4, de mínimo local, e no sentido da conformação $\beta$ de segundo mínimo de energia, mostrada na Figura 5.5. A altura da barreira de inversão calculada é de $0,03854$ a.u. $(24,18\,\mathrm{kcal.mol^{-1}})$ pela esquerda, no sentido $\alpha \rightarrow \beta$. Pela direita, o sentido $\beta \rightarrow \alpha$, o valor da barreira é de $0,03522$ a.u. $(22,10\,\mathrm{kcal.mol^{-1}})$.

Valores da barreira energética obtidos com utilização de pseudo-potencial no nível Hartree-Fock com base n-31G provavelmente não reproduzem os valores experimentais da barreira de energia. O principal resultado obtido é uma geometria do estado de transição que deve ser utilizada como ponto inicial para cálculos com bases mais adequadas para caracterização de estados excitados. Resultados iniciais com bases pequenas podem diminuir sensivelmente o tempo total de cálculo necessário na busca pela geometria estado de transição e do perfil energético da relaxação do estado de transição.

O algorítimo utilizado [145] para seguir a coordenada de reação foi Runge-Kutta de 4ªordem em substituição ao algorítimo de González-Schlegel de 2ªordem. A substituição não é vantajosa do ponto de vista computacional, já que o algorítimo de González-Schlegel é mais eficiente, sendo necessária apenas nas imediações do ponto de sela, uma região em que o método de 2ª ordem não foi efetivo.

Para modelar o mecanismo da eliminação que converte estes compostos para a forma biologicamente ativa, foi usado o seguinte procedimento: A partir de um ponto de sela já calculado, adicionou-se um íon $H^+$ na molécula, próximo ao oxigênio ligado ao enxofre, (a posição indicada pelo mecanismo publicado, mostrado na Figura 5.2) e procedeu-se uma nova busca pelo ponto de sela, o cálculo da hessiana no ponto de sela, e a busca pelo ponto de sela final. A geometria deste ponto foi, mostrada na Figura 5.6, foi utilizada como ponto inicial para a simulação da reação de eliminação. A diferença de energia entre os dois estados de transição, protonado e não protonado, corresponde a $-0,0950$ a.u $(-59,64\,\mathrm{kcal.mol^{-1}})$.

A eliminação do grupo 2,4-dimetoxibenzila a partir do ponto de sela da Figura 5.6, com o grupo sulifinil protonado, é o caminho natural para os produtos da reação. Esta eliminação não concorda com o mecanismo publicado para esta reação catalisada por

ácido mostrado na Figura 5.2. O mecanismo aponta a ciclização do nitrogênio do grupo amida com o átomo de enxofre como a primeira etapa da reação de eliminação/ciclização.

Na Figura 5.7 é mostrada a geometria obtida seguindo a coordenada de reação a partir do ponto de sela, protonado na região do oxigênio, no sentido dos produtos. Eliminando os átomos correpondentes ao fragmento 2,4-dimetoxibenzila resta o frgmento que deve ciclizar-se para dar origem ao produto final da reação, denominado **22-f1**. A cisão da ligação carbono-enxofre é heterolítica, ficando o fragmento **22-f1** com carga neutra em função do próton adicionado anteriormente ao átomo de oxigênio durante a busca do estado de transição.

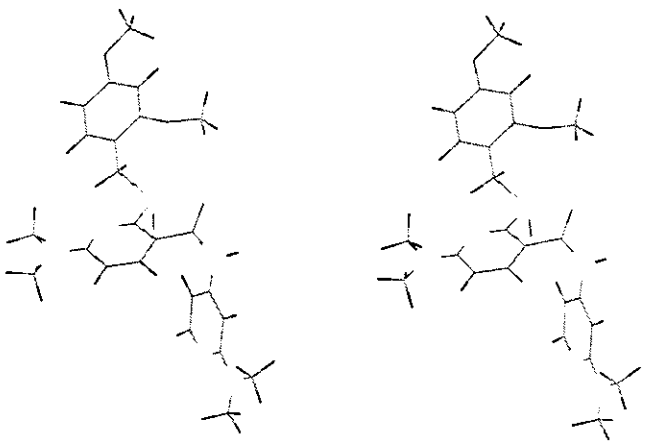


Figura 5.4: Estereograma da conformação mais estável ($\alpha$) do composto **nm-22** otimizada com método *ab initio* utilizando pseudo potencial SBKJC no cerne e base 31G na camada de valência. Um estereograma consiste no conjunto de duas figuras rotacionadas em cerca de 5°, que formam uma imagem tridimensional quando observadas adequadamente.

O fragmento **22-f1** obtido da eliminação, mostrado na Figura 5.7 foi utilizado para tentar simular o fechamento do anel. Este é o passo final para se obter os produtos mostrados na Figura 5.2. Partindo do estado de transição da Figura 5.7 foram calculados quatrocentos e quarenta pontos seguindo os modos de vibração com raiz imaginária. O valor de máximo na matriz de forças foram selecionados, e são denominados pelo número do ponto. A altura da barreira de energia entre os pontos cinqüenta e noventa corresponde a $-0,007428$ u.a. ($-4,66$ kcal.mol$^{-1}$). Entre os pontos número duzentos e número quatrocentos é de $-0,006689$ u.a. ($-4,20$ kcal.mol$^{-1}$). Entre os pontos número quatrocentos e número quatrocentos e quarenta é de $0,005687$ u.a. ($3,57$ kcal.mol$^{-1}$). Em cada um dos pontos foi recalculada a matriz hessiana e iniciada nova busca pelo estado de transição. Não foi encontrado nenhum intermediário que pudesse seguir um caminho de reação no sentido da ciclização.

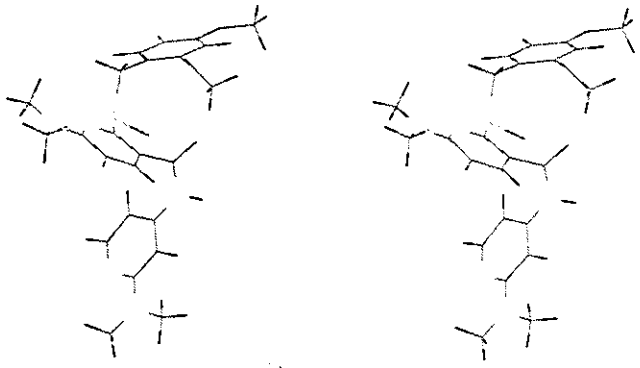


Figura 5.5: Estereograma da conformação $\beta$ do composto **nm-22** otimizada com método *ab initio* utilizando pseudo potencial SBKJC no cerne e base n-31G na camada de valência.

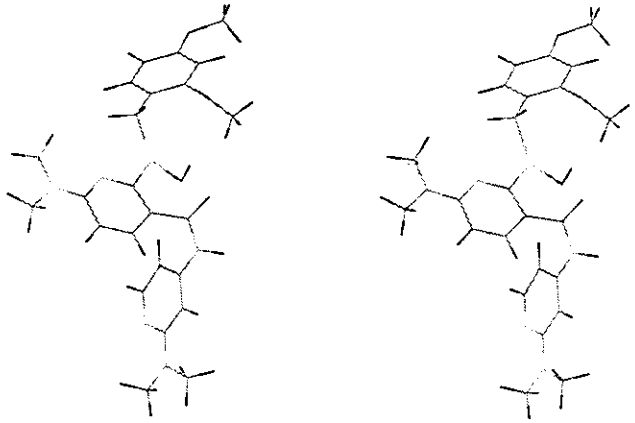


Figura 5.6: Estereograma do ponto de sela do composto $\mathbf{H^+}$**-nm-22** otimizada a partir do ponto de sela da espécie não protonada.

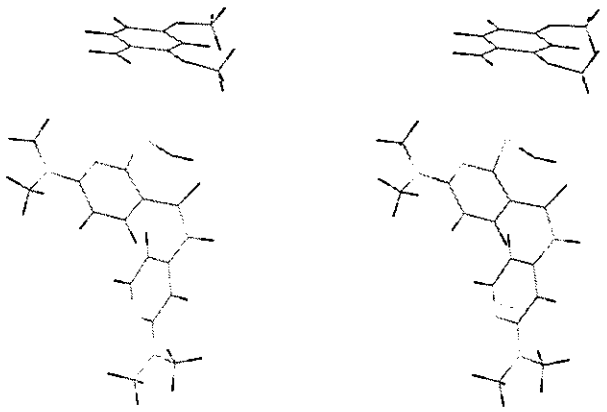


Figura 5.7: Estereograma do composto **nm-22** após a primeira etapa do mecanismo, a eliminação do grupo dimetóxi-benzil, otimizada com método *ab initio* e pseudo potencial SBKJC.

## 5.4 Conclusões

Os mecanismos mostrados nas Figuras 5.2 e 5.3 propõem que a facilidade de conversão para a forma ativa é preponderante sobre a afinidade droga–sítio na atividade das drogas pertencentes a estas classes. Na Tabela 5.1 vemos que os compostos com maior atividade são aqueles com um tempo de meia vida intermediário, capazes de permenecerem inativos até estarem nas condições encontradas nos canalículos da parede estomacal.

O modelo desenvolvido prossegue até o um intermediário semelhante àquele denominado 'Ativo 1' na Figura 5.3, não havendo caminho de reação que produza um intermediário semelhante ao denominado 'Ativo 2' na figura, que seria o caminho mostrado na Figura 5.2, e considerado o mais provável para esta série de compostos. A possibilidade de estabilizar a forma intermediária 'Ativo 1' é um fator que deve ser considerado na criação de modelos consistentes do mecanismo da atividade destes compostos na modulação da secreção gástrica.

# Capítulo 6

# Estudo da enzima $H^+$,$K^+$-ATPase

## 6.1 Introdução à modelagem de proteínas

A bomba de prótons $H^+$,$K^+$-ATPase consiste em sub-unidades $\alpha$ e $\beta$. A sub-unidade $\alpha$ é responsável pela catálise com dez domínios transmembrana. A sub-unidade $\beta$ tem apenas um domínio transmembrana, é glicoproteica e é responsável pela estabilização do complexo $\alpha/\beta$ como uma enzima funcional e pelo transporte intracelular da enzima. Há sete sítios putativos de N-glicosilação na sub-unidade $\beta$, e a remoção seqüencial desses sítios causa a perda também seqüencial da atividade do complexo enzimático $H^+$,$K^+$-ATPase até a perda completa [146].

A sub-unidade catalítica $\alpha$ tem massa molecular de 114 kDa, e contém um sítio de ligação de ATP, um sítio de acil-fosforilação e diversos sítios de ligação para inibidores da troca de íons. A $H^+$,$K^+$-ATPase é covalentemente inibida por piridil-metilsulfinil-benzimidazóis (como o Omeprazol), que se convertem em compostos tiofílicos no meio fortemente ácido da face luminal da parede estomacal e se ligam a cisteínas e metioninas

O inibidor SCH28080 também pode ligar-se à enzima. Sabe-se que a ligação de SCH28080 evita a inibição irreversível da enzima por Omeprazol, porém é incerto se isso acontece ligando-se ao mesmo sítio ou por mediação alostérica em outro sítio. Em mutações com expressão estável de células HEK293, as cisteínas que podem reagir com benzimidazóis são as seguintes (na enzima de coelhos): As Cisteínas situadas nas hélices transmembrana C321 (TM3), C813 (TM5), C822 (TM6), e C892 (TM8)[1]. Estes resíduos foram substituídas pelos amino-ácidos encontradas na enzima $Na^+$,$K^+$-ATPase resistente à SCH28080 por Sachs e colaboradores (entre outros grupos), e os parâmetros cinéticos da atividade da $H^+$,$K^+$-ATPase foram analisados. Mutações na C822 e C892 têm efeito insignificante nas constantes de equilíbrio $K_i$(app), $K_m$(app) ou $V_{max}$. As mutações da cisteína 813 por treonina (C813T) e da cisteína 321 por alanina (C321A) resultam variação na cinética de inibição, ainda com a alta afinidade pela forma livre de cátions,

[1] O código de uma letra para a nomenclatura de $\alpha$-amino-ácidos é A: Alanina; R: Arginina; D: Ácido aspártico; N: Asparagina; C: Cisteína; E: Ácido glutâmico; Q: Glutamina; G: Glicina; H: Histidina; I: Isoleucina; L: Leucina; K: Lisina; M: Metionina; F: Fenilalanina; P: Prolina; S: Serian; T: Treonina; W: Triptofano; Y: Tirosina; V: Valina. Freqüentemente é utilizado seguido do número do resíduo na cadeia — cisteína 321: C321

característica da fosfoenzima. O anel fenilmetoxi dos inibidores imidazo-piridínicos liga-se à TM1/2. As mutações em TM-6, substituindo cisteína por treonina (C813T), assim como no final de TM-3, C321A, diminuem a afinidade da enzima por SCH28080. As hélices TM1/2, TM3 e TM6 agrupam-se em distâncias ao redor de 16Å. Esta distância entre elas coincide com o comprimento da conformação ativa, estendida de SCH28080. Os resultados mostram que ao menos as cisteínas C321, C813, C822, C892 ligam-se aos benzimidazóis, e que a ligação em C813 é, provavelmente, a responsável pela inibição da enzima pelos benzimidazóis.

Um esquema da ligação dos benzimidazóis é mostrado na Figura 6.1. Tanto os inibidores do tipo reversível (SCH28080) quanto os do tipo irreversível (Omeprazol) tem sua forma ativa como um cátion permanente, que pode ser uma sulfenamida (nos benzimidazóis) ou a forma protonada ou quaternária do nitrogênio (imidazopiridinas). Ambas as classes ligam-se à enzima na face luminal, e têm comprimento entre 12Å e 16Å [119,120].

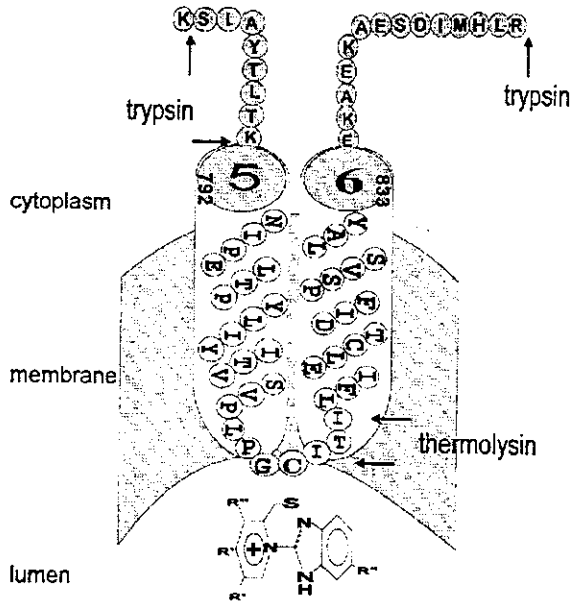



Figura 6.1: Esquema ilustrativo das hélices transmembrana TM5/6 sub-unidade $\alpha$ da enzima $H^+$,$K^+$-ATPase. O sítio comum de ação para os compostos tipo benzimidazolínicos é a Cys-813. O esquema mostra também como a Cys-822 fica mais protegida dentro da membrana plasmática.

A enzima $H^+$,$K^+$-ATPase completa na sua forma nativa em coelhos tem 1035 resíduos, distribuídos em dez sub-unidades transmembrana (TM), cinco laços intracelulares e cinco laços extracelulares. O maior fragmento é o laço intracelular entre os segmentos TM6/7. Acredita-se que o sítio de ação da droga SCH28080 seja localizado nos segmentos TM1/2 e TM5/6 [119,120,141,147–150].

O segmento extracelular entre os segmentos TM7/8 [151] também é considerado participante do mecanismo de atividade da droga SCH28080, porém não há informação sobre análise de mutantes, para determinar quais resíduos são participantes putativos do sítio de ligação. O diagrama da estrutura da proteína, com os pontos de inserção na mem-

brana e os laços dentro e fora do citoplasma é mostrado na Figura 6.2. O sequenciamento completo da proteína é mostrado na Figura 6.3.

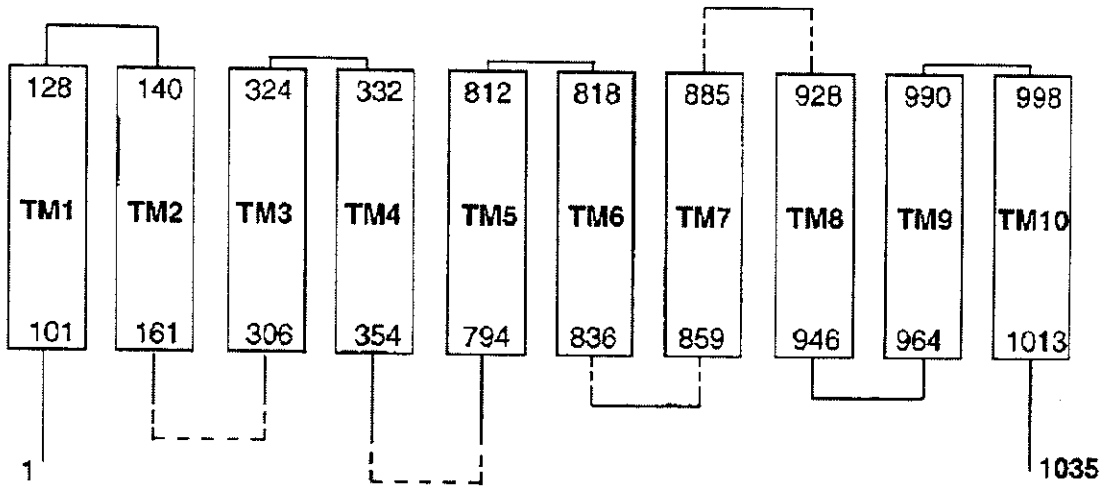



Figura 6.2: Diagrama da estrutura completa da enzima $H^+$,$K^+$-ATPase gástrica de coelhos, mostrando as posições de inserção de cada segmento na membrana e o comprimento dos laçoes na face luminal da célula epitelial gástrica (acima na figura) e no citoplasma (abaixo na figura).

## 6.2 Objetivos

- Relacionar regiões da proteína onde haja possíveis sítios receptores para compostos entre as classes de drogas já estudadas.
- Obter simulações da interação entre a macromolécula e os compostos mencionados.
- Determinar fatores potencialmente importantes na afinidade droga-receptor.

## 6.3 Modelagem estrutural dos segmentos transmembrana da $H^+$,$K^+$-ATPase

Inicialmente foram estudados os segmentos TM1/2 e TM5/6, para os quais são disponíveis dados de afinidade da droga SCH28080 pela enzima, na forma nativa e com várias mutações induzidas na região de ligação proposta.

Os segmentos TM1/2 e TM5/6 foram estudados por Asano e colaboradores por análise de mutações induzidas na enzima de coelhos e cães. Os autores obtiveram mutações em regiões determinadas, e mediram as variações da afinidade dos mutantes em relação à forma nativa, para cada mutação. Estes resultados permitiram selecionar regiões que de

alguma forma relacionam-se ao mecanismo de interação, mesmo sem conhecer a estrutura terciária da proteína. Estas regiões são chamadas de sítios putativos de ação da droga.

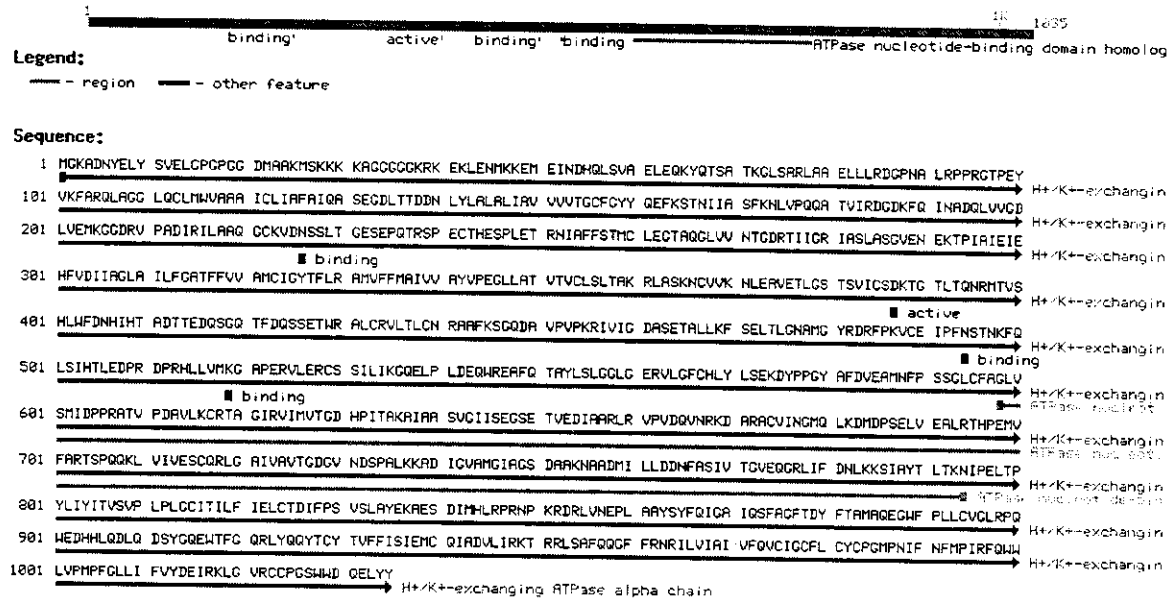



Figura 6.3: Sequenciamento completo dos resíduos de amino-ácidos de enzima $H^+$,$K^+$-ATPase gástrica de coelhos. Esta seqüência de resíduos foi a base para a construção dos segmentos transmembrana utilizados nos cálculos. A ilustração da Figura 6.1 mostra um detalhamento da região dos segmentos TM-5/6.

### 6.3.1 Métodos e procedimentos para simulação de fragmentos de proteína com dinâmica molecular

Os segmentos TM1/2 e TM5/6 da enzima foram construídos através do seqüenciamento da Figura 6.3 e do diagrama de inserção na membrana da Figura 6.2. Para os amino-ácidos inseridos na membrana a construção foi em $\alpha$-hélice. Os laços foram construídos de forma a reproduzir as distâncias aproximadas entre os pontos de inserção entre os segmentos trans-membrana anterior e posterior ao laço. Estes segmentos assim construídos tiveram a energia minimizada em campo de força MM2 usando efeito de solvatação em clorofórmio, durante cinco mil iterações. O campo de força MM2 utilizado no programa `MacroModel` [152] tem o número de parâmetros de alta, média ou baixa qualidade para este fragmento de proteína mostrados na Tabela 6.1.

A geometria resultante da minimização com MM2 foi utilizada para a organização da estrutura terciária do fragmento de acordo com dados estruturais com cálculos de Dinâmica Molecular (MD). Foram realizadas simulações em 20 ps com passo de 0,001 ps. O resumo dos resultados obtidos está listado na Tabela 6.2. A porção terminal das segmentos TM foi mantida a uma distância de $\approx 10$ Å com a aplicação de uma interação não ligante com constante de força de 10 kJ/mol.Å entre os átomos de oxigênio das ligações

Tabela 6.1: Qualidade dos parâmetros do campo de forças MM2, utilizado no `MacroModel` para os cálculos com o fragmento TM1/2 da $H^+$,$K^+$-ATPase de coelho construída.

| | Alta | Média | Baixa |
|---|---|---|---|
| *bend* | 509 | 230 | 0 |
| *stretch* | 891 | 468 | 3 |
| *torsion* | 868 | 956 | 24 |

Tabela 6.2: Resultaods da etapa de equilibração de geometria do fragmento TM1/2 da proteína $H^+$,$K^+$-ATPase. O tempo total de simulação é de 20 ps e os passos são de 1 fs.

| Tempo | $E^{total}$ | $T^{0,5ps}$ | $H^m$ | $T^m$ | $H^m_{300K}$ |
|---|---|---|---|---|---|
| ps | kJ/mol | K | kJ/mol | K | kJ/mol |
| 2,001 | 1829,4 | 296,5 | -128,1 | 297,6 | -112,7 |
| 4,001 | 1956,0 | 299,0 | -38,4 | 297,3 | -21,0 |
| 6,000 | 2096,0 | 300,8 | 45,1 | 297,5 | 61,2 |
| 8,001 | 2167,8 | 299,0 | 92,3 | 297,8 | 106,6 |
| 10,000 | 2165,1 | 300,5 | 128,8 | 298,0 | 141,5 |
| 12,000 | 2012,6 | 300,2 | 139,6 | 298,5 | 149,4 |
| 14,001 | 2110,2 | 299,7 | 148,6 | 298,6 | 157,7 |
| 16,000 | 2092,5 | 302,7 | 153,1 | 298,8 | 160,6 |
| 18,000 | 2094,0 | 301,7 | 155,5 | 299,0 | 162,0 |
| 20,001 | 2087,1 | 300,1 | 157,5 | 299,0 | 163,9 |

peptídicas do primeiro e do último amino ácido das cadeia. Os valores da temperatura medida no intervalo e 0,5 ps decrescem. A diminuição da variação da energia total escalada para 300 K ($E^m_{300K}$ na tabela) para valores em torno de 2 kJ/mol ($\approx$0,5 kcal.mol$^{-1}$) foi considerado o sinal de parada na minimização.

A outra etapa de simulação com MD foi realizada sem interações do tipo não ligantes entre as estremidades do segmento. A simulação foi por um período de 300 ps e passos de 1,5 fs. A geometria de menor energia durante todo o período de simulação em clorofórmio foi considerada adequada para esse fragmento. O resumo dos resultados está listado na Tabela 6.3. A diminuição da temperatura média é quase constante, prevendo um esfriamento do sistema. O valor da energia potencial ao fim do tempo de simulação escalado para 300 K ($H^m_{300K}$, na coluna da tabela), é de 115,7 kJ/mol. A maior varição nos primeiros passos da simulação corresponde à acomodação da estrutura na região da eliminação das interações não ligantes. Como não nos preocupa um mapeamento da energia conformacional e a estrutura terciária do fragmento deve ser mantida, a estabilização na queda do valor dessa energia para valores próximos de 2 kJ.mol$^{-1}$ é considerado um indicador de parada da minimização da energia.

Tabela 6.3: Resumo da simulação com Dinâmica Molecular com o fragmento TM1/2 de enzima $H^+$,$K^+$-ATPase de coelho. O tempo total de simulação é de 300 ps e os passos são de 1,5 fs.

| Tempo ps | $E^{total}$ kJ/mol | $T^{0,5ps}$ K | $H^m$ kJ/mol | $T^m$ K | $H^m_{300K}$ kJ/mol |
|---|---|---|---|---|---|
| 20,001 | 2213,3 | 301,2 | 264,4 | 301,2 | 256,7 |
| 40,000 | 2010,7 | 296,8 | 230,3 | 300,4 | 227,6 |
| 60,000 | 2009,0 | 298,6 | 196,6 | 300,2 | 195,3 |
| 80,001 | 1978,0 | 298,1 | 180,3 | 300,1 | 179,8 |
| 100,001 | 2102,7 | 296,4 | 164,4 | 300,0 | 164,5 |
| 120,002 | 2008,5 | 298,4 | 154,2 | 300,0 | 154,5 |
| 140,001 | 2022,5 | 301,2 | 147,1 | 299,9 | 147,5 |
| 160,001 | 2000,5 | 297,5 | 142,4 | 299,9 | 143,0 |
| 180,001 | 1924,6 | 301,4 | 135,5 | 299,9 | 136,2 |
| 200,001 | 1995,2 | 298,9 | 129,7 | 299,9 | 130,5 |
| 220,001 | 1819,4 | 299,3 | 125,9 | 299,9 | 126,7 |
| 240,001 | 1950,8 | 298,7 | 122,2 | 299,8 | 123,2 |
| 260,001 | 1991,7 | 301,4 | 119,4 | 299,8 | 120,4 |
| 280,000 | 2002,4 | 298,8 | 117,0 | 299,8 | 118,1 |
| 300,001 | 2057,7 | 300,7 | 114,6 | 299,8 | 115,7 |

### 6.3.2 Resultados obtidos na simulação de fragmentos da enzima

O gráfico de Ramachandran mostrado na Figura 6.4 mostra que não há problemas de distorção das ligações peptídicas na estrutura obtida, usando o procedimento de cálculo descrito. Os resíduos de $\alpha$-hélice aparecem todos no 1º ou no 4º quadrante, dentro da região permitida delimitada pelas linhas mais claras no gráfico. Esse resultado está em conformidade com uma estrutura sem tensão torsional nas ligações peptídicas $\Phi$ e $\Psi$. Os resíduos que aparecem na região permitida menos favorável, dentro do espaço das linhas mais escuras do gráfico, são os resíduos de aspartato A27, leucina L28, treonina T29 e treonina T30 localizados na região do laço extracelular. Estes resíduos não estão formando $\alpha$-hélice, e portanto podem estar localizados nesta região sem que isto signifique grande tensão torsional das ligações peptídicas. As Glicinas são mostradas como quadrados no gráfico mostrado, no 1ºquadrante ou no 2ºquadrante.

Os valores de atividade para a forma nativa da enzima e para os mutantes preparados por substituições de resíduos na região TM1/2 são listados na Tabela 6.4. Na Figura 6.3 da página 132 é mostrada toda a seqüência de amino ácidos da proteína. O fragmento construído, que contém os segmentos TM1/2 e o laço extracelular ligado a eles, tem sessenta amino-ácidos no total. Inicia-se no resíduo de valina V101 e termina no resíduo de glutamina Q161.

As modificações feitas no resíduo F126 do segmento TM2 diminuem a atividade quando o resíduo de Fenilalanina (F–cadeia lateral aromática), é trocado por Leucina (L–cadeia lateral alifática) ou por Alanina (A–cadeia lateral alifática). A cadeia lateral

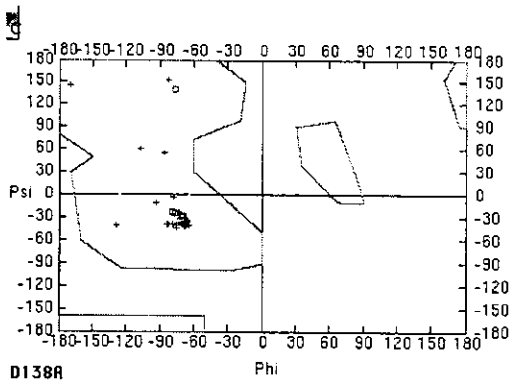



Figura 6.4: Gráfico de Ramachandran para o segmento TM1/2 da proteína, obtido da estrutura minimizada com MM2 seguida de simulação de DM pelo procedimento descrito. As regiões de menor energia no campo estérico são mostradas delimitadas em linhas claras. As regiões de energia intermediária delimitadas pelas linhas escuras, e as regiões pouco prováveis situam-se ao redor. Os resíduos de Glicina e de Prolina aparecm marcados por quadrados. Estes resíduos têm regiões favoráveis bastante diferentes dos demais, podendo ocorrer fora das regiões determinadas na figura por causa de sua estrutura peculiar.

Tabela 6.4: Valores da atividade medidos para a enzima $H^+$,$K^+$-ATPase de coelho, em sua forma nativa e nos mutantes preparados para avaliar os possíveis sítios de ligação da droga SCH28080 no segmento TM1/2.

| Tipo | Atividade | Porcentagem de Inibição | |
|---|---|---|---|
| | $\mu$mol/mg/h | 10 $\mu$molL$^{-1}$ | 50 $\mu$molL$^{-1}$ |
| Nativa | 1,19±0,02 | 80 | 100 |
| F126A | 0,72±0,04 | 98 | 100 |
| F126L | 0,48±0,05 | 80 | 84 |
| F126Y | 0,91±0,07 | 75 | 77 |
| D138A | 0,79±0,04 | 75 | 94 |
| D138E | 0,82±0,05 | 82 | 83 |
| D138N | 0,94±0,07 | 78 | 79 |
| D138V | 0,89±0,05 | 85 | 90 |
| F126A/D138A | 0,76±0,07 | 76 | 85 |
| F126Y/D138N | 0,83±0,05 | 81 | 95 |
| E132A | 0,96±0,08 | 70 | 93 |
| G133A | 0,80±0,03 | 62 | 100 |
| G133E | 0,81±0,08 | 77 | 88 |
| D134A | 0,61±0,06 | 56 | 83 |
| L135A | 0,65±0,06 | 85 | 95 |
| T136A | 1,27±0,03 | 74 | 94 |
| T137A | 1,18±0,06 | 78 | 97 |

da Leucina é mais volumosa que a da Alanina, e ambas são menos volumosas que a Fenilalanina. A substituição da Fenilalanina por Tirosina (Y–cadeia lateral aromática) afeta menos a atividade, ainda com pequena diminuição. A sensibilidade das formas mutantes à droga SCH28080 é afetada. O mutante F126A é mais sensível à droga em pequenas doses que a forma nativa, e os mutantes F126L e F126Y são menos sensíveis tanto em alta como em baixa dose de SCH28080.

As substituições no resíduo D138 afetam a atividade dentro do limite de confiança das medidas, e a única conclusão é que diminuem ligeiramente a atividade da enzima, e a sensibilidade à droga SCH28080. As substituições conjuntas F126A/D138 e F126Y/D138N também diminuem a atividade da enzima, com valores indistinguíveis dentro do limite de confiança. A sensibilidade à SCH28080 é mais reduzida no mutante F126Y/D138N. A substituição E132A diminui ligeiramente a atividade e a sensibilidade da enzima. As mutações G133A e G133E também diminuem a atividade, porém só a mutação G133E diminui significativamente a sensibilidade por SCH28080. A mutação D134A diminui sensivelmente a atividade e a sensibilidade. A mutação L135A diminui significativamente a atividade, sem alterar significativamente a sensibilidade. A mutação T136A aumenta a atividade em relação à forma nativa, com uma pequena diminuição da sensibilidade. A mutação T137A não altera a atividade, nem a sensibilidade.

### 6.3.3 Conclusões sobre a análise dos fragmentos da proteína

Para o segmento na forma nativa e para alguns mutantes são mostradas as figuras com a superfície molecular e o potencial eletrostático, calculado com o programa SPDBV [153] utilizando a parametrização do GROMOS96. Analisando os gráficos mostrados na Figura 6.5 podemos chegar a algumas conclusões: Os mutantes T136A e T137A têm a superfície de potencial eletrostático negativo estendendo-se ao longo de todo o segmento TM2. Isto os difere dos demais, e os torna mais semelhante à enzima nativa. Este fator pode ser responsável por uma maior exposição deste segmento ao meio extracelular, porque a tendência é que regiões mais polares fiquem fora da membrana. Esta é uma suposição baseada nas informações vindas da modelagem de um fragmento isolado da enzima, e pode não corresponder aos fatos, se analisada no contexto de um modelo da enzima toda.

Os valores medidos para os mutantes na região TM5/6 são listados na Tabela 6.5. Não foram publicados dados do percentual de inibição da enzima para os mutantes E822L e E822Q. Pode-se observar que a substituição do resíduo de Glutamato (E – negativamente carregado) por Aspartato (D – negativamente carregado) diminui um pouco a atividade da enzima. A substituição por Alanina (A – alifático) também reduz a atividade. Se o resíduo é substituído por Leucina (L – alifático) ou por Glutamina (Q – polar com grupo amina) a atividade é praticamente inexistente. A conclusão é que este resíduo deve fazer parte do sítio de ação da enzima, e que deve ser um resíduo negativamente carregado preferencialmente. A análise dos mapas de potencial eletrostático da Figura 6.6 mostra exatamente a mesma conclusão. Os mutantes E822L e E822Q têm a superfície de potencial eletrostático restrita à região final do segmento TM6, enquanto a forma nativa e o mutante E822D têm a superfície estendendo-se por grande parte do segmento TM6. O mutante E822A têm uma superfície de potencial eletrostático que assemelha-se mais

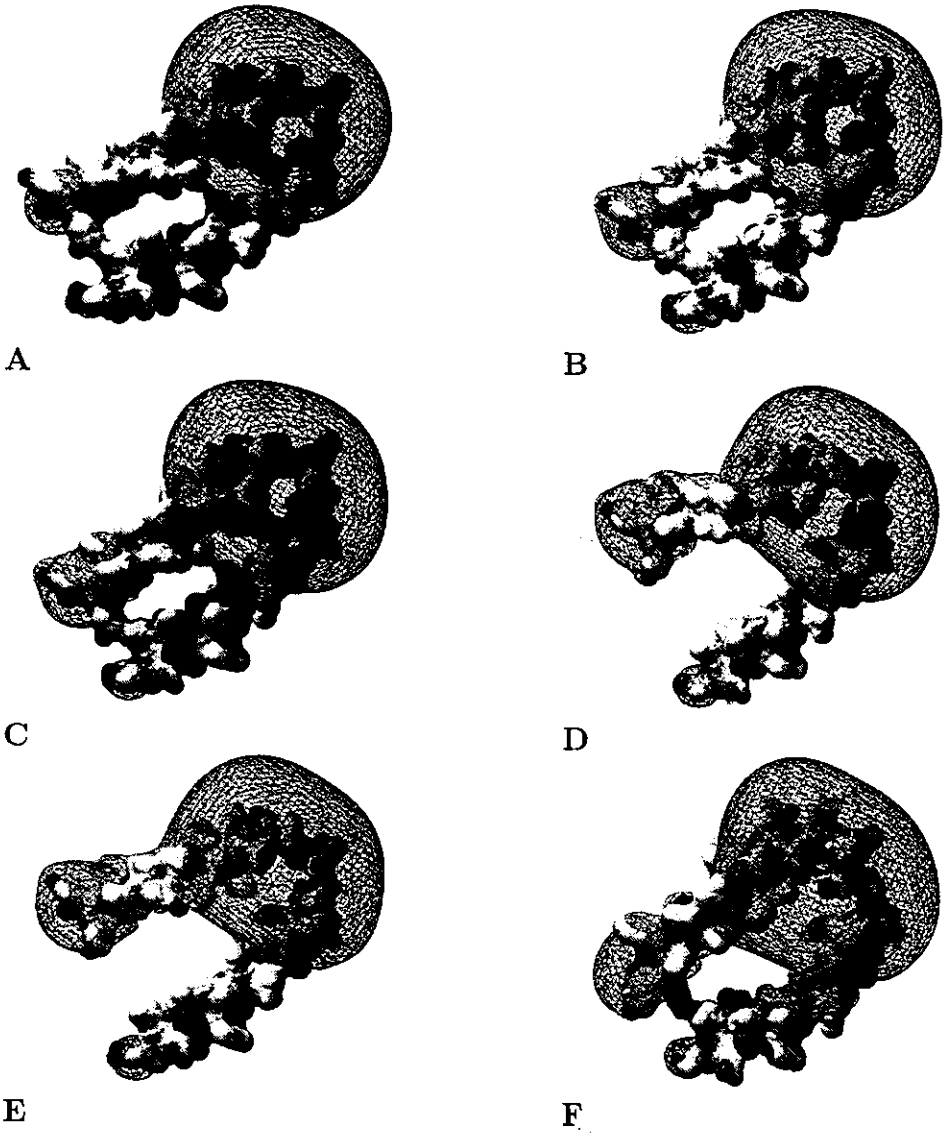



Figura 6.5: Superfície molecular de van der Waals e mapa do potencial eletrostático para os segmentos mutantes F126A (**A**), F126L (**B**), F126Y (**C**), T136A (**D**), T137A (**E**) e na forma nativa (de coelho) (**F**). O mapa mostra a superfície de potencial eletrostático mais volumoso, estendendo-se por todo o segmento TM2, nas mutações T136A, T137A e na forma nativa, que são as mais ativas, como pode ser observado nos dados mostrados na Tabela 6.4.

Tabela 6.5: Valores da atividade medidos para a enzima $H^+$,$K^+$-ATPase de coelho, em sua forma nativa e nos mutantes preparados para avaliar os possíveis sítios de ligação da droga SCH28080 no segmento TM5/6.

| Tipo | Atividade | Porcentagem de Inibição | |
|---|---|---|---|
| | $\mu$mol/mg/h | 50$\mu$M | 10$\mu$M |
| Nativa | 0,76±0,05 | 18 | 0 |
| E822A | 0,23±0,05 | 20 | 10 |
| E822D | 0,31±0,05 | 60 | 15 |
| E822L | -0,13±0,04 | — | — |
| E822Q | 0,02±0,01 | — | — |

aos mutantes menos ativos E822L e E822Q. Este modelo não é capaz de ilustrar este fato, nem de ajudar a compreendê-lo. Estes mapas obtidos ilustram, da mesma forma que para o segmento TM1/2, as conclusões obtidas com a análise do segmento isolado do resto da enzima. As mesmas restrições já apontadas são válidas também para as conclusões apontadas aqui.

## 6.4 Procedimentos de alinhamento ligante–sítio nos fragmentos de proteína

O segmento TM5/6 é considerado o principal sítio de ação da droga SCH28080 na enzima $H^+$,$K^+$-ATPase gástrica. Este segmento modelado conforme o procedimento apresentado na página 131 com a seqüencia de resíduos da proteína humana **P20648**, obtido pelo serviço **ExPASy** [154]. A partir da seqüência da amino-ácidos foi construída a estrutura terciária do segmento utilizando um procedimento análogo ao descrito para a construção do segmento TM1/2 da proteína de coelhos, descrito na Secção 6.3.1. Esta seqüência foi utilizada para alinhamento e mapeamento dos sítios de ligação com o composto mostrado na Figura 6.7, pertencente à classe dos indolil-guanidinotiazóis estudados na Secção 3.4.

Para analizar a interação ligante–macromolécula foi utilizado o programa `AutoDock` [42]. Este programa realiza o alinhamento entre droga e enzima baseado em interações eletrostáticas e repulsão entre campos de força representando superfícies moleculares (campo estérico). O programa tem rotinas de acesso a bancos de dados para classificar átomos, resíduos de amino-ácidos e resíduos de DNA e RNA, baseando-se no campo de força `AMBER95`.

Este programa cria um *grid* como o da Figura 6.8 e calcula valores de potencial eletrostático e energia potencial de van der Waals através das interações obtidas entre cada átomo de prova e cada átomo da macromolécula e do ligante. A formulação utilizada é desenvolvida levando em conta resultados empíricos de afinidade entre ligantes e macromoléculas, visando reproduzir os dados e tornar os modelos úteis em QSAR.

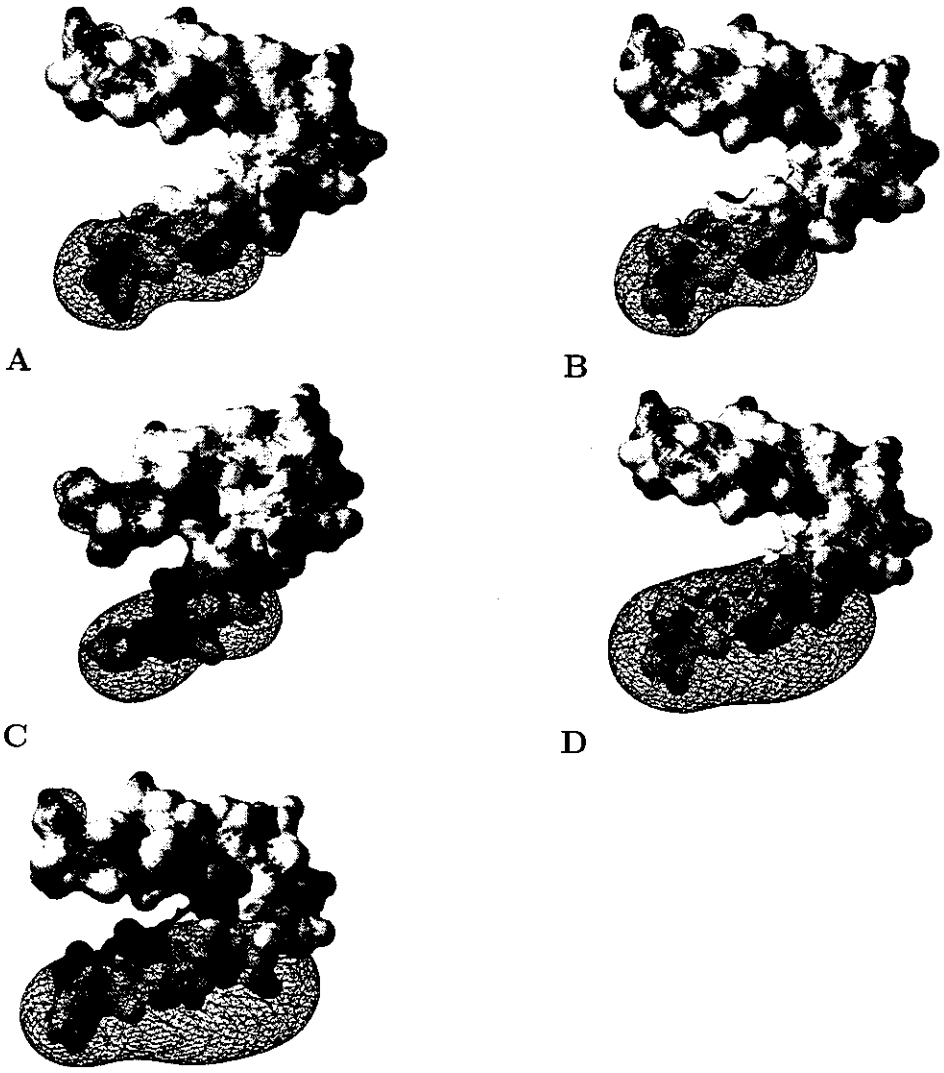



Figura 6.6: Superfície molecular de potencial eletrostático para os segmentos mutantes E822Q (**A**), E822L (**B**), E822A (**C**), E822D (**D**), e na forma nativa (de coelho) (**E**). O mapa mostra a superfície de potencial eletrostático mais volumoso, estendendo-se por todo o segmento TM6, na mutação E822D e na forma nativa, que são as mais ativas, como pode ser observado nos dados mostrados na Tabela 6.5.

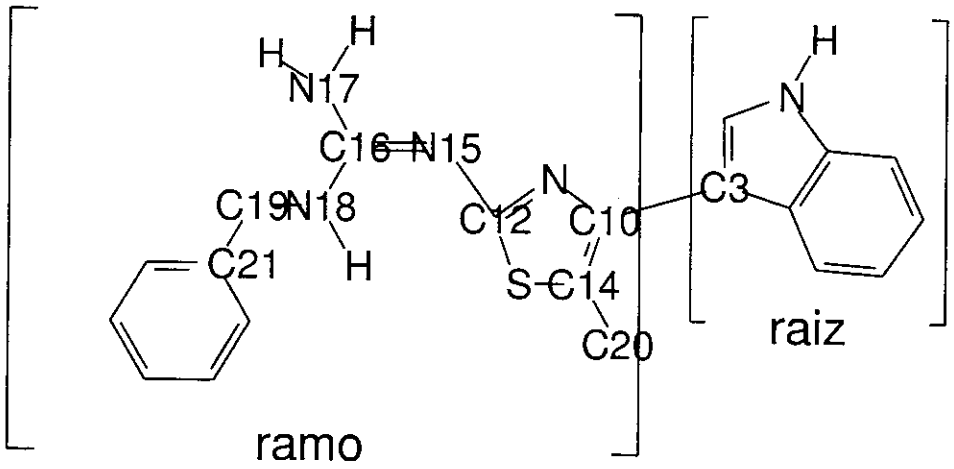



Figura 6.7: Esquema de seleção dos modos de liberdade rotacional da estrutura do composto **igt-45** da Tabela 3.7 da página 91, utilizado no alinhamento com fragmentos da enzima $H^+$ATPase. O ligantes são subdivididos em **raiz**, **ramo** e **galhos** para sorteio de conformações no programa `AutoDock`. Os **galhos** correspondem às ligações entre os átomos numeradso na molécula.

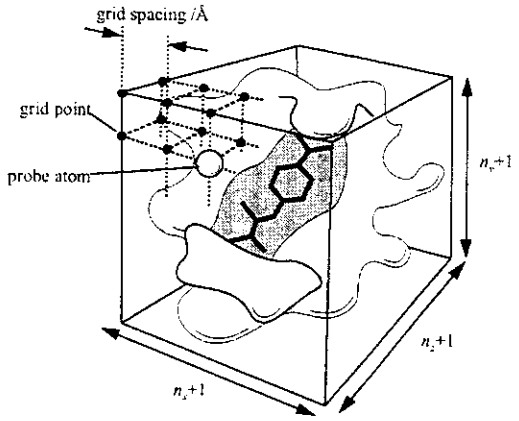



Figura 6.8: Gráfico mostrando um *grid* genérico utilizado pelo programa `AutoDock` para calcular potenciais eletrostáticos e energias potenciais de van der Waals para interação ligante–macromolécula.. O método determina diversos *grids* para os campos estéricos e eletrostáticos de cada tipo de átomo classificado e combina os resultados a funções empíricas para reprodução da energia livre de ligação de sistemas bem conhecidos.

### 6.4.1 Procedimentos utilizados para alinhamento sítio–ligante e resultados obtidos

Para o ligante mostrado na Figura 6.7 foi necessário definir uma topologia molecular. Os raios de van der Waals foram atribuídos de acordo com o tipo atômico, utilizando dados dos amino-ácidos como estimativa para os parâmetros. Para os átomos de carbono fora do sistema conjugado, foram utilizados os raios de carbonos alifáticos; para os carbonos em conjugação, os raios de carbonos aromáticos; para os nitrogênios das aminas, os raios de aminas de cadeias laterais; para os nitrogênios em conjugação, os raios dos nitrogênios das ligações peptídicas; para o enxofre, os raios da ligação do enxofre da Metionina (M). Para o cloro há parâmetros próprios na base de dados interna do programa. As cargas utilizadas para o ligante foram obtidas em cálculo *ab initio* HF(6-31*G) com método CHELPG.

Para a macromolécula foram utilizadas as cargas do campo de força AMBER. Um programa auxiliar denominado `PMOL2Q` [155] foi utilizado para adicionar os parâmetros de carga ao arquivo PDB em conformidade aos arquivos de entrada do programa `AutoDock`. É necessário definir quais hidrogênios são adicionados à estrutura da proteína obtida no banco de dados. Este passo implica determinar as condições de pH. Na região dos canalículos que ligam a superfície externa das células ao interior do estômago através da camada de muco que recobre o estômago, há um gradiente de pH entre meio muito ácido do interior do estômago ($\approx$ pH 2) e o meio tamponado da camada de muco que recobre a parede estomacal ($\approx$ pH 7). Considerando o pKa=4,5 do ácido aspártico e do ácido glutâmico, foram adicionados prótons às carboxilas.

Para realização dos cálculos de alinhamento, a geometria da macromolécula é congelada, e o ligante tem alguma liberdade rotacional, previamente estabelecida. A liberdade rotacional do ligante é atribuída relacionando-se dois fragmentos: **raiz** e **ramo**. O fragmento **raiz** não tem liberdade rotacional, e o **ramo** tem graus de liberdades, denominados **galhos**. Para o composto **igt-45** da Tabela 3.7 de página 91 a definição do separação entre ramo e raiz foi feita na ligação C3–C10, onde o grupo indolil constitui a **raiz**. O restante da molécula foi considerado o **ramo**. As ligações C12–N15, N15–N16, C16–N17, C16–N18, N18–C19, C19–C21 e C20–C13 no ramo foram definidas como **galhos**. A ligação ramo–raiz (C3–C10) também foi definida com liberdade rotacional O esquema é mostrado na Figura 6.7.

O *grid* utilizado tem $0,2\,\text{Å}$ de distância entre os nós. Para os *grids* de potencial eletrostático são colocados corpos de prova do tipo $H^+$. Os alinhamentos foram realizados usando *Monte Carlo Simulated Annealing* para sorteio de geometrias, que utiliza duas fases. O método de sorteio realiza translações grandes do ligante em fases denominadas **corridas**. Nas fases denominadas **ciclos** o programa realiza translações pequenas no ligante, rotações do ligante e rotações dos **galhos**. O número máximo de **corridas** é 128. O número máximo de **ciclos** é 9999. Os valores máximos de **corridas** e **ciclos** foram utilizados nas simulações. O limite de aceitação do tamanho das translações grandes e pequenas é definido pelo programa levando em conta o tamanho do *grid*.

Um alinhamento provável ocorre entre a droga e o segmento TM5 na região próxima aos resíduos de isoleucina I795, prolina P796, leucina L798 e treonina T799. Este ali-

nhamento é mostrado na Figura 6.9. A posição do grupo indolil da droga, próxima à hidroxila da treonina, permite a formação de ponte de hidrogênio nessa região. Os grupos guanidino e tiazol do ligante ficam próximos aos grupos isoleucina e prolina, sem que haja grande impedimento estérico. A energia total associada a este alinhamento é de $-4,32\,\mathrm{kcal.mol^{-1}}$. Não foram obtidos outros alinhamentos com energia livre mais negativa que $-4,00\,\mathrm{kcal.mol^{-1}}$ para este resíduo. Este valor corresponde a uma barreira de energia pouco provável de ser ultrapassada em temperatura de 300 K, se aplicado o procedimento para calcular a estabilidade relativa entre conformações mostrado na página 76.

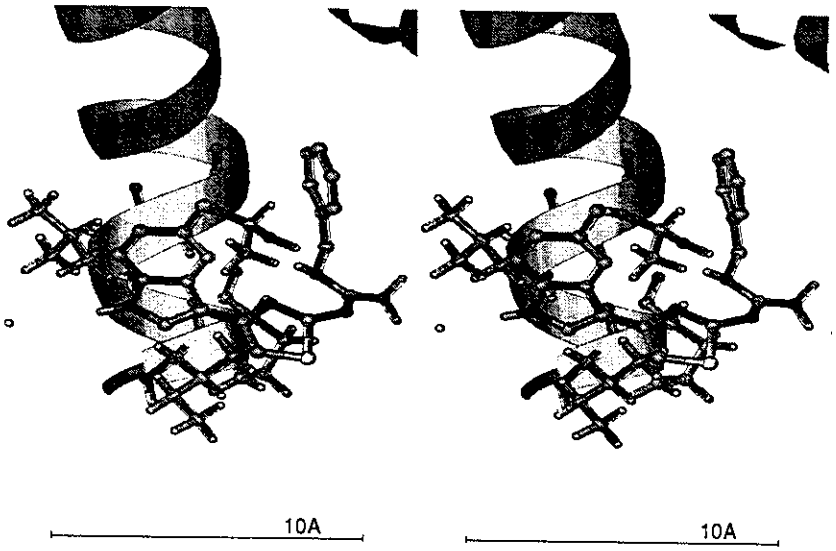



Figura 6.9: Estereograma mostrando um alinhamento droga-sitio na região do segmento TM5. Os resíduos I795, P796, L798 e T799 ficam próximos à droga, constuindo um possível sítio de ligação. O resíduo T799 tem o hidrogênio da hidroxila em posição favorável para realizar ponte de hidrogênio com o nitrogênio do grupo indolil do composto igt-45. Os demais resíduos apenas encaixam-se sem que haja grande impedimento estérico.

### 6.4.2 Conclusões sobre o estudo estrutural dos fragmentos da $H^+$,$K^+$-ATPase

Os resultados obtidos indicam que o mecanismo de atividade e das enzima $H^+$,$K^+$-ATPase está ligado ao modo de inserção da enzima na membrana das células apicais da parede estomacal. A inserção da enzima deve ser controlada pela hidrofilicidade dos resíduos em cada segmento, e pela interação entre segmentos próximos, na estrutura terciária da enzima. Pequenas variações na seqüencia de resíduos causam mudanças estruturais nos modelos. Estas mudanças também devem ter papel fundamental para a atividade e para a afinidade da enzima com as drogas, efeito que não pode ser mensurado em análises de fragmentos. Considerando apenas o fagmento TM5/6 temos um possível sítio de ligação na região dos resíduos isoleucina I795, prolina P796, leucina L798 e treonina T799, baseado principalmente em ponte de hidrogênio.

## 6.5 A estrutura completa da proteína $H^+$-ATPase

A elucidação da estrutura da enzima $H^+$-ATPase do fungo *Neurospora crassa* por Kühlbrandt e colaboradores [156] através de modelagem molecula, baseando-se na estrutura da $Ca^{2+}$-ATPase de coelhos obtida a uma resolução de $2,6\,Å$ por Toyoshima e colaboradores [157], torna possível o trabalho com fragmentos maiores da proteína, porquê permite alinhar corretamente os segmentos transmembrana em uma estrutura terciária mais próxima àquela encontrada no meio biológico. O seu desenho é mostrado na Figura 6.10. O seqüenciamento completo das duas proteínas, mostrando os resíduos homologos e aqueles substituídos por resíduos semelhantes, pode ser visto na Figura 6.11. A estrutura terciária depositada no banco de dados do *Protein data Bank* tem o código de acesso 1MHS. A parte relativa á secção mostrada no plano **C** de Figura 6.10 foi utilizada na construção de um fragmento da porção transmembrana da proteína.

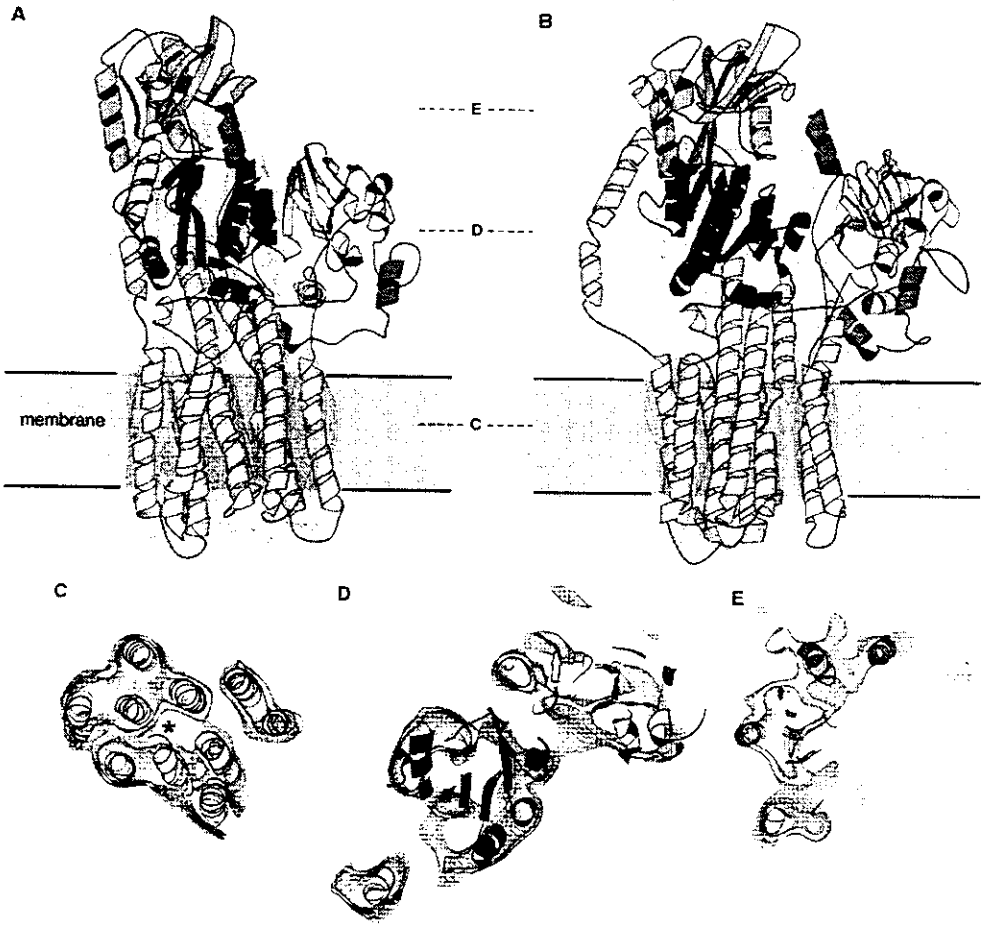



Figura 6.10: Desenho da enzima 1MHS do *Protein Data Bank*. A visualização da inserção da proteína na membrana celular em **A** e **B** é complementada por secções paralelas à membrana nos planos marcados em **C**, **D** e **E**. A perspectiva mostrada no plano **C** é a mesma da visualização em detalhe dos segmentos transmembrana da Figura 6.12.

A seqüência de amino-acidos da proteína mostrada na Figura 6.11 foi truncada no resíduo alanina A674 onde inicia-se o segmento TM5 e no resíduo isoleucina I677 onde termina o segmento TM10. O fragmento obtido é mostrado na Figura 6.12. A estrutura é visualizada no sentido lúmen–citoplasma. Este fragmento foi utilizado em cálculos de alinhamento ligante–sítio por método de *Monte Carlo Simulated Annealing* com o programa `AutoDock`, aplicando o mesmo procedimento mostrado na página 141 para o segmento TM5/6. As condições de pH foram consideradas tais que os resíduos ácidos ficam protonados.

O ligante utilizado foi o composto **igt-44** da Tabela 3.7 da página 91. Este composto foi dividido em **raiz**, **ramo** e **galhos** da mesma forma mostrada na Figura 6.7, exceto que pela ausência do grupo benzila não foram definidos os graus de liberdade N18–C19 e C19–C21. O processo resulta em seis alinhamentos que têm energia total de ligação com a proteína com valores mais baixos que $-4,00\,\mathrm{kcal.mol^{-1}}$. A geometria ligante–sítio obtida é mostrada na Figura 6.13 com todos os alinhamentos sobrepostos nas regiões dos sítios encontrados. Podemos observar as ligações de todos os alinhamentos em três sítios distintos.

Na posição inferior da Figura 6.13 temos a ligação mais efetiva no sítio denominado **Dock-1**, com energia total de ligação igual a $-5,88\,\mathrm{kcal.mol^{-1}}$. Esta é a região próxima da interface membrana–citoplasma do segmento TM9, e que também aproxima-se do laço intracelular entre os segmentos TM6/7. Na Figura 6.14 temos o detalhamento da região, mostrando os resíduos ac. aspártico D739, asparagina N740 e os resíduos serina S820, serina S821, isoleucina I822, prolina P823 e serina S824. A ponte de hidrogênio entre o nitrogênio N17 do grupo indolil o a hidroxila dos resíduos de serina S820 e S821 são as grandes responsáveis pela energia de alinhamento. Também deve ser assinalada a interação entre a região polarizada dos carbonos C5 e C6 do grupo indolil apontada pelo modelo QSAR na página 96 como de importância na atividade biológica, e a região de maior potencial eletrostático dos oxigênios da asparagina N740.

A Figura 6.15 mostra em detalhes a região do segundo sítio de ligação **Dock-2** mostrado na Figura 6.13. A energia total de ligação mais negativa foi $-4,45\,\mathrm{kcal.mol^{-1}}$ para este sítio. Mais dois alinhamentos com energias de $-4,20\,\mathrm{kcal.mol^{-1}}$ e $-4,05\,\mathrm{kcal.mol^{-1}}$ foram encontrados para este sítio, e não foram mostrados nas figura. O alinhamento mostrado é o de energia mais negativa, que tem uma grande proximidade do átomo de enxofre do ligante com a região das hidroxilas da serina S864. A Figura 6.16 mostra em detalhes a região do terceiro sítio de ligação mapeado na região transmembrana da Figura 6.13. Foram mapeados alinhamentos em duas conformações diferentes nessa região, com energia total de ligação de $-4,47\,\mathrm{kcal.mol^{-1}}$ e $-4,03\,\mathrm{kcal.mol^{-1}}$. O alinhamento mostrado é o que tem energia mais negativa, devendo certa parte da energia de estabilização do complexo à formação de pontes hidrogênio entre o nitrogênio N17 do ligante e o hidrogênio do ácido aspártico D835 e entre o hidrogênio ligado ao nitrogênio N19 e o oxigênio da ligação peptídica do leucina L799. Outra parte deve-se às interações hidrofóbicas entre o grupo carbono C17 do grupo metil ligado ao anel de tiazol do ligante e os resíduos valina V857 e triptofano W861.

A Figura 6.17 mostra em detalhes a região do sítio de ligação denominado **Dock-4**. Este sítio fica localizado na mesma região que o sítio **Dock-1** da Figura 6.13. O composto

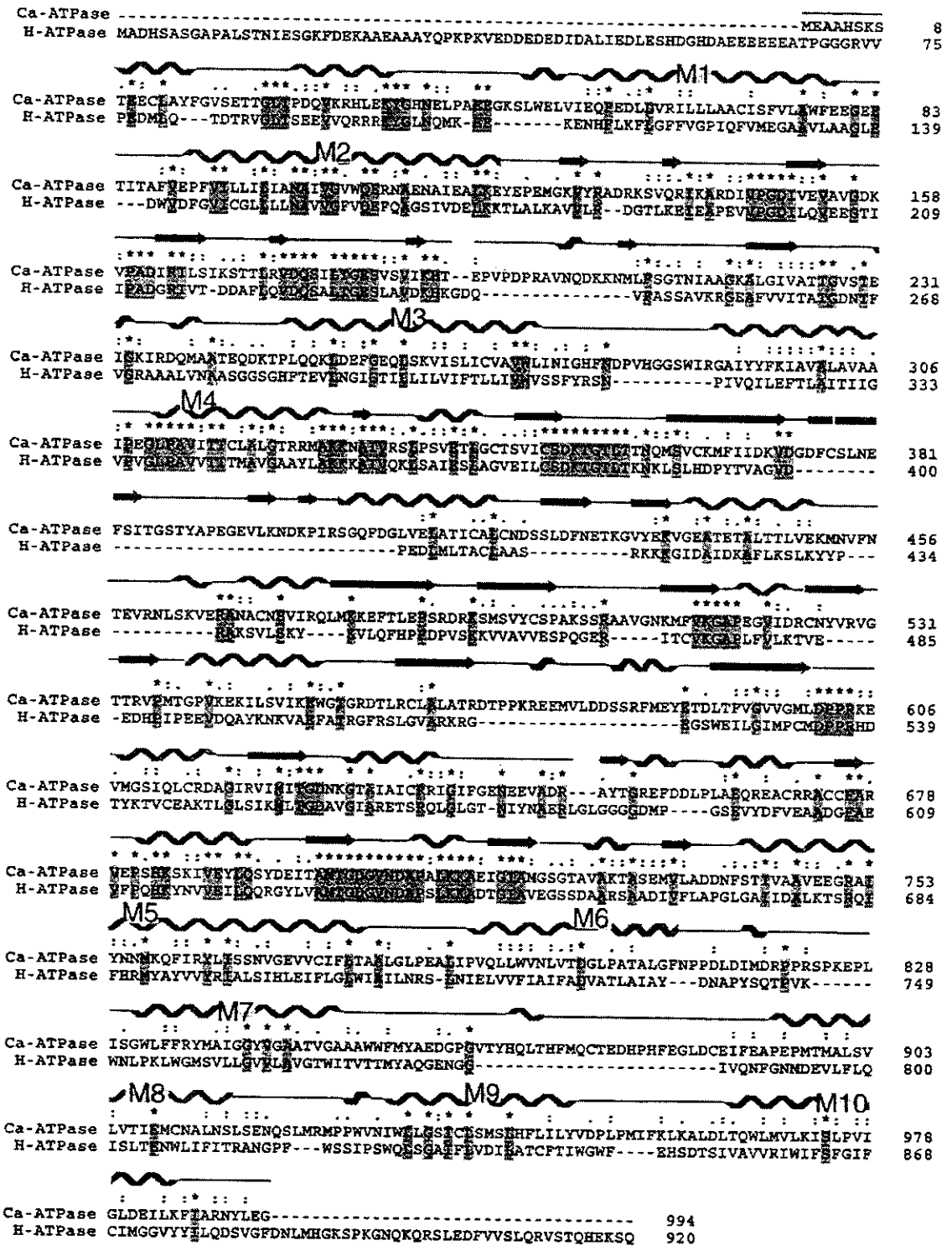



Figura 6.11: Seqüenciamento completo e homologia entre as enzimas $H^+$-ATPase e $Ca^{2+}$-ATPase. Cada par de resíduos marcados em negrito sob o sinal de asterísco corresponde a um ponto de coicidência entre resíduos das cadeias. Pares de resíduos sob o sinal 'dois pontos' correspondem a substituições muito favoráveis e os pares sob 'um ponto' àquelas favoráveis sob determinadas condições de solvatação.

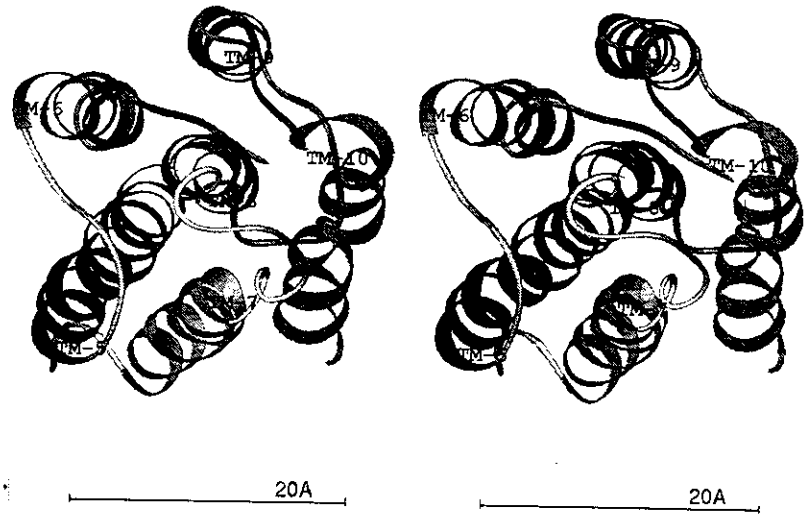



Figura 6.12: Estereograma da estrutura do fragmento contendo os resíduos compreendidos entre o início do segmento TM5 e o fim do segmento TM10, utilizado no mapeamento de sítios de ligação da proteína com o composto **igt-44** da Tabela 3.7 da página 91. O posição da visualização do estereograma é da região luminal, com o citoplasma ao fundo.

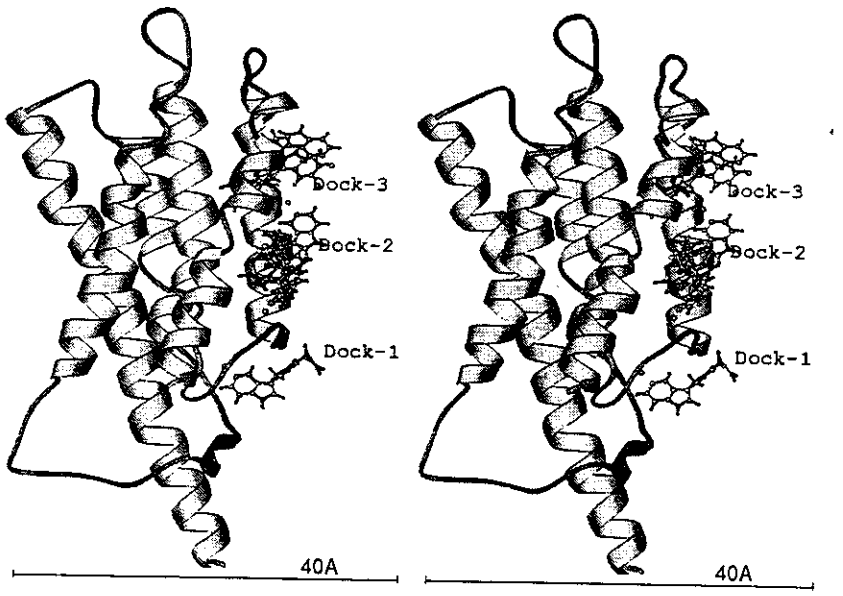



Figura 6.13: Estereograma da estrutura do fragmento da região TM5/10 onde foram mapeados três sítios de alinhamento para o ligante **igt-44** no fragmento da proteína $H^+$-ATPase com o programa `AutoDock`. Na parte inferior é mostrado o alinahmento no sítio denominado **Dock-1**, ao centro o alinhamento no sítio **Dock-2** e na parte superior o alinhamento no sítio **Dock-3**.

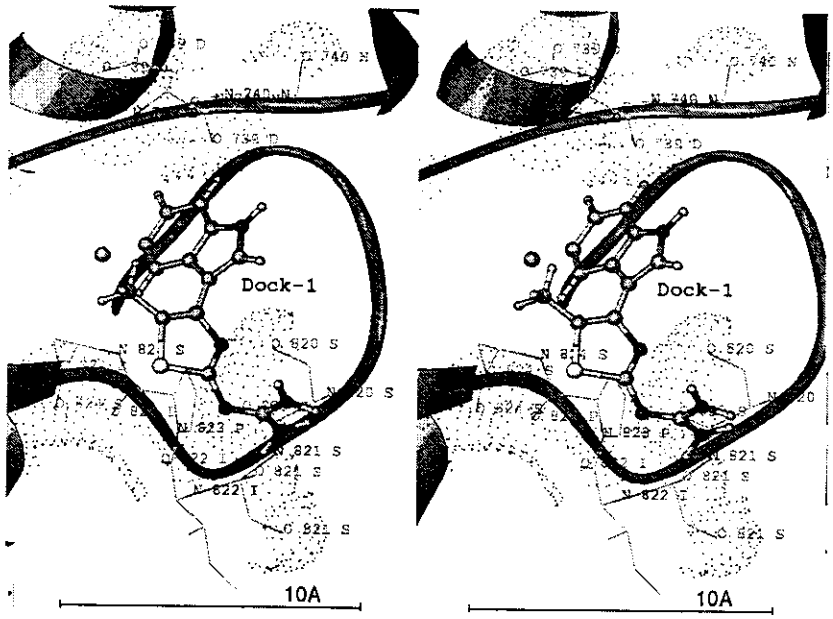



Figura 6.14: Estereograma da estrutura do fragmento da proteína $H^+$-ATPase na região do alinhamento mais favorável para o sítio **Dock-1**. A estrutura dos resíduos D739, N740, S820, S821, I822, P823 e S824 é mostrada em modelo tipo *wireframe*, e as regiões com maior potencial eletrostático de cada resíduo são mostradas com pontilhados nos raios de van der Waals. O ligante é mostrado em modelo *stick*.

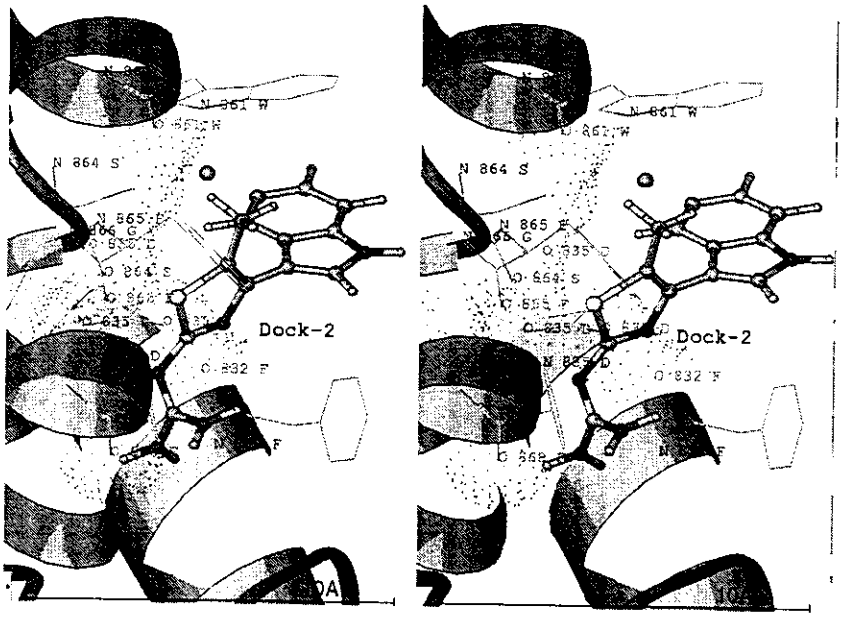



Figura 6.15: Estrutura do fragmento da $H^+$-ATPase mostrando a região do sítio **Dock-2** com o alinhamento mais favorável. A estrutura dos resíduos D739, N740, S820, S821, I822, P823 e S824 é mostrada em *wireframe*, e o maior potencial eletrostático em pontilhados. O ligante é mostrado em *stick*.

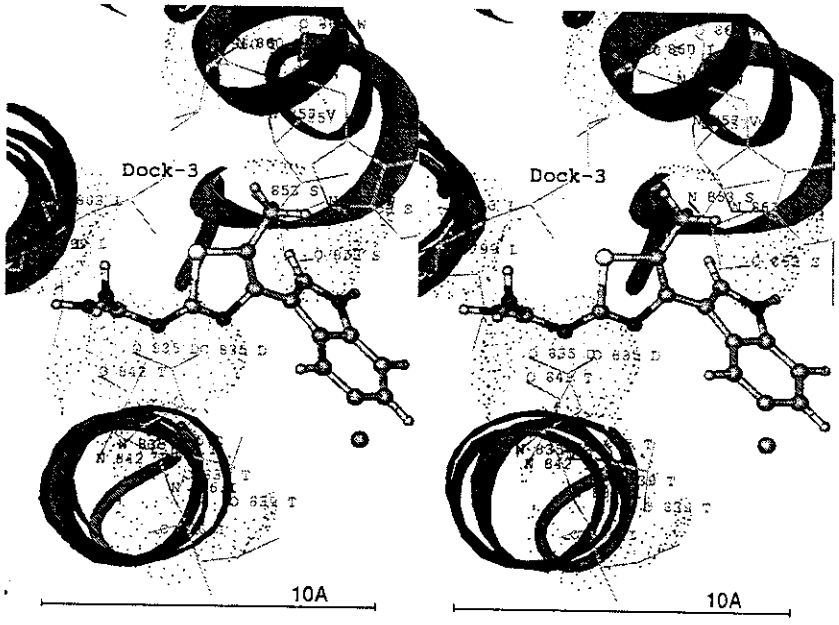



Figura 6.16: Estrutura do alinhamento mais provável no sítio **Dock-3**. A estrutura dos resíduos L799 e L803 do segmento TM-8, D835, I836, T839 e T842 do segmento TM-9 e S853, V857, I860 e W861 no segmento TM-10 são mostradas em *wireframe* recortado, e o potencial eletrostático pontilhado. O ligante é mostrado em *stick*.

**igt-45** da Tabela 3.7 da página 91 foi utilizado como ligante para obter este alinhamento. A energia total de ligação para o alinhamento é de $-7,48\,\mathrm{kcal.mol^{-1}}$. As principais contribuições para a estabilização do complexosão: A ponte de hidrogênio formada entre o nitrogênio do anel de tiazol e o hidrogênio ligado ao nitrogênio da cadeia lateral da arginina 813. O hidrogênio formada ligado ao nitrogênio do anel de indolil também fica a distâncias que permitem a formação de pontes de hidrogênio com as carbonilas da ligação peptídica da tirosina 738 e da cadeia lateral da asparagina 740. também podem estabilizar o complexo as interações do tipo hidrofóbico entre a região do anel de indolil do ligante e o anel de fenol da tirosina 738.

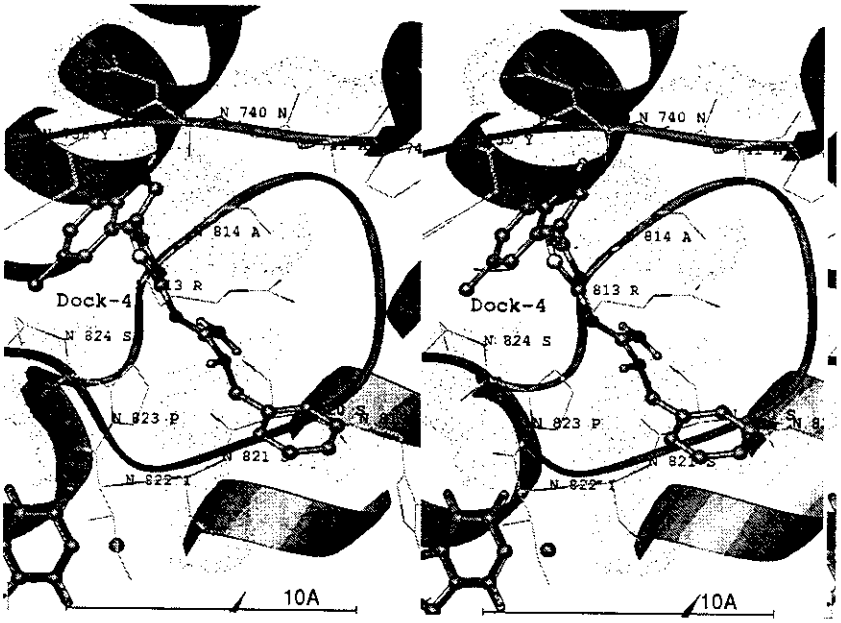



Figura 6.17: Estrutura do alinhamento mais provável do composto **igt-45** na mesma região do sítio **Dock-1** da Figura 6.13, aqui denominado **Dock 4**. A estrutura dos resíduos Y738, N740 e R813 é mostrada em *wireframe* e o potencial eletrostático em pontilhado e o ligante é mostrado em *stick*.

### 6.5.1 Conclusões sobre alinhamento

- A inserção dos segmentos transmembrana na membrana celular está relacionada à atividade das enzimas de transporte de prótons estudadas.
- É possível estudar aspectos relevantes das interações ligante–proteína utilizando fragmentos da proteína completa.
- O método de alinhamento *Monte Carlo Simulated Annealing* permite um mapeamento eficiente da superfície da proteína. Aparentemente não resulta alinhamentos aceitáveis em regiões mais internas da macromolécula.

# Capítulo 7

# Estudo QSAR de Inibidores da Secreção Gástrica e Simulação Molecular da Inibição

## 7.1 O desenvolvimento de modelos

O estudo de QSAR pretende ser uma ferramenta para auxiliar no desenvolvimento racional de drogas. Partido da estrutura da droga, devemos relacionar quais são as características moleculares relevantes para que uma dada série de compostos químicos, e que são responsávies pela sua atividade biológica. Consideramos aqui que a atividade biológica é obtida pela ligação de uma droga a um sítio específico em uma proteína, e que esta ligação é a única responsável pelas alterações no metabolismo que resultam na atividade biológica. Temos então conjuntos de características que devem ser direcionadas para que a interação entre a droga e o receptor sejam as mais efetivas.

Este processo é conhecido como QSAR baseado na droga, e tem diversas limitações. Uma primeira limitação é que uma mesma droga pode ter (e na maioria dos casos tem) vários sítios de atividade, numa mesma proteína ou em diversas proteínas, que fazem parte ou não do mesmo ciclo metabólico relacionado àquela atividade biológica que nos interessa. Esta limitação é crítica, porque baseando-se em métodos estatísticos para reconhecimento das características da droga que relacionam-se a uma dada atividade, podemos ser confundidos com características relativas a sua ligação com qualquer outro sítio receptor que tenha alguma participação no ciclo metabólico que produz aquela atividade que é de interesse. Nos casos estudados, o exemplo é pertinente. As drogas das séries dos fenil e indolil-guanidinotiazóis têm ação em pelo menos duas enzimas ligadas ao processo da secreção ácida do estômago: A $H^+$,$K^+$-ATPase e os Receptores H2 de histamina. As séries dos compostos derivados de quinolina e nicotinamida aparentemente não se ligam aos receptores H2 de histamina.

Uma possibilidade de contornar esta limitação é trabalhar com dados *in vitro* da afinidade da droga por uma dada enzima, freqüentemente obtidos como a concentração necessária para inibição de 50% da atividade da enzima ($IC_{50}$). Fazendo desta maneira, estamos limitados àquela enzima, que ainda assim pode ter mais de um sítio receptor para

a mesma droga. Também estamos limitados a relacionar as características que fazem da droga um bom ligante, ignorando completamente todas as demais que compõem o quadro geral da atividade *in vivo*. No caso específico de drogas administradas por via oral, temos que as condições de pH do estômago são uma barreira capaz de promover a degradação de certa parte das drogas. No caso de qualquer medicamento, temos que mecanismos de complexação com componentes do plasma sangüíneo, metabolização (particularmente no fígado em um complexo enzimático denominado citocromo P-450) e excreção diminuem os níveis séricos das drogas, e sua eficiência.

As possibilidades para desenvolvimento de QSAR baseado na droga dependem também da acuracidade da descrição da droga pelo conjunto descritor escolhido. Fatores estéricos, de lipofilicidade, eletrônicos e eletrostáticos participam concomitantemente do equilíbrio químico que denominamos afinidade droga-enzima. Um conjunto de propriedades que seja capaz de descrever acuradamente estas propriedades em uma droga tem melhores chances de resultar modelos que não sejam apenas fruto do acaso. Modelos casuais são perfeitamente possíveis, tratando-se de métodos estatísticos de regressão multivariada bastante complexos, em que muitos descritores são relacionados a um conjunto relativamente pequeno de atividades biológicas das drogas.

A primeira etapa do trabalho realizado foi justamente uma validação dos descritores que foram aplicados nas etapas seguintes. Uma propriedade física relativamente simples de ser obtida, disponível na literatura para um grande número de compostos orgânicos e que tem relação com fatores estéricos, lipofílicos e eletrostáticos é o coeficiente de partição octanol/água, $logP_{o/w}$. A validação do conjunto descritor foi feita construindo um modelo QSPR para $logP_{o/w}$ utilizando os descritores mecânico-quânticos calculados. As mesmas metodologias estatísticas e de selação de variáveis que são aplicadas na elaboração do modelo QSPR são utilizados ao longo do trabalho de modelagem QSAR. Os resultados obtidos em validação cruzada para este modelo QSPR tem significância estatística, e uma estimativa de erro de previsão da mesma ordem daquela obtida na mediçao experimental dos dados de $logP_{o/w}$, permitindo concluir que os descritores aplicados são de fato capazes de descrever acuradamente as interações presentes no equilíbrio que resulta o valor de $logP_{o/w}$. Considerando-se esta validação, foi decidido que seriam utilizados apenas descritores de origem mecânico-quântica com o objetivo de explorar as possibilidades da descrição de valores de atividades biológicas, e também as possibilidades de relacionar fatores estruturais aos descritores mecânico-quânticos, para viabilizar a sugestão de alterações nas estruturas das drogas para aumentar a sua afinidade pelo sítio de ligação na proteína alvo.

Os modelos QSAR construídos baseando-se nos descritores mecânico-quânticos mostraram-se capazes de prever valores de $\log IC_{50}$ para os compostos das séries estudadas, dentro de critérios estatísticos que confirmam a validade dos modelos. As previsões, no entanto, não têm uma uma estimativa de desvio padrão suficientemente pequena para determinar qual seria o mais ativo entre um conjunto de compostos de alta atividade. Os menores valores de $IC_{50}$ correspondem às afinidades mais altas entre droga-enzima, situando-se na casa de variações nanomolares da concentração das drogas. Portanto os modelos denominados quantitativos são na verdade semi-quantitativos, apontando a direção para maior afinidade, sem contudo poder concluir sobre qual seria o melhor ligante.

Então estes modelos devem ser utilizados como uma análise discriminante, em que se tem como resposta real a classificação dos compostos em bons e melhores ligantes, de acordo com o valor previsto de $IC_{50}$ para cada droga. Esta limitação não é crítica, pois o surgimento de novas drogas mais potentes pode ser utilizado na reconstrução dos modelos de regressão, baseados nos mesmos conjuntos de descritores já selecionados ou através de uma nova seleção, com exclusão de moléculas menos potentes. Procedendo desta maneira pode-se refinar os modelos, selecionando cada vez mais as características que tornam a molécula um bom ligante e excluindo as demais. Uma observação é que bons ligantes têm entre sí grande similaridade: Compartilham as propriedades necessárias para que sejam perfeitamente complementares ao sítio de ligação da proteína alvo. Ligantes menos efetivos têm conjuntos de características diversas que impedem a perfeita complementaridade, que são indefinidos. Qualquer propriedade pode ser capaz de impedir a complementaridade, não havendo limite *a priori* para a dissimilaridade entre moléculas.

Também as decodificação dos descritores mecânico-quânticos em estruturas químicas não é inequívoca. Grupamentos de átomos diferentes podem ter propriedades semelhantes em relação à complementaridade com o receptor, um fenômeno denominado bioisosterismo. A limitação da decodificação dos descritores em estruturas químicas é um problema menor, podendo ser considerada uma potencialidade, já que baseando-se em bioisosterismo podemos propor novas estruturas de moléculas pela substituição de átomos ou ligantes e realizar os cálculos dos descritores mecânico-quânticos, introduzindo estes dados nos modelos criados.

A inclusão de novos elementos em modelos já definidos representa pouco esforço, se comparado a síntese de uma nova molécula. Os resultados de cálculos obtidos para novas moléculas potencialmente ativas podem então ser analizados dentro dos modelos QSAR para decidir sobre qual caminho é mais promissor, principalmente pela eliminação dos compostos considerados de pouca afinidade pelo modelo. Muitos fatores serão levados em conta ao decidir sobre a síntese de um novo composto, a dificuldade e o custo estão entre os principais.

No caso específico dos compostos inibidores da enzima $H^+$,$K^+$-ATPase da série das nicotinamidas, que têm um mecanismo de ligação que depende da transformação da droga administrada, uma pró-droga, na espécie que realmente liga-se à enzima dentro do organismo, nas condições do meio gástrico, modelos QSAR com valores das atividades *in vitro* e *in vivo* baseados na estrutura química da pró-droga não puderam ser obtidos. Estes resultados mostram a necessidade de entender melhor o mecanismo de transformação da pró-droga antes de prosseguir no estudo de sua afinidade.

Um mecanismo publicado para a ligação da forma ativa ao sítio receptor pôde ser reforçado, ainda que utilizando cálculos mecânico-quânticos de baixa qualidade. Os cálculos descartam outro mecanismo, também publicado, para a reação de transformação da pró-droga na espécie bioativa. Este estudo define as geometrias dos intermediários que podem efetivamente ligar-se ao sítio de ação da droga segundo o mecanismo proposto, necessitando ainda de refinamento para que os resultados possam ser correlacionados aos tempos de meia vida ($t_{1/2}$) dos compostos em meio ácido, uma forma de mensurar a capacidade de cada molécula para transformar-se na forma bioativa. Devemos lembrar-nos que os tempos de meia vida são uma medição realizada da taxa da decomposição dos

compostos, não necessariamente através do mecanismo que passa pelos intermediários capazes de ligar-se à enzima, e portanto devem ser considerados com cautela quando comparados com uma única proposta para o mecanismo de decomposição.

O estudo detalhado do mecanismo da decomposição de cada uma das moléculas da série é necessário para que sejam obtidos resultados que auxiliem o desenvolvimento de drogas que decomponham-se pelo mecanismo adequado à ligação com o receptor em desejado. Os resultados obtidos são apenas um primeiro passo, apenas a obtenção das geometrias de dos estados durante a decomposição do melhor ligante da série. Estes resultados também podem ser utilizados de outras formas. O conhecimento detalhado da estrutura do sítio de ligação pode permitir a aplicação imediata das geometrias obtidas para o estado de transição dos intermediários de reação em cálculos de afinidade droga-enzima, por alguma metodologia capaz de considerar os efeitos relevantes na formação de ligação química. Particularmente os efeitos eletrônicos devem ser acuradamente descritos, visto que estes compostos são inibidores irreversíveis da enzima, e ligam-se covalentemente ao sítio. O estudo detalhado da interação eletrônica entre pequenas moléculas e macromoléculas constitui um campo aberto ao desenvolvimento teórico e tecnológico.

O estudo da interação entre a droga e um determinado sítio receptor, num processo conhecido com QSAR baseado no sítio, é o caso mais favorável ao desenvolvimento racional de novas drogas. Diversas metodologias podem ser exploradas para obter alinhamentos entre drogas e macromoléculas e calcular a energia de ligação da geometria resultante. A aplicação das metodologias de alinhamento pressupõe o conhecimento da estrutura da droga e da macromolécula (ou pelo menos a região do sítio de ligação). Grande parte das metodologias aplica campos de força para calcular a energia de interação através de mecânica molecular, com resultados que dependem da acuracidade do compo de forças determinado. Também dependem da capacidade de sortear e encontrar os alinhamentos favoráveis.

Uma boa descrição molecular através de campos de força requer dados de origem espectroscópica e/ou cálculos teóricos de nível muito elevado para os compostos, para determinar constantes de força de ligação, torsionais e tantas outras quantas o campo de força escolhido necessite para caracterizar cada tipo atômico presente na molécula. Conhecer quais são as regiões da molécula mais relevantes na ligação droga-macromolécula pode ser de grande valia ao se propor a definição de novos tipos atômicos para utilização em campos de força, já que muitas vezes são definidos fragmentos contendo dois ou mais átomos (por exemplo o grupo $CH_3$ terminal em uma alanina, ou o grupo CH aromático em uma fenil-alanina). Este ponto aproxima as abordagens baseadas na droga das abordagens baseadas no sítio.

Diante da impossibilidade de realizar cálculos de mecânica quântica envolvendo grandes porções de macromoléculas, devemos pelò menos ser capazes de discriminar alguns alinhamentos droga-macromolécula potencialmente importantes, para então partir para um refinamento. Se um campo de forças for bem definido para os amino-ácidos que compõem a proteína (e muitos o são), resta definir bons parâmetros para a droga, e assim obter alinhamentos promissores. Portanto os dados obtidos determinando as regiões de ligação da droga em estudos QSAR baseados na droga podem e devem ser utilizados para auxiliar na definição da topologia das drogas, se houver necessidade de definir novos tipos

de átomos para utilizar a droga em alinhamentos.

## 7.2 Considerações finais sobre o trabalho

Fica claro que a utilização de descritores de origem mecânico-quântica é viável e eficiente para obtenção de modelos QSPR e QSAR no caso dos compostos utilizados para previsão de $logP_{o/w}$ de compostos com funcionalização diversa, e dos valores de $IC_{50}$ das drogas estudadas. A utilização conjunta de descritores de outro tipo, de conectividade ou topológicos por exemplo, poderia trazer melhoras na capacidade de previsão dos modelos. O motivo da sua não inclusão é a simplicidade. Utilizar apenas um tipo de descritor pode resultar numa melhor atribuição das relações entre os descritores mecânico-quânticos e as estruturas químicas, servindo ao propósito de melhor conhecer o significado dos descritores escolhidos.

A possibilidade de utilização dos descritores mecânico-quânticos para proposição de alterações estruturais não é tão clara. No caso dos compostos utilizados para criação do modelo para $logP_{o/w}$ podemos afirmar apenas quais são as principais classes de compostos pior representadas com o modelo proposto, sem uma conclusão sobre a relação entre os descritores e as características apontadas para os tipos de compostos que têm previsão ruim (grupos hidroxilados e/ou volumosos). As atribuições das limitações dos descritores que podem relacionar-se à má previsão são circusntancias. Sabemos que cargas permanentes e momentum de dipolo elevados são associados a baixa solubilidade em fase apolar, e que grandes volumes moleculares desestabilizam solutos que têm pouca interação com solventes polares nessa fase. Podemos considerar que a má previsão de grupos hidroxilados, capazes de realizar pontes de hidrogênio efetivas, aponta para uma representação pobre da formação de pontes de hidrogênio pelo modelo de solvatação tipo *continum* D-PCM. Consideradas assim, as relações entre estrutura e propriedade são razoáveis do ponto de vista químico, ainda que obtidas baseando-se em modelos restritos a algumas classes de compostos, apresentados sem uma extensa discussão teórica que mostre claramente como cada descritor representa cada tipo de interação intermolecular presente.

No caso dos modelos QSAR dos compostos fenilguanidinotiazóis, temos que os descritores apontados como mais relevantes são os momenta, polarizabilidades, cargas líquidas em algumas posições atômicas e densidades eletrônicas em orbitais de fronteira em algumas posições atômicas. Sem o conhecimento da estrutura complementar à droga no sítio receptor, as cargas, os momenta e polarizabilidades devem ser associadas a facilidade de penetrar na membrana, fato que também tem sentido químico, pois sabe-se que para ligar-se à $H^+$,$K^+$-ATPase, mesmo em sua face luminal, a droga deve ser capaz de penetrar na membrana ou na porção lipofílica da enzima, já que os sítios de ligação propostos não situam-se exatamente na face luminal (como pode fazer pensar o desenho da Figura 6.1 da página 130). Também é mostrado que diferenças na lipofilicidade das regiões transmembrana onde encontram-se sítios putativos de ligação, obtidas por mutações de determinados resíduos, têm relação com a atividade da droga SCH28080 na enzima, afirmando a possibilidade da importância de características ligadas à permeabilidade e

lipofilicidade.

As densidades em orbitais de fronteira, sem o conhecimento de alguma estrutura que seja capaz de interagir com as drogas através de pontes de hidrogênio ou outro tipo de interação eletrônica devem ser consideradas apenas indicadores de substituição no grupo amina, se primária ou secundária, fazendo aqui o papel de um descritor topológico. Também as cargas na região do anel aromático podem estar mostrando a influência dos substituintes do anel, portando-se como descritores topológicos.

A inclusão dos efeitos de solvatação no modelo QSAR dos fenilguanidinotiazóis não melhorou significativamente os resultados, resultando a seleção de descritores bastante parecidos aos utilizados na construção do modelo em vácuo, em número ligeiramente menor. O aumento da complexidade dos cálculos aparentemente não trouxe a contrapartida da melhora da representação molecular pelos descritores calculados.

No caso do modelo QSAR para os compostos indoliguanidinotiazóis a utilização de modelos concorrentes, elaborados a partir de conformações diferente, permitiu optar pela conformação mais extendida da molécula, como sendo a que melhor adaptaria-se ao sítio de ligação na enzima, sem conhecer a estrutura do sítio. Os melhores resultados de correlação entre o modelo obtido utiliando descritores calculados na conformação extendida foi utilizado como critério para optar pela conformação extendida, já que através variação da energia conformacional elas seriam indistinguiveis. A decisão mostrou-se acertada, já que as dimensões da conformação extendida são as que melhor se aproximam das dimensões de outros tipos de ligantes e das distâncias estimadas entre alguns pontos de interação da estrutura do farmacóforo, proposta para o sítio receptor por diversos autores independentemente.

Os descritores relacionados são momenta, campos elétricos, cargas líquidas sobre átomos, densidades eletrônicas em orbitais de fronteira e energia de orbitais de fronteira. Aos momenta, campos elétricos e cargas atômicas é razoável atribuir a descrição da lipofilicidade, pelos mesmos motivos que no caso dos fenilguanidinotiazóis. Às densidades eletrônicas nos orbitais de fronteira e energia dos orbitais aparentemente não podem ser atribuído nenhum significado, sem o conhecimento da estrutura complementar em que a droga deve ligar-se. Uma hipótese é que o orbital HOMO-10 seja responsável por restrições conformacionais na região de ligação entre os anéis indolil e tiazol. os resultados obtidos na previsão da atividade de compostos desconhecidos, definidos com base em bioisosterismo e para os quais foram realizados cálculos mecânico-quânticos mostram a viabilidade da utilização dos modelos como auxiliar do desenvolvimento racional de drogas. Os resultados obtidos durante a modelagem desta classe de compostos também foram úteis para definir sua topologia na definição de tipos atômicas, necessária aos cálculso de alinhamento realizados posteriormente para verificar se os descritores calculados atendem ao quesito da complementaridade em regiões da macromolécula que contêm o sítio receptor.

O estudo QSAR dos comostos da séria das quinolinas mostra resultados mais consistentes. Os modelos obtidos incluem moléculas com variações estruturais relativamente grandes e a validade estatística dos modelos é maior devido à grande quantidade de compostos da série. As conclusões sobre o significado de cada descritor e baseiam-se em suposições sobre a porção complementar, no sítio receptor. Temos que regiões onde a

carga líquida nas posições atômicas alternam-se com densidade eletrônicas totais indicam relevância de fatores eletrostáticos na sua ligação. As energias dos orbitais de fronteira apontam para orbitais com densidade eletrônica na região do anel benzênico, da carbonila e na região de ligação entre os anéis. Sem conhecer a estrutura do receptor, pode-se presumir que estas regiões podem ligar-se a sítios aceptores de elétrons, e que no caso da densidade na ligação entre os anéis relaciona-se a restrição da liberdade rotacional.

O estudo mecanístico da reação de eliminação e formação dos intermediários ativos dos derivados de nicotinamida mostra grande conformidade em relação ao mecanismo proposto para a reação, mesmo utilizando cálculos de baixo nível. O tamanho da molécula e sua estrutura torna a realização de cálculos com bases maiores muito dispendiosa, ou que sejam realizados cálculos com apenas uma parte da molécula, eliminando substituintes. No caso praticular do composto mostrado, o mais ativo, as aminas terciárias substituídas nos anéis participam do formação do estado de transição, alternando-se ao modo de vibração na região do enxofre durante o busca do estado de transição.

O desenvolvimento do estudo do mecanismo está relacionado á utilização de métodos capazes de quebrar da molécula em fragmentos que utilizem níveis diferentes de cálculo, para que se possa aumentar significativamente a representação na região de interesse, próxima ao átomo de enxofre, sem aumentar desnecessariamente o número de primitavas utilizadas, tornando os cálculos menos dispendiosos. Este método, denominado ONIOM, pode permitir a realização dos cálculos para um conjunto significativo de compostos, de maneira que seja possível relacionar a energia de ativação de cada uma a sua estabilidade relativa, mediada pelo tempo de meia vida, apoiando o mecanismo proposto na literatura através dos cálculos mecânico-quânticos.

A obtenção de alinhamentos entre as drogas inibidoras da secreção gástrica e as macromoléculas que contém os receptores deve solucionar diversos problemas cruciais antes de tornar-se, de fato, um mecanismo para o planejamento racional de drogas. As enzimas envolvidas nos ciclos bioquímicos da secreção ácida são proteínas inseridas na membrana celular, com porções lipofílicas grandes, que dificilmente cristalizam-se para permitir análises por Raios-X ou difração de neutrons, por exemplo. A solução é obter as estruturas terciárias destas proteínas através de modelagem molecular. Trata-se um trabalho arriscado, por quê não há como certificar-se que a estrutura obtida por modelagem corresponde àquela que realiza as funções no organismo vivo De fato, não é possível certificar-se que a estrutura foi completamente preservada nem mesmo para aquelas que cristalizam-se e são mapeados espectroscopicamente.

O depósito da estrutura da enzima $H^+$-ATPase do organismo *Neurospora crassa* inicia uma nova fase no estudo e desenvolvimento das drogas relacionadas, todavia. A estrutura terciária da enzima pode servir de base para a construção das outras enzimas com alto grau de homologia, como a $H^+$,$K^+$-ATPase humana. A construção das enzimas a partir de outras é um procedimento complicado. A utilização de mecânica molecular para minimização da estrutura após a substituição e/ou inclusão de novos resíduos estrutura pode afetar significativamente sua estrutura terciária, mesmo se realizado com critério e cuidado.

A construção dos fragmentos transmembrana da enzima $H^+$,$K^+$-ATPase humana foi realizada a partir do seqüenciamento da enzima e de diversas estruturas da enzima publi-

cadas, em resolução menor que a necessária para mostrar a posição de cada átomo, porém suficiente para mostrar como é o alinhamento entre as hélices transmembrana, particularmente. Desta forma, as porções terminais das $\alpha$-hélices construídas são posicionadas para reproduzir as distâncias obtidas a partir das estruturas publicadas e mantidas na posição pela aplicação de interações entre átomos não ligados. O restante da estrutura tem sua estrutura otimizada pela minimização da energia total, de forma que sejam evitadas grandes repulsões entre as cadeis laterais dos amino-ácidos. Um gráfico de Ramachandran é utilizado para certificar-se que também as ligações peptídicas situam-se em ângulos aceitáveis para cada tipo de amino-ácido. Atingido um grau de convergência, as restrições de distância entre as extremidades das cadeias são relaxadas e alguns ciclos mais de minimização são executados. Poucos ciclos podem não ser suficientes para corrigir as distorções moleculares introduzidas pela restrição de distância entre átomos não ligados, e muitos ciclos certamente distanciam a estrutura daquela que se pretende construir. Assim podem ser modificadas as seqüências de amino-ácidos das estruturas de proteínas obtidas em bancos de dados para produzir outras proteínas, ainda que não seja possível comprovar a perfeição destas novas estruturas construídas.

Construindo fragmentos através do procedimento descrito, e aplicando campos de força consagrados, foi possível realizar modelos que mostraram a importância da lipofilicidade das regiões transmembrana, na região do sítio receptor putativo das drogas nestas enzimas. Esta importância era óbvia e foi posteriormente reportada na literatura, porém a conclusão obtida através dos modelos criados permite acreditar na veracidade dos modelos construídos. A observação das regiões mais e menos lipofílicas nos segmentos mostrados nas páginas 139 e 137 e sua relação com a afinidade da droga SCH28080 a estas enzimas em estudos publicados em diversas revistas de grande aceitação é um resultado interessante, mostrando que a metodologia aplicada na construção dos fragmentos pode ser testada em outros casos.

No entanto, obter alinhamentos com pequenos fragmentos da enzima não parece ser razoável. Uma grande proximidade entre as cadeias laterais dos amino-acidos na enzima completa pode invalidar completamente as tentativas de alinhamento como aquela mostrada na Figura 6.9 da página 142 para um compostos da série dos indolilguanidinotiazóis e o fragmento transmembrana TM5/6 da enzima. No caso o trabalho de alinhamento foi realizado apenas para teste do procedimento e do programa `AutoDock`. A partir da aplicação do procedimento da definição das características e tipos de átomos da droga, que foi escolhida por quê tem o fragmento indolil facilmente adaptado da estrutura do aminio-ácido triptofano, pode-se partir para um estudo com um fragmento maior da enzima, obtido diretamente pela edição do arquivo de coordenadas no formato `.pdb` para a $H^+$-ATPase de *Neurospora crassa* disponibilizada no *Protein Data Bank.*

Os alinhamentos obtidos utilizando o mesmo procedimento resultaram em energias de ligação bastante pequenas, possivelmente por causa da má representação de pontes de hidrogênio. A alternativa é utilizar estes alinhamentos em métodos de cálculo alternativos, valendo-se de campos de força que sejam capazes de uma melhor representação de todos os aspectos da interação molécula-macromolécula, ou mesmo por métodos que utilizam mecânica-quântica na região de ligação da droga no receptor e mecânica molecular no restante da macromolécula, conhecidos com QMMM. A situação permite mostrar que a

metodologia aplicada para sorteio de alinhamentos e conformações pode ser eficiente para mapear diversos sítios possíveis de ligação ao longo da superfície da enzima, sem contudo permitir uma definição de qual é o melhor alinhamento, utilizando o programa `AutoDock` e seu método de cálculo de energia de alinhamento.

## 7.3 Conclusões finais

O trabalho realizado desenvolveu um modelo QSPR para $logP_{o/w}$ com estimativa de desvio comparável àquela obtida por métodos experimentais. Podemso concluir que os descritores mecânico-quânticos propostos são adequados ao desenvolvimento dos modelos. É bastante relevante mostrar que através de cálculos mecânico-quânticos podemos modelar propriedades termodinâmicas com grande conteúdo entrópico, que tradicionalmente são modeladas pela aplicação de métodos de dinâmica molecular utilizando mecânica molecular para obter os valores das energias de interação entre átomos e moléculas.

Os resultados de QSAR obtidos têm possibilidades de utilização direta como auxiliar na decisão sobre a síntese de novos compostos das séries estudadas. O trabalho mostra inequivocamente que os descritores mecânico-quânticos são suficientemente acurados para obter modelos QSAR, e que os fatores eletrostáticos e eletrônicos estão entre os mais relevantes na interação entre as drogas ativas na modulação da secreção gástrica e seus sítios de ação. A modesta capacidade de sugestão de novas estruturas a partir da análise multivariada pode ser compensada pela facilidade de obtenção de valores de atividades de outros novos compostos através dos modelos propostos. Os modelos mostram-se mais robustos com o aumento dos conjuntos de moléculas utilizados para sua elaboração, o que permite considerar que são elaborados a partir de relações pertinentes e não casuais entre os descritores mecânico-quânticos e as atividades biológicas medidas *in vitro*.

Os modelos mecanísticos podem ser refinados, tratando-se de um problema de ordem computacional obter resultados com a aplicação de cálculos mecânico-quânticos de alto nível para os casos estudados. Os procedimentos para realização dos cálculos e os detalhes da aplicação dos métodos foram resolvidos.

Os procedimentos de alinhamento são apenas etapas iniciais de um trabalho que deve prosseguir para que sejam obtidos resultados úteis para orientar o desenvolvimento racional de novas drogas. Trata-se de um campo aberto de pesquisa, não cabendo aqui conclusões de fato. A aplicação dos métodos disponíveis em diversos programas e o interfaceamento entre eles foi a etapa desenvolvida durante este trabalho. Foram estabelecidos os procedimentos para realização da varredura da superfície da enzima em busca de sítios de ligação e para a obtenção da energia associada ao alinhamento, assim como o procedimento para obtenção dos gráficos que representam os resultados de alinhamento. Estes procedimentos utilizam programas fornecidos gratuitamente pelos autores, podendo ser úteis a toda a comunidade, independentemente da disponibilidade de recursos para aquisição de programas, fato relevante.

# Apêndice A

# Programas desenvolvidos

## A.1 Geração de geometrias com o método da Matriz de distâncias métricas

O processo de geração das geometrias a partir de uma semente utilizando o módulo `distgeom` do `TINKER` envolve diversas conversões entre os formatos de arquivos de coordenadas dos átomos nas moléculas. Para gerenciar e automatizar o processo foram realizados *shell scripts* na linguagem *PERL*. Os programas de geração de geometrias foram desenvolvidos por outro pesquisador [99], associado ao grupo, e foram tanscritos aqui apenas para documentação

Este programa utiliza um arquivo de entrada de dados que deve conter o nome do arquivo de semente que será utilizado para a geração das estruturas e algumas opções de entrada do programa `distgeom`. O arquivo com os dados de entrada deve ser chamado de `csp.in`, e um exemplo com o formato está transcrito a seguir. Este *script* utiliza os programas `MOPAC`, `BABEL` [158] e `TINKER`, que devem estar acessíveis no diretório `/usr/local/bin` do computador.

### cs.pl

```
#!/usr/bin/perl -w
#
$input=$ARGV[0];
#
################################################
#Settings - This program needs babel-1.6 and Mopac
#
$tinker_dir="/usr/local/bin";
$babel_dir="/usr/local/bin";
$mopac_dir="/usr/local/bin";
#
# Keywords to mopac
#
$keywords='AM1 PRECISE NOINTER MMOK';
#
# Send an e-mail to me after finishing
#
$email="eborges\@iqm.unicamp.br";
#
################################################
```

```
#
#########################################################################
#Structure files are generated by TINKER using distance geometry methods
#
 `cat $input | $tinker_dir/distgeom > cs.log`;
#
#########################################################################
#
################################################
#Root names are extracted and storaged in a vector
#
$tmp= `ls *.001`;
$tmp=~ s/.001//;
chomp($tmp);
@tmp=<$tmp*>;
pop(@tmp);
#
#################################################################
#Each tinker file is converted to mopac internal coordinate file
#in order to minimize its geometry
#
while (@tmp){
    $arq=shift(@tmp);
    `$tinker_dir/xyzsybyl $arq  > gbg`;
    ($babel_in,$babel_out)=($tmp.".mol2",$tmp.".dat");
    `$babel_dir/babel -imol2 $babel_in  -omopint $babel_out "$keywords" > gbg`;
    `$mopac_dir/rmopac $tmp`;
    rename($tmp.".out",$arq.".out");
#
#Remove some rubbish
#
    unlink($tmp.".arc");
    unlink($tmp.".dat");
    unlink($tmp.".log");
    unlink($tmp.".mol2");
}
#################################################################
#
#########################
#Remove some rubbish again
#
unlink("gbg");
#########################

open MAIL,"|mail $email";
     print MAIL "Terminou a analise conformacional\n";
close MAIL;
```

A seguir é mostrado um exemplo de arquivo de entrada de parâmetros para geração de geometrias através do programa `cs.pl`, que utilza como semente o arquivo `22.xyz`, que tem corrdenadas cartesianas no formato do TINKER, podendo ser gerado com o programa `molden` [159], ou convertido com o BABEL para o formato necessário. O programa vai gerar mil conformações, com imposição de quiralidade em átomos tetraédricos, imposição de quiralidade ou planaridade em átomos trigonais, imposição de restrição na liberdade rotacional da ligação entre dois átomos trigonais e pré otimização com método de *Simulated Annealing*.

**csp.in**

`22.xyz`

```
1000
y
n
y
y
n
n
A
```

### A.1.1 Programa para comparar, eliminar e classificar confôrmeros

**uniquefy.pl**

```
#$Id: uniquefy.pl,v 1.1 2000/07/12 01:42:20 kesley Exp kesley $#
#$Log: uniquefy.pl,v $
#Revision 1.1  2000/07/12 01:42:20  kesley
#Initial revision
require('getopts.pl');
Getopts('te:');
#####################################################
#This program is appropriated for Mopac output files
#####################################################
$babel_dir="/usr/local/bin";
$tinker_dir="/usr/local/bin";
$tinker_par="/usr/local/share/tinker/params/mm3.prm";
$email="eborges\@iqm.unicamp.br";
#
###
# For field_1 two Options are Available :  (1) Fit atoms "M" through "N" from structur
e 1
# to the corresponding atoms of structure 2. Enter "1,M,N" to use this option.
# If "N" is omitted, the fit uses atoms 1 through "M". If both "M" and "N" are
# omitted, the fit uses all atoms; or (2) Individual entry of atom range pairs
# to be used in the fitting procedure.
###
#
$field_1="1,0,0";
#
############################################################
#For field_2 include Hydrogens Atoms in the Fitting [Y/N]
############################################################
#
$field_2="Y";
#
############################################################
#For field_3 use Mass- or Unit-Weighted Coordinates (M/U)
############################################################
#
$field_3="U";
#
#############################################
format SUPERPOSE =
@<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
$current_file
@<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
$tinker_par
@<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
$other_files
@<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
$tinker_par
```

```
@<<<<<<
$field_1
@<
$field_2
@<
$field_3
N
0.0
.
##########################################
#
##########################################
#Mopac output files are storaged in a vector
##########################################
#
@tmp=<*.out>;
#
########################################
#Extracts energy values from output files
########################################
#
foreach $arq (@tmp){
    open(MOPAC,"<$arq");
    while (<MOPAC>) {
        if (/(\s+FINAL HEAT OF FORMATION =)\s+([0-9+-.]+)/) {
            open(TMP,">>energy.dat");
            print TMP $arq," ",$2,"\n";
            close(TMP);
        }
    }
    close(MOPAC);
}
open(TMP,"<energy.dat");
while(<TMP>){
    @fld=split(" ",$_,10);
    $energy{"$fld[0]"}=$fld[1];
}
close(TMP);
#
##############################
#The files are sorted by energy
##############################
#
@sorted_energy = sort by_energy keys(%energy);
#
##########################################
#The files are converted from Mopac to TINKER
##########################################
#
foreach (@sorted_energy) {
    if ($opt_t) {
        $babel_out = $_;
        $babel_out =~ s/.out/.xyz/;
        `$babel_dir/babel -imopout $_ -otinker $babel_out`;
        push(@tinker_file,$babel_out);
        $uniq{$babel_out} = 1;
    } elsif ($opt_e) {
        $limit = $opt_e + $energy{$sorted_energy[0]};
        $value = $energy{$_};
        print $energy{$sorted_energy[0]},"\n";
        if ($value <= $limit) {
            $babel_out = $_;
            $babel_out =~ s/.out/.xyz/;
            `$babel_dir/babel -imopout $_ -otinker $babel_out`;
            push(@tinker_file,$babel_out);
            $uniq{$babel_out} = 1;
```

```
        } else {
            $babel_out = $_;
            $babel_out =~ s/.out/.xyz/;
            push(@tinker_file,$babel_out);
            $uniq{$babel_out} = 0;
        }

    } else {
        print "Uso:","\n";
        print "uniqueify -opcao","\n";
        print "opcoes: ","\n";
        print "       -t ==> todos os arquivos sao comparados","\n";
        print "       -e num ==> compara todos os arquivos com energia de (num) acima
do minimo","\n";
        exit;
    }

}
#
########################################
#RMS values between files are calculated
########################################
#
while (@tinker_file) {
    $current_file=shift(@tinker_file);
    if ($uniq{$current_file}==1){
        foreach $other_files (@tinker_file) {
            if ($uniq{$other_files}==1) {
                open(SUPERPOSE,">superpose.in");
                write (SUPERPOSE);
                close(SUPERPOSE);
                `cat superpose.in | $tinker_dir/superpose > superpose.log`;
                open(RMS,"<superpose.log");
                while (<RMS>) {
                    if (/( IMPOSE  --  After Rotation)\s{19}([0-9.]+)/) {
                        if ($2 <= 0.4) {
                            $uniq{$other_files}=0;
                        }
                        open(TABLE,">>table_rms.dat");
                        print TABLE "The r.m.s. between ",$current_file," and ",$other
_files," is ",$2," \n";
                        close(TABLE);
                    }
                }
                close(RMS);
            }
        }
    }
}
#
####################################
#Only unique structures are extracted
####################################
#
open(UNIQ,">unicas.dat");
foreach $file (@sorted_energy) {
    $tmp = $file;
    $tmp =~ s/.out/.xyz/;
    if ($uniq{$tmp}==1) {
        print UNIQ $file," ",$energy{$file},"\n";
    }
}
close(UNIQ);
#
sub by_energy {
return $energy{$a} <=> $energy{$b};
```

```
}
#
####################
#Remove some rubbish
####################
#
unlink("superpose.in");
unlink("superpose.log");
unlink(<*.xyz>);

open MAIL,"|mail $email";
     print MAIL "Terminou a remocao de arquivos duplicados\n";
close MAIL;
```

## A.2 Programas para automatizar a realização de cálculos

### A.2.1 Programa para execução serial de vários cálculos para QSAR

Este *script* realiza otimização de geometria e cálculo de propriedades para séries de moléculas, seqüencialmente. A estrutura de diretório deve ser mantido: Um diretório para cada compostos da série (no caso os diretórios são `mol1 mol2 mol3`). Arquivos de entrada em coordenadas internas para cada conformação utilizada, com nome composto pelo nome do composto e conformação (no caso cada conformação do composto `mol1` deve ser nomeada `mol1_a.zmt` e `mol1_b.zmt`). Partindo destes arquivos o programa cria corretamente os arquivos com os nomes adequados. Ao fim da execução os diretórios de cada composto deverão conter os arquivos com as propriedades calculadas de acordo com o conteúdo do arquivo `${HD}/${PPT}.hed` e um arquivo com as densidades eletrônicas calculadas para cada orbital molecular entre HOMO-10 e LUMO. O arquivo com as propriedades terá o nome formado pelo conteúdo das variáveis `${NM}_${CF}_${OPT}_${PPT}.log` (no caso `mol1_a_ecp_321Gpp.log`), e as densidades no arquivo `${NM}_${CF}_${OPT}_${PPT}.den`. Os demais arquivos intermediários que não são apagados seguem a mesma lógica para composição dos nomes.

#### runall.csh

```
#$Id: runall_am1_prop,v 1.1 2000/07/11 13:09:55 eborges Exp eborges $
set OT = 'pwd'
set OPT = ecp
set PPT = 321Gpp
set HD = ${OT}/head
cd $OT/work
foreach NM (mol1 mol2 mol3)
  cd $NM
  foreach CF (a b)
    gzip -d *.gz>>&$OT/${OPT}${PPT}.err
    if (-f ${NM}_${CF}_${OPT}.log) then
      set FLAG = 'grep NORMALLY ${NM}_${CF}_${OPT}_${PPT}.log|cut -d\  -f6'
    else
      set FLAG = ERR01
    endif
    if (($FLAG == "NORMALLY")&&(-f ${NM}_${CF}_${OPT}_${PPT}.den)) then
      gzip ${NM}_${CF}_${OPT}_${PPT}.log >>& $OT/${OPT}${PPT}.err
      echo ${NM}_${CF} completo>>$OT/${OPT}${PPT}.log
```

```
    else if (-z ${NM}_${CF}_${OPT}.zmt) then
      rm -f ${NM}_${CF}_${OPT}.zmt >>& $OT/${OPT}${PPT}.err
      if (-f ${NM}_${CF}.zmt) then
      cat ${HD}/${OPT}.hed ${NM}_${CF}.zmt ${HD}/tail>${NM}_${CF}_${OPT}.inp
      gms ${NM}_${CF}_${OPT} 01 1>&${NM}_${CF}_${OPT}.log
      set FLAG3 = `grep "END OF GEOMETRY SEARCH"|cut -d\  -f2|cut -d. -f7`
      if (FLAG3 == "END") then
        babel -igamout ${NM}_${CF}_${OPT}.log -ogzmat
${NM}_${CF}_${OPT}.tmp
        sed /Variables:/s///
${NM}_${CF}_${OPT}.tmp>${NM}_${CF}_${OPT}.zmt
        rm ${NM}_${CF}_${OPT}.tmp>>&$OT/${OPT}${PPT}.err
      endif
    else if (-f ${NM}${CF}${OPT}.zmt) then
      cat ${HD}/${PPT}.hed  ${NM}${CF}${OPT}.zmt ${HD}/tail>${NM}_${CF}_${OPT}_${PPT}.inp
      rm -f ${NM}_${CF}_${OPT}_${PPT}.log ${NM}_${CF}_${OPT}_${PPT}.dat >>& $OT/${OPT}${PPT}.err
      gms ${NM}_${CF}_${OPT}_${PPT} 01 1>& ${NM}_${CF}_${OPT}_${PPT}.log
      set FLAG2 = `grep NORMALLY ${NM}_${CF}_${OPT}_${PPT}.log|cut -d\
-f6`
      if ($FLAG2 == "NORMALLY") then
        set HOMO = `grep ALPHA ${NM}_${CF}_${OPT}_${PPT}.log|awk
      '{print $7}'`
        @HOMO ++
        cat ${HD}/${PPT}den.hed|sed -e "/XX/s//`echo $HOMO`/">tmp.hed
        set LIN = `grep -n END ${NM}_${CF}_${OPT}_${PPT}.dat|awk 'BEGIN {FS=":"} {print $1}'`
        tail -n +$LIN[1] ${NM}_${CF}_${OPT}_${PPT}.dat|tail -r>vec
        set LIN = `grep -n END vec|awk 'BEGIN {FS=":"} {print $1}'`
        tail -n +$LIN[1] vec|tail -r>tmp.vec
        cat tmp.hed ${NM}_${CF}_${OPT}.zmt ${HD}/tail tmp.vec>de.inp
        rm -f vec de.dat de.log>>&  $OT/${OPT}${PPT}.err
        gms de 01 1>&de.log
        set AT = `grep ATOMS de.log|awk '{print $6}'`
        set AT2 = `expr $AT + 3`
        grep -B 1 -A $AT2 'ELECTRON DENSITY' de.log>${NM}_${CF}_${OPT}_${PPT}.den
        rm -f de.log de.dat >>& $OT/${OPT}${PPT}.err
        foreach OB (1 2 3 4 5 6 7 8 9 10 11)
          set HOMO = `expr $HOMO - $OB`
          cat ${HD}/${PPT}den.hed|sed -e "/XX/s//`echo $HOMO`/">tmp.hed
          cat tmp.hed  ${NM}_${CF}_${OPT}.zmt ${HD}/tail tmp.vec>de.inp
          gms de 01 1>&de.log
          grep -B 1 -A $AT2 'ELECTRON DENSITY' de.log>>${NM}_${CF}_${OPT}_${PPT}.den
          rm -f de.log de.dat de.inp >>& $OT/${OPT}${PPT}.err
        end
      endif
    endif
  end
  cd ..
  gzip --best -r $NM
end
#$Log: runall_am1_prop,v $
```

## A.2.2 Programa para incluir efeitos de solvatação em geometria otimizada

Este *script* utiliza a geometria gerada em coordenadas internas do definida no bloco `$data$` para o GAMESS, otimizada previamente com nível de cálculo adequado à inclusão do efeito de solvatação proposto. As instruções para identificação de nomes de arquivos e diretórios são dadas em comentários dentro do programa, que deve ser modificado para adequar-se a outros exemplos práticos.

**runsolv.csh**

```
#!/bin/csh
#
#  C-shell script to execute sequential GAMESS jobs.
#  Invoke this by typing 'runsolv.csh'.
#  Needs headfiles, and zmatrix coordinates for GAMESS run above the
# following paths. The GAMESS setup script name is 'gms', and must
# exist in any searchable path. Follows some settings.

#  Here is the root directory:
set OT = 'pwd'
#  Logfiles and error files are generated here.

#  Headfiles must exist above the next path:
set HD = $OT/head
# The name of any headfile must be an expression formed with two
# identifier variables followed by the '.hed' suffix.
# The identifiers currently used are {$AQV} for the runlevel and {$SO}
# for the solvent effect were added inn. So their names must be
# must be 6311Gaq.hed 321Gac.hed 6311Gch.hed with current settings.

#  Work directory must exist above the next path:
set LO = $OT/work
#  Optimized geometrical coordinates of the molecules must be provided
# on files above this path. Their names must be an expression formed
# with a runlevel identifier variable followed by the suffix '.zmt'.
# All GAMESS output files must be stored above this directoy until the
# end of all the calculations of the entire series of compounds.
# Corrupted or abnormally output files must be mannually removed.

#  This is an identifier the this run
set AQV = 321G

#  Here goes the script name for log files
set SC = runall_solv

cd $LO
foreach NU (mol1 mol2 mol3)
 foreach SO (ac aq ch)
  if (-f $NU$AQV${SO}.log.gz) then
   echo "arquivo $NU$AQV${SO}.log.gz JÁ EXISTE ">$OT/$SC.err
  else
   cat $HD/$AQV$SO.hed ${NU}.zmt > tmp.inp
   gms tmp 01 1 > & $NU$AQV${SO}.log
   echo  $NU$AQV${SO}: 'grep NORMALLY $NU$AQV${SO}.log | \
   cut -d\ -f6 -f8 -f9 -f11'>>$OT/$SC.log
   gzip --best $NU$AQV${SO}.dat>>&$OT/$SC.err
  endif
 end
end
```

## A.3 Programas para buscar dados nos arquivos de saída

### A.3.1 Tabulação e conferência dos conjuntos de dados para criação do modelo QSPR de $\log P_{o/w}$

. Este programa localiza descritores e dados necessários para calcular os valores de partição em arquivos de dados de saída do GAMESS calculados utilizando nos cabeçalhos as diretivas listadas na página 56 de Secção 2.3.1 para calcular propriedades moleculares com efeitos de solvatação com cavidade do tipo denominado 'S1'. Utiliza o programa funções escritas para o comando awk, denominadas ak80, ak25 e charmin.

**Programa utilizado para tabular e calcular valores de partição**

**particaoS1.csh**

```
#!/bin/csh
set AWK = ~/prog/awk_lib
set ATOM = `fgrep "TOTAL NUMBER OF ATOMS" $1|awk '{print $6}'`
set NATOM = `expr ${ATOM} + 4`
fgrep " TOTAL FREE ENERGY IN SOLVENT =" $1 $2| awk -f ${AWK}/ak80>tmp1
fgrep " TOTAL FREE ENERGY IN SOLVENT =" $1 $3| awk -f ${AWK}/ak80>>tmp1
fgrep " TOTAL FREE ENERGY IN SOLVENT =" $2 $3| awk -f ${AWK}/ak80>>tmp1
fgrep "DIPOLE MOMENT" $1|awk '{print $5}'>>tmp1
fgrep "DIPOLE MOMENT" $2|awk '{print $5}'>>tmp1
fgrep "DIPOLE MOMENT" $3|awk '{print $5}'>>tmp1
fgrep 'TOTAL NUMBER OF TESSERAE' $1 $2 $3| awk ' NF>7 {print $9}'>>tmp1
fgrep -A${NATOM} 'NET CHARGES' $1|awk -f ${AWK}/charmin>>tmp1
fgrep -A${NATOM} 'NET CHARGES' $2|awk -f ${AWK}/charmin>>tmp1
fgrep -A${NATOM} 'NET CHARGES' $3|awk -f ${AWK}/charmin>>tmp1
awk -f ${AWK}/ak25 tmp1>>$4
exit
```

**Funções para awk** Estas funções, denominadas `ak80`, `charmin` e `ak25` devem ser criadas em arquivos com este nome no diretório `~/prog/awk_lib`

```
#ak80
{n[NR]=$8}
END {logpaq_ch=log(exp(1)^(.16244367*(n[1]-n[2])))/log(10)
     print logpaq_ch}

#ak25
{campo=campo $1" "}
END {print campo}

#charmin
{if ( (($1=="C")||($1=="N")||($1=="S")||($1=="CL")||($1=="BR")    \
                ||($1=="F")||($1=="O")||($1=="SI")||($1=="P" ) ) \
   && ($2 < c1) ) c1 = $2 }
END {print c1}
```

## A.3.2 Programa utilizado para tabular os descritores mecânico-quânticos

A estrutura do programa divide a tarefa em três *shell scripts*. O primeiro denominado `datall.csh` gerencia a busca na estrutura de diretórios. O segundo localiza e imprime o

bloco de dados para cada arquivo de saída, denominado `buscadados.csh`. O terceiro é um conjunto de pequenos *scripts* para o `awk` que fazem a seleção e formatação dos dados a partir das linhas do arquivo de saída do GAMESS

**datall.csh** Este programa gerencia o processo de busca dos dados calculados como GAMESS através da árvore de diretórios criada para organizar os cálculos necessários para o estudo QSAR de uma série de compostos. Este programa utiliza a mesma estrutura de diretórios que o Programa A.2.1, denominado `runall.csh`, listado na Secção A.2

```
#!/bin/csh

#  Realiza o gerencia a busca de dados com o script buscadados
#  Algumas informações sobre caminhos e arquivos de destino
#  devem ser inicializadas nas linha a seguir.
#  A sintaxe é: datall
#  Não aceita argumentos de entrada.

#  Define o diretório onde ficam os diretórios referentes a cada uma
#  das moléculas, e dentro dos diretório das moléculas os arquivos
#  de saída do GAMESS
set RAIZ = `pwd`

#  Define o diretório onde ficam os diretórios correspondentes a cada
#  molécula na árvore. O termo abinitio pode ser substituído.
set LO = $RAIZ/abinito

#  Os nomes dos arquivos de saída do GAMESS são formados
#  obrigatoriamente da seguinte junção: {$NUM}{$AQV}.log
#  NUM devem ser os nomes dos diretórios das moléculas onde ficam
#  os arquivo de saída do GAMESS correspondente àquela molécula.
#  Os nomes dos diretório mol1 mol2 e mol3 devem ser substituídos
#  no comando logo abaixo: foreach (mol1 mol2 mol3)

#  AQV: Varível que define o tipo de cálculo realizado
set AQV = 6311G

#  Define qual o script de busca de dados, no caso buscadados.csh
set BUSCA = ~/prog/buscadados.csh

#  Define para onde o arquivo com os dados tabulados será direcionado
set SA = $RAIZ/dados

cd {$LO}
```

```
foreach NUM (mol1 mol2 mol3)
  if (-f {$NUM}/{$NUM}{$AQV}.log.gz) then
    echo +++Descomprimindo o arquivo $NUM/$NUM$AQV.log.gz+++
    gzip -d {$NUM}/{$NUM}{$AQV}.log.gz>& $RAIZ/run{$AQV}.run
  endif
  if (-f {$NUM}/{$NUM}{$AQV}.log) then
    cd $NUM
    ${BUSCA} ${NUM}${AQV}.log tmp3
    cat tmp3>>../tmp3
    rm tmp3
    cd ..
    echo "$BUSCA $NUM$AQV.log">& $RAIZ/run{$AQV}.log
    gzip --best {$NUM}/{$NUM}{$AQV}.log
    echo $NUM$AQV>>tmp4
  else
    echo "FALHA: $NUM$AQV.log">& $RAIZ/run{$AQV}.run
  endif
end
mv tmp3 {$SA}/dados{$AQV}.txt
mv tmp4 {$SA}/lista{$AQV}.txt
rm tmp2
```

### buscadados.csh

Este é o programa que lê os dados em cada arquivo de saída do **GAMESS** e chama as funções de formatação das linhas de dado correspondentes a cada uma das moléculas que serão utilizadas em QSAR.

```
#!/bin/csh
# A sintaxe é:
# buscadados arquivo1 arquivo2
# arquivo1 é a saída do gamess, arquivo2 é onde os dados serão escritos

# Define o caminho das funções de biblioteca
set AWK = ~/prog/awk_lib
# Define o número de átomos onde as propriedades são interessantes
set NA = 18

# O comando fgrep não dispõe da opçao -A na implementação do
# AIX, devendo ser modificado.
fgrep -A 14 "MULTIPOLE MOMENTS:" $1|awk -f ${AWK}/ak22>>tmp1

# Localizao os campos elétricos (se) calculados.
set na = `expr $NA \* 5 + 3`
fgrep -A $na "    ELECTRIC FIELD    " $1|awk -f ${AWK}/ak20>>tmp1
```

```
# Localiza as cargas calculadas com o método CHELPG
set na = `expr $NA + 3`
fgrep -A $na "NET CHARGES:" $1|awk -f ${AWK}/ak21>>tmp1

# Localiza as densidades eletrônicas calculadas.
# Normalmente o programa GAMESS calcula a densidade total,
# exceto se definido o orbital no arquivo de entrada.
set na = `expr $NA + 3`
fgrep -A $na "ELECTRON DENSITY" $1|awk -f ${AWK}/ak24>>tmp1

# Localiza as energias dos últimos 11 orbitais ocupados
set HOMO = `fgrep ALPHA $1|awk '{print $7}'`
set LIMI = `expr $HOMO - 11`
fgrep -B 1 'A          A          A          A          A' $1 \
      | awk -f $AWK/ak_enorb | sed -n "$LIMI,${HOMO}p" \
      | awk '{printf("%12.6f\n",$1)}'>>tmp1

# Organiza os dados em uma única linha para entrada na matriz dos descritores
awk -f ${AWK}/ak25 tmp1>>$2
#
rm tmp1
exit
```

### Funções para awk

Estas funções realizam a formatação dos dados localizados pelo *script* A.3.2. Sua implementação em arquivos separados separada permite maior mudularidade na execução das funções. Os arquivos devem ser criados no caminho determinado no *script*, atualmente determinado em um sistema UNIX para ~/prog/awk_lib.

```
# Funções em awk, que devem ser escritas em arquivos separados
# denominados  ak22, ak20, ak21 e ak_enorb
# localizados no diretório ~/prog/awk_lib.

#ak22
{if ($6 !="VALUE" && $6!="") print $5}

#ak20
{if ($1 !~"[A-Z]" && $2 !~"[A-Z]" && $4!=0) print $4}

#ak21
{if ($2 !~ "C" && $2!="") print $2}

#ak_enorb
{for (i=1;i<=NF;i=i+1) {if ($i!~/[A-Z]/&&$i!="--") print $i}}
```

# Referências Bibliográficas


[1] Drossman, D. A.; Thompson, W. G.; Talley, N. J.; Funch-Jensen, P.; Janssens, J.; Whitehead, W. E. *Gastroenterology International* **1990,** *3,* 159-172.

[2] Talley, N. J.; Collin-Jones, D.; Koch, K. L.; Koch, M.; Nyrén, O.; V.Stanghellini, *Gastroenterology International* **1990,** *4,* 145-160.

[3] Nyrén, O.; Lindberg, G.; Lindström, E. *Pharmacoeconomics* **1992,** *1,* 312.

[4] Coelho-Filho, J. M.; de O. Lima, J. W.; Furtado, G. B.; Castelo, A. *Revista da Associação Médica Brasileira* **2000,** *Janeiro-Março,* 30-39.

[5] Guariento, M. E. Departamento de Clínica Médica, FCM - UNICAMP, 2004.

[6] Quigley, A. Health Behavior News Service, 2002.

[7] Andersen, I. B.; Bonnevie, O.; Jorgensen, T.; Sorensen, T. I. *Ugeskr Laeger* **1999,** *161,* 1589-1594.

[8] Higham, J.; Kang, J. K.; Majeed, A. *Gut* **2002,** *50,* 460-464.

[9] Zelmanowicz, R. U. ABC da Saúde e Prevenção Ltda., 2004.

[10] Lehmann, F.; Hildebrand, P.; Beglinger, C. *Drugs* **2003,** *63,* 1785-1797.

[11] Lewis, G. IMS World Review, 2002.

[12] Sawaguchi, A.; McDonald, K. L.; Forte, J. G. *J. Histochemistry and Citochemistry* **2004,** *52,* 77-86.

[13] Asano, S.; Morii, M.; Takeguchi, N. *Biological Pharmaceutical Bulletin* **2004,** *27,* 1-12.

[14] Schneider, G. *Neural Networks* **2000,** *13,* 15 F. Hoffmann-La Roche Ltd., Pharmaceuticals Research, CH-4070 Basel, Switzerland.

[15] Gaudio, A. C. *Relações entre Estrutura Química e Atividade Biológica de Inibidores da Timidina-Cinase do Vírus Herpes Simplex,* Doutorado thesis, Universidade Estadual de Campinas, 1998.

[16] Kubinyi, H. Lecture, 2001.

[17] Klebe, G.; Mietzner, T. *J-CAMD* **1994,** *8,* 583-606.

[18] Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996,** *96,* 1027-1043.

[19] Baumann, K. *TrAC* **1999,** *18,* 36-46.

[20] Todeschini, R.; Lasagni, M.; Marnego, E. *J. Chemom.* **1994,** *8,* 263-273.

[21] Clerc, J.-T.; Terkovics, A. L. *Anal. Chim. Acta* **1990,** *235,* 93-102.

[22] Zupan, J.; Novic, M. *Anal. Chim. Acta* **1997,** *384,* 409-418.

[23] Zupan, J.; Novic, M. *Anal. Chim. Acta* **1999,** *388,* 243-250.

[24] Hong, S. Y.; Park, T. G.; Lee, K.-H. *Peptides* **2001,** *22,* 1669-1674.

[25] Niño, A.; Muñoz-Caro, C. *Biophys. Chem.* **2001,** *91,* 49-60.

[26] Moore, V. A.; Irwin, W. J.; Timmins, P.; Lambert, P. A.; Chong, S.; Dando, S. A.; Morrison, R. A. *Int. J. Pharmac.* **2000,** *210,* 29-44.

[27] Estrada, E. *Mut. Res.* **1998,** *420,* 67-75.

[28] Coluci, R. V. V. R.; Braga, R.; Galvão, D. *J. Mol. Struct. THEOCHEM* **2002,** *619,* 195-205.

[29] Tomoda, S.; Senju, T. *Tetrahedron* **1999,** *13,* 3871-3882.

[30] Tomoda, S.; Senj, T. *Tetrahedron* **1999,** *55,* 5303-5318.

[31] Chang, C. M. *J. Mol. Struct. THEOCHEM* **2003,** *622,* 249-255.

[32] Dinur, U. *J. Mol. Struct. THEOCHEM* **1994,** *303,* 227-237.

[33] Zhu, W.; Jiang, Y. *J. Mol. Struct. THEOCHEM* **2000,** *496,* 67-72.

[34] Makovskaya, V.; Dean, J.; Tomlinson, W.; Comber, M. *Anal. Chim. Acta* **1995,** *315,* 193-200.

[35] Yang, M.; Jiang, Y. *Chem. Phys.* **2001,** *274,* 121-130.

[36] Uhrig, U.; Höltje, H.-D.; Mannholdb, R.; Weber, H.; Lemoine, H. *J. Mol. Graph. Modell.* **2002,** *21,* 37-45.

[37] Foresman, J. B.; Frisch, Æ. *Exploring chemistry with electronic structure methods;* Gaussian: Carnegie Office Park, Building 6 – Pittsburgh, PA 15106 USA, 2nd ed.; 1996 ISBN 0-9636769-3-8.

[38] Taylor, R. *Acta Cryst.* **2002,** *D58,* 879-888.

[39] van der Spoel, D.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Hess, B.; Feenstra, K. A.; Lindahl, E.; van Drunen, R.; Berendsen, H. J. C. "Gromacs User Manual version 3.1", 2002.

[40] Mohamadi, F.; Richards, N.; Guida, W.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. *J. Comput. Chem.* **1990,** *11,* 440-467.

[41] Kaminski, G.; Jorgensen, W. L. *J. Phys. Chem.* **1996,** *100,* 18010-18013.

[42] Morris, G. M.; Goodsell, D. S.; Huey, R.; Hart, W. E.; Halliday, S.; Belew, R.; Olson, A. J. "Automated docking of flexible ligands to receptors", The Scripps Research Institute, `http://www.scripps.edu/pub/olson-web/doc/autodock`, 3.0.3 ed.; 1999.

[43] Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. *J. Med. Chem.* **1989,** *32,* 1083-1094.

[44] Lii, J.-H.; Allinger, N. L. *J. Comput. Chem.* **1998,** *19,* 1001-1016.

[45] Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980,** *102,* 939-947.

[46] Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. *J. Am. Chem. Soc.* **1982,** *104,* 2797-2803.

[47] Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980,** *72,* 650-654.

[48] Stevens, W. J.; Bash, H.; Krauss, M. *J. Chem. Phys.* **1984,** *81,* 6026-6033.

[49] Cundari, T. R.; Stevens, W. J. *J. Chem. Phys.* **1993,** *98,* 5555-5565.

[50] Wadt, W. R.; Hay, P. J. *J. Chem. Phys.* **1985,** *82,* 284-298.

[51] Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999,** *99,* 2161-2200.

[52] Huron, M.-J.; Claverie, P. *J. Phys. Chem* **1972,** *76,* 2123-2133.

[53] Tomasi, J.; Cammi, R. *J. Chem. Phys.* **1994,** *100,* 7495-7502.

[54] Huron, M.-J.; Calverie, P. *J. Phys. Chem.* **1974,** *78,* 1853-1861.

[55] Huron, M.-J.; Calverie, P. *J. Phys. Chem.* **1974,** *78,* 1862-1867.

[56] Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997,** *106,* 5151-5158.

[57] Chipman, D. M. *J. Chem. Phys.* **1996,** *104,* 3276-3289.

[58] Questel, J. Y. L.; Berthelot, M. *J. Chem. Soc. Perkin Trans. 2* **1997,** *12,* 2711-2717.

[59] Dewar, M. J. S.; Thiel, W. J. *J. Am. Chem. Soc.* **1977,** *99,* 4899-4907.

[60] Dewar, M. J. S.; Zoebish, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985,** *107,* 3902-3909.

[61] Saunders, M.; Houk, K. N.; Wu, Y.-D.; Still, C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc.* **1990,** *112,* 1419-1427.

[62] Howard, A. E.; Kollman, P. A. *J. Med. Chem.* **1988,** *31,* 1669-1675.

[63] Parish, C.; Lombardi, R.; Sinclair, K.; Smith, E.; Goldberg, A.; Rappleye, M.; Dure, M. *Journal of Molecular Graphics and Modelling* **2002,** *21,* 129-150.

[64] Wengner, J. C.; Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1982,** *22,* 29-34.

[65] Dyott, T. M. *Utilization of Stereochemistry and Other Aspects of Computer-Assisted Synthetic Design,* Thesis, Princeton University, 1973.

[66] Crippen, G. M.; Havel, T. F. *Acta Cryst.* **1978,** *A34,* 282-284.

[67] Hodsdon, M. E.; Ponder, J. W.; Cistola, D. P. *J. Mol. Biol.* **1996,** *364,* 585-602.

[68] Santagata, L. N.; Suvire, F. D.; Enriz, R. D. *J. Mol. Struct. THEOCHEM* **2001,** *571,* 91-98.

[69] Santagata, L.; Suvire, F.; Enriz, R. *J. Mol. Struct. THEOCHEM* **2001,** *536,* 173-188.

[70] Santagata, L.; Suvire, F.; Enriz, R. *J. Mol. Struct. THEOCHEM* **2000,** *507,* 89-95.

[71] Santagata, L. N.; Suvire, F. D.; Enriz, R. D.; Torday, L. L.; Csizmadia, I. G. *J. Mol. Struct. THEOCHEM* **1999,** *465,* 33-67.

[72] Cramer-III, R. D.; Patterson, D. E.; Bunce, J. D. *JACS* **1988,** *110,* 5959.

[73] van der Graaf, P. H.; Nilsson, J.; van Schaick, E. A.; Danhof, M. *J. Pharm. Sci.* **1999,** *88,* 306-312.

[74] Bakshi, B. R.; Chatterjee, R. *J. Alloys Comp.* **1998,** *279,* 39.

[75] Schleifer, K.-J. *J-CAMD* **2000,** *14,* 467-475.

[76] Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987,** *2,* 37-52.

[77] Strang, G. *Linear Algebra and its Applications;* Academic Press: San Diego, 3ª ed.; 1988.

[78] Wise, B. M.; Gallagher, N. B. *PLS Toolbox for use with Matlab;* Eigenvector Technologies: Manson, 1996.

[79] Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O. *Química Nova* **1999,** *22,* 724-731.

[80] Massart, D. L. *Chemometrics: a textbook;* volume 2 of *Data handling in science and technology* Elsevier: Amsterdam, 1990 488 p.

[81] Brereton, R. G., Ed.; *Multivariate pattern recognition in chemometrics, illustrated by case studies;* volume 9 of *Data handling in science and technology* Elsevier: Amsterdam, 1992 325 p.

[82] Geladi, P.; Kowalski, B. R. *Analytica Chimica Acta* **1986,** 1-17.

[83] Wise, B. M.; Gallagher, N. B. "PLS Toolbox Version 1.5", Eigenvector Technologies, P.O. Box 483, 196 Hyacinth Avenue, Manson, WA, 98831,.

[84] Eaton, J. W. "Octave 2.0.13: A High Level Language for Numerical Computations", 1998.

[85] Martens, H.; Naes, T. *Multivariate Calibration;* John Willey & Sons: Chichester, 1989 419 p.

[86] Geladi, P.; Kowalski, B. R. *Analytica Chimica Acta* **1986,** 19-32.

[87] Kubinyi, H. *QSAR* **1994,** *13,* 285-294.

[88] Kubinyi, H. *QSAR* **1994,** *13,* 393-401.

[89] Kellogg, G. E.; Abraham, D. J. *Eur. J. Med. Chem.* **2000,** *35,* 651-661.

[90] Leo, A. J. *Chem. Rev.* **1993,** *93,* 1281-1306.

[91] Eyck, M. H. V.; Ducarme, P.; Benhabiles, N.; Thomas, A.; Brasseur, R. *Eur. J. Anal. Chem.* **1999,** *27,* 6-14.

[92] Kellogg, G. E.; Abraham, D. J. *Eur. J. Anal. Chem.* **1999,** *27,* 19-23.

[93] Verhaar, H. J. M.; van Leeuwen, C. J.; Hermens, J. L. M. *Chemosphere* **1992,** *4,* 471-491.

[94] Henk, J. M.; Ramos, E. U.; Hermens, J. L. M. *J. Chemom.* **1996,** *10,* 149-162.

[95] Borges, E. G.; Takahata, Y. *Quimica Nova* **2002,** *25,* 1061-1066.

[96] Gramatica, P.; Navas, N.; Todeschini, R. *Trends Anal. Chem.* **1999,** *18,* 461-471.

[97] "MOPAC", `http://home.att.net/~mrmopac`, 1999.

[98] "TINKER", `http://dasher.wustl.edu/tinker`, 2000.

[99] de Oliveira, K. M. G. Tese de Doutoramento em andamento–Instituto de Química–UNICAMP.

[100] "GAMESS", http://www.msg.ameslab.gov/GAMESS/GAMESS.html, 1999.

[101] Pierotti, R. A. *Chem. Rev.* **1976,** *76,* 717-726.

[102] Langlet, J.; Claverie, P.; Caillet, J.; Pullman, A. *J. Phys. Chem.* **1988,** *92,* 1617 - 1631.

[103] Amovilli, C.; Mennucci, B. *J. Phys. Chem. B* **1997,** *B101,* 1051-1057.

[104] Vaes, W. H. J.; Ramos, E. U.; Verhaar, H. J. M.; Cramer, C. J.; Hermens, J. L. M. *Chem. Res. Toxicol.* **1998,** *11,* 847-854.

[105] Wang, H.; Ben-Naim, A. *J. Phys. Chem. B* **1997,** *101,* 1077 - 1086.

[106] Breneman, C. M.; Wiberg, K. *J. Comp. Chem.* **1990,** *11,* 361-373.

[107] *CRC Handbook of Chemistry and Phisics, 67th ed.;* CRC Press Inc.: Boca Raton, Florida, 1987.

[108] Lai, L. *J. Chem. Inf. Comput. Sci.* **1997,** *37,* 615-621.

[109] Borges, E. G. "PLST1.0", edilson@cenapad.unicamp.br, 2001.

[110] "OCTAVE", ftp://ftp.che.wisc.edu/pub/octave,.

[111] Ichikawa, H. "PSDD: Perceptron-type Neural Network Simulator", Quantum Chemistry Program Exchage - QCPE, Hoshi College of Pharmacy, 2-4-41 Ebara, Shinagawa, Tokyo 142 Japan, 1989.

[112] do Amaral, A.; Malvezzi, A.; Gonçalves, R. S. Aspectos Intermoleculares no Coeficiente de Partição em uma Série de Análogos da Procainamida. In *Livro de Resumos da 24 Reunião da SBQ*; Sociedade Brasileira de Química: Av. Prof. Lineu Prestes, 784 - Instituto de Química da USP, bloco 3 superior - Cidade Universitária, São Paulo - SP CP 26.037 CEP 05513-970, 2001 - Comunicação verbal com os autores.

[113] Huang, J.-Q.; Hunt, R. H. *Bailliéri's Clin. Gastroent.* **2001,** *15,* 355-370.

[114] Satoh, K.; Nagai, F.; Kano, I. *Biochem. Pharmac.* **2000,** *59,* 881-886.

[115] Fort, F. L.; Miyajima, H.; Ando, T.; Suzuki, T.; Yamamoto, M.; Yamashima, T.; Sato, S.; Kitazaki, T.; Mahony, M. C.; Hodgen, G. D. *Fund. Appl. Toxicol.* **1995,** *26,* 191-202.

[116] Hawkey, C. J. *Bailliéri's Clin. Gastroent.* **2000,** *14,* 173-192.

[117] Pope, A. J.; Boehm, M. K.; Leach, C.; Ife, R. J.; Keeling, D.; Parsonss, M. E. *Biochem. Pharmac.* **1995,** *50,* 1543-1549.

[118] LaMattina, J. L.; Lawrence, P. A.; Holt, W. F.; Yeh, L. A. *J. Med. Chem.* **1990,** *33,* 543-552.

[119] Munson, K. B.; Lambrecht, N.; Sachs, G. *Biochemistry* **2000,** *39,* 2997-3004.

[120] Lambrecht, N.; Munson, K. B.; Vagin, O.; Sachs, G. *J. Biol. Chem.* **2000,** *275,* 4041-4048.

[121] Rhee, K.-H.; Scarborough, G.; Henderson, R. *The EMBO Journal* **2002,** *21,* 3582-3589.

[122] Vagin, O.; Denevich, S.; Munson, K.; Sachs, G. *Biochemistry* **2002,** *41,* 12755-12762.

[123] Cenaand, C.; Lolli, L. M.; Lazzarato, L.; Guaita, E.; Morini, G.; Coruzzi, G.; McElroy, S. P.; Megson, I. L.; Fruttero, R.; Gasco, A. *J. Med. Chem.* **2003,** *46,* 747-754.

[124] Desiraju, G. R.; Gopalakrishnan, B.; Jetti, R. K. R.; Nagaraju, A.; Raveendra, D.; Sarma, J. A. R. P.; Sobhia, M. E.; Thilagavathi, R. *J. Med. Chem.* **2002,** *45,* 4847-4857.

[125] Selinsky, B. S.; Gupta, K.; Sharkey, C. T.; Loll, P. J. *Biochemistry* **2001,** *40,* 5172-5180.

[126] Augelli-Szafran, C. E.; Blankley, C. J.; Jaen, J. C.; Moreland, D. W.; Nelson, C. B.; Penvose-Yi, J. R.; Schwarz, R. D.; Thomas, A. J. *J. Med. Chem.* **1999,** *42,* 356-363.

[127] Böhme, T. M.; Keim, C.; Kreutzmann, K.; Linder, M.; Dingermann, T.; Dannhardt, G.; Mutschler, E.; Lambrecht, G. *J. Med. Chem.* **2003,** *43,* 856-867.

[128] Goodman, L. S.; Gilman, A. G. *Goodman & Gilman's the pharmacological basis of therapeutics;* McGraw-Hill: New York, 9th ed. ed.; 1996.

[129] Borges, E. G.; Takahata, Y. *J. Mol. Struc.:THEOCHEM* **2001,** *539,* 245-251.

[130] Cho, S. J.; Tropsha, A. *J. Med. Chem.* **1995,** *38,* 1060-1066.

[131] Jobson, J. D. *Applied Multivariate Data Analysis; Volume 1: Regression an Experimental Design;* Springer-Verlag: 175 Fifth Avenue, New York, NY 10010, USA, 1991.

[132] Borges, E. G.; Takahata, Y. *J. Mol. Struc.:THEOCHEM* **2002,** *580,* 263-270.

[133] Leach, C. A.; Brown, T. H.; Ife, R. J.; Keeling, D. J.; Parsons, M. E.; Theobald, C. J.; Wiggall, K. J. *J. Med. Chem.* **1995,** *38,* 2748-2762.

[134] Ife, R. J.; Brown, T. H.; Blurton, P.; Keeling, D. J.; Leach, C. A.; Meeson, M. L.; Parsons, M. E.; Theobald, C. J. *J. Med. Chem.* **1995,** *38,* 2763-2773.

[135] Ife, R. J.; Brown, T. H.; Keeling, D. J.; Leach, C. A.; Meeson, M. L.; Parsons, M. E.; Reavill, D. R.; Theobald, C. J.; Wiggall, K. J. *J. Med. Chem* **1992**, *35*, 3413-3422.

[136] Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423-1433.

[137] de L. Romero, M.; Mendez, F. *J. Phys. Chem. A* **2003**, *107*, 4526-4530.

[138] Terauchi, H.; Tanitame, A.; Nakamura, K.; Seto, K.; Nishimura, Y. *Chem. Pharm. Bull.* **1997**, *45*, 1027-1038.

[139] Terauchi, H.; Tanitame, A.; Tada, K.; Nakamura, K.; Seto, Y.; Nishikawa, Y. *J. Med. Chem.* **1997**, *40*, 313-321.

[140] Terauchi, H.; Tanitame, A.; Tada, K.; Nakamura, K.; Seto, Y.; Nishikawa, Y. *Chem Pharm. Bull.* **1997**, *45*, 1177-1182.

[141] Bensacon, M.; Simom, A.; Sachs, G.; Shin, J. M. *J. Biol. Chem.* **1997**, *272*, 22438-22446.

[142] Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899-926.

[143] Gung, B. W. *Tetrahedron* **1996**, *52*, 5263-5301.

[144] Wipf, P.; Jung, J.-K. *Chem. Rev.* **1999**, *99*, 1460-1480.

[145] Baldridge, K.; Pederson, L. *Pi Mu Epsilon Journal* **1993**, *9*, 513-521.

[146] Asano, S.; Kawada, K.; Kimura, T.; Grishini, A. V.; Caplani, M. J.; Takeguchi, N. *J. Biol. Chem.* **2000**, *275*, 83248330.

[147] Asano, S.; Matsuda, S.; Tega, Y.; Shimizu, K.; Sakamoto, S.; Takeguchi, N. *J. Biol. Chem.* **1997**, *272*, 17668-17674.

[148] Lambrecht, N.; Corbett, Z.; Bayle, D.; Karlish, S. J. D.; Sachs, G. *J. Biol. Chem.* **1998**, *273*, 13719-13728.

[149] Asano, S.; Kimura, T.; Uenoi, S.; Kawamura, M.; Takeguchi, N. *J. Biol. Chem.* **1999**, *274*, 22257-22265.

[150] Munson, K. B.; Lambrecht, N.; Shin, J. M.; Sachs, G. *J. Exp. Biol.* **2000**, *203*, 161-170.

[151] Farley, R. A.; Schreiber, S.; Wang, S.-G.; Scheiner-Bobis, G. *J. Biol. Chem.* **2001**, *276*, 2608-2515.

[152] Inc., S. "MacroModel", `http://www.schrodinger.com/Products/macromodel.html`, 2004.

[153] Guex, N.; Schwede, T.; Peitsch, M. C.; Diemand, A. "Swiss PDBViewer 3.7 (beta)", `http://www.expasy.ch/spdbv/mainpage.html`, 2001 Glaxo Welcome Experimental Research.

[154] Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. *Nucleic Acids Res.* **2003,** *31,* 3784-3788.

[155] Wang, S.-H. "PMOL2Q: A specific file format converter for protein pdb file to pdbq and pdbqs data format according AutoDock", `gentamicin@pchome.com.tw`, v2.1.1.

[156] Kühlbrandt, W.; Zeelen, J.; Dietrich, J. *Scince* **2002,** *297,* 1692-1696.

[157] Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. *Nature* **2000,** *405,* 647-655.

[158] "OpenBabel-1.100.1", `http://openbabel.sourceforge.net/`, 2003.

[159] Schaftenaar, G. "Molden—Programa para pré e pós processamento de estruturas moleculares e eletrônicas", CAOS/CAMM Center, Holanda, `http://www.caos.kun.nl/schaft/molden`.