UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTOS DE QUÍMICA, FÍSICA E ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

Simulação computacional de processos de formação de pares de bases biológicas livres considerando critérios probabilísticos, geométricos e energéticos

por

Sani de Carvalho Rutz da Silva

Tese submetida como requisito parcial
para a obtenção do grau de

Doutor em Ciência dos Materiais

Prof. Dr. Dimitrios Samios
Orientador

Prof. Dr. Álvaro de Bortoli
Co-orientador

Porto Alegre, Outubro de 2003.

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

de Carvalho Rutz da Silva, Sani

Simulação computacional de processos de formação de pares de bases biológicas livres considerando critérios probabilísticos, geométricos e energéticos */ Sani de Carvalho Rutz da Silva.—Porto Alegre: PGCIMAT da UFRGS, 2003.*

135 p.: il.

Tese (Doutorado) —Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Ciência dos Materiais, Porto Alegre, 2003.
Orientador: Samios, Dimitrios; Co-orientador: de Bortoli, Álvaro

Tese: Materiais Poliméricos
método de Monte Carlo, formação de pares de bases, critérios geométricos e energéticos

Simulação computacional de processos de formação de pares de bases biológicas livres considerando critérios probabilísticos, geométricos e energéticos

por

Sani de Carvalho Rutz da Silva

Tese submetida ao Programa de Pós-Graduação em Ciência dos Materiais da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de

Doutor em Ciência dos Materiais

Linha de Pesquisa: Materiais Poliméricos

Orientador: Prof. Dr. Dimitrios Samios

Co-orientador: Prof. Dr. Álvaro de Bortoli

Banca examinadora:

Prof. Dr. Pedro Geraldo Pascutti
Universidade Federal do Rio de Janeiro

Profa. Dra. Rita M. Cunha de Almeida
Instituto de Física/UFRGS

PhD. Paulo Ricardo Zingano
PPGMap/IM/UFRGS

Prof. Dr.Paolo R. Livotto
PPGQ/IQ/UFRGS

Tese apresentada e aprovada em
3 de outubro de 2003.

Profa. Dra. Raquel Santos Mauler
Coordenador

AGRADECIMENTO

Desejo expressar minha sincera e profunda gratidão para as pessoas que contribuiram para a realização deste trabalho: Professores Dimitrios Samios e Álvaro Luiz de Bortoli por suas valiosas orientações, pela dedicação e incentivos permanentes, tanto no âmbito profissional como pessoal e pelo apoio oferecido e convívio gratificante. Ao grande amigo e professor Paulo Netz que dispôs de seus valioso tempo para auxiliar na realização de meu trabalho e também ao inesquecível amigo Dagoberto Rizzoto Justo, que mesmo distante dedicou auxílio em toda parte computacional deste trabalho.

Aos queridos professores do Instituto de Matemática Paulo Ricardo Zingano, Maria Cristina Varriale, Maria Paula, Mark Thompson, Vilmar Trevisan, Waldir Roque, Jacques Aveline, José Afonso, Julio Clayssen, especialmente ao professor Rudnei Dias da Cunha que sempre mostrou carinho e atenção em todos os momentos de minha vida, e outros que porventura tenha esquecido apenas de citar o nome aqui.

Aos amigos do SRC pela grande amizade, especialmente pelo carinhoso apelido "cachinhos dourados", Luiz Fabricio, Claudete, Celso, Greice, Edsandro, Marcos, Priscila, João, muito obrigado....

Aos também amigos da Secretaria do Instituto de Matemática, Vocês são a minha família, Adriane, Ana, Marta, Augusto, Giovani, Leonardo, Antunes, que tenham todo o sucesso do mundo.

Aos colegas do LICC pelo companheirismo e amizade, especialmente a minha amiga Heloísa e aos grandes e inesquecíveis amigos do Instituto de QUÍMICA, principalmente Giovana, que sempre me trataram bem, obrigado pelo carinho....

Ao CNPq que possibilitou a realização deste trabalho.

Conteúdo

LISTA DE FIGURAS VIII
LISTA DE TABELAS XI
LISTA DE SíMBOLOS XII
RESUMO XIII
ABSTRACT XIV

1 INTRODUÇÃO 1
1.1 *Objetivos gerais e específicos* 15

2 UMA BREVE INTRODUÇÃO À BIOLOGIA MOLECULAR . 17
2.1 Estrutura do DNA e do RNA 19
2.2 Tipos de DNA e suas propriedades físicas e químicas 28
2.3 Replicação do DNA 29
2.4 Código Genético 32
2.5 Mutações 33
2.6 Tipos de pareamento de bases em DNA 36

3 IMPLEMENTAÇÃO DO MODELO: O ALGORITMO 42
3.1 Fluxograma do algoritmo 42
3.2 Definição e evolução do sistema 42
3.3 O método de Monte Carlo 45
3.4 Critério geométrico 48
3.4.1 Representação geométrica das bases (A,T,G e C) - SRB 48
3.4.2 Representação geométrica das bases (A,T,G e C) - SRU 52
3.4.3 Condições de contorno periódicas 61
3.5 Critério energético 64
3.5.1 Otimização da geometria molecular 65

3.5.2 Modelagem molecular usando o software CAChe 67

3.5.3 Procedimentos para Modelagem usando CAChe 69

3.6 Verificação das ligações de hidrogênio considerando o critério geométrico 73

3.7 Verificação das ligações de hidrogênio considerando o critério energético 74

4 RESULTADOS E DISCUSSÃO **79**

4.1 Parâmetros de simulação: domínio, iniciadores e varmax . . . 80

4.2 Simulações com o critério geométrico 83

4.2.1 Ligações simples e duplas entre as bases Adenina e Timina 85

4.2.2 Ligações simples e duplas entre as bases Adenina e Adenina 90

4.2.3 Ligações simples e duplas entre as bases Timina e Timina 92

4.2.4 Ligações simples e duplas entre as bases Guanina e Citosina 93

4.2.5 Ligações simples e duplas entre as bases Guanina e Guanina 95

4.2.6 Ligações simples e duplas entre as bases Citosina e Citosina 96

4.3 Simulações com o critério energético 96

4.3.1 Ligações simples e duplas entre as bases Adenina e Timina 97

4.3.2 Ligações simples e duplas entre as bases Adenina e Adenina 103

4.3.3 Ligações simples e duplas entre as bases Timina e Timina 106

4.3.4 Ligações simples e duplas entre as bases Guanina e Citosina 107

4.3.5 Ligações simples e duplas entre as bases Guanina e Guanina 111

4.3.6 Ligações simples e duplas entre as bases Citosina e Citosina 115

5 CONCLUSÕES E PERSPECTIVAS **118**

5.1 Conclusões dos resultados obtidos 118

5.2 Contribuições da tese 121

5.3 Sugestões para trabalhos futuros 122

REFERÊNCIAS 125

APÊNDICE A ENERGIA DE ESTABILIZAÇÃO DE LIGAÇÕES DE HIDROGÊNIO DOS PARES DE BASES . . 136

Lista de Figuras

Figura 2.1 Estrutura do Nucleotídeo 19

Figura 2.2 Estrutura da Desoxirribose 20

Figura 2.3 Par de bases: Adenina e Timina 21

Figura 2.4 Par de bases: Guanina e Citosina 21

Figura 2.5 Estrutura da dupla-hélice [39] 23

Figura 2.6 Ligações fosfodiéster 26

Figura 2.7 Estrutura do RNA 27

Figura 2.8 Duplicação do DNA 30

Figura 2.9 Pares de bases ATWC e ATrWC 37

Figura 2.10 Pares de bases ATH e ATrH 38

Figura 3.1 Fluxograma do algoritmo 43

Figura 3.2 Pirimidinas: Citosina(C) e Timina(T); Purinas: Adenina(A) e Guanina (G) 50

Figura 3.3 Representação das coordenadas dos átomos da base i 55

Figura 3.4 Movimentação das bases 57

Figura 3.5 Movimentação das bases fora do sistema de referência 58

Figura 3.6 Distribuição de Boltzmann [63] 76

Figura 3.7 Processo de decisão 77

Figura 4.1 Análise da conversão na formação dos pares de bases: experimento A1 83

Figura 4.2 Ligações de hidrogênio totais entre A e T com iniciadores: B1, B2 e B3. 84

Figura 4.3 Ligações de hidrogênio totais entre A e T: experimentos A2 e B2. 85

Figura 4.4 Ligações simples entre A e T: experimento C2 87

Figura 4.5 Ligações duplas entre A e T: experimentos A1 e B3 88

Figura 4.6 Ligações duplas entre A e T: experimentos C2 e C3 88

Figura 4.7 Ligações simples e duplas entre A e T: experimento C2 89

Figura 4.8 Ligações simples entre A e A: experimento C2 90

Figura 4.9 Ligações duplas entre A e A: experimento C2 91

Figura 4.10 Ligações duplas entre T e T: experimentos B3 e C1 92

Figura 4.11 Ligações de hidrogênio simples entre G e C: experimento C2 . . 93

Figura 4.12 Ligações de hidrogênio múltiplas entre G e C: experimento C2 . 94

Figura 4.13 Ligações de hidrogênio simples e duplas entre G e C: experimento C2 95

Figura 4.14 Ligações de hidrogênio duplas entre G e G: experimento C2 . . 96

Figura 4.15 Ligações de hidrogênio simples e duplas entre A e T com critérios geométrico e energético: experimento C2. 99

Figura 4.16 Ligações de hidrogênio simples entre A e T com critérios geométrico e energético: experimento C2,(g) geométrico-(e) energético 101

Figura 4.17 Ligações de hidrogênio duplas entre A e T com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 102

Figura 4.18 Ligações de hidrogênio simples entre A e A com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 104

Figura 4.19 Ligações de hidrogênio duplas entre A e A com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 105

Figura 4.20 Ligações de hidrogênio simples e duplas entre T e T com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 107

Figura 4.21 Ligações de hidrogênio simples e duplas entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 109

Figura 4.22 Ligações de hidrogênio simples entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 110

Figura 4.23 Ligações de hidrogênio duplas entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético . . . 111

Figura 4.24 Ligações de hidrogênio simples e duplas entre G e G com os critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 112

Figura 4.25 Ligações de hidrogênio simples entre G e G com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético 114

Figura 4.26 Ligações de hidrogênio duplas entre G e G com critério energético: experimento C2. 114

Figura 4.27 Ligações de hidrogênio simples e duplas entre C e C com critérios geométrico e energético: experimento C2. 116

Figura 4.28 Ligações de hidrogênio duplas entre C e C com critérios geométrico e energético: experimento C2. 116

Lista de Tabelas

Tabela 2.1 Características geométricas de diferentes pares entre T e T . . . 38

Tabela 2.2 Características geométricas de diferentes pares entre G e C . . . 39

Tabela 2.3 Características geométricas de diferentes pares entre C e C . . . 39

Tabela 2.4 Características geométricas de diferentes pares entre A e T . . . 39

Tabela 2.5 Características geométricas de diferentes pares entre G e G . . . 40

Tabela 2.6 Características geométricas de diferentes pares entre A e A . . . 40

Tabela 3.1 Coordenadas atômicas das bases Adenina e Timina no plano cartesiano 51

Tabela 3.2 Coordenadas atômicas das bases Guanina e Citosina no plano cartesiano 52

Tabela 4.1 Planilha de experimentos 79

Tabela 4.2 Valores dos fatores de Boltzmann entre A e T 98

Tabela 4.3 Valores dos fatores de Boltzmann entre A e A 104

Tabela 4.4 Valores dos fatores de Boltzmann entre T e T 106

Tabela 4.5 Valores dos fatores de Boltzmann entre G e C 108

Tabela 4.6 Valores dos fatores de Boltzmann entre G e G 112

Tabela 4.7 Valores dos fatores de Boltzmann entre C e C 115

Tabela A.1 Energia de estabilização (em Kcal/mol) dos pares de bases (ligações múltiplas) obtidos pelo método semi-empírico AM1. 136

Tabela A.2 Energia de estabilização (em Kcal/mol) dos pares de bases (ligações simples) obtidos pelo método semi-empírico AM1. . 137

LISTA DE SÍMBOLOS

A	Adenina
$AM1$	Austin Model 1
$ATWC$	Ligação entre Adenina e Timina de Watson e Crick
$ATrWC$	Ligação entre Adenina e Timina reversa de Watson e Crick
ATH	Ligação entre Adenina e Timina de Hoogsteen
$ATrH$	Ligação entre Adenina e Timina reversa de Hoogsteen
$AA(I, II, III, IV)$	Ligação entre Adenina e Adenina
C	Citosina
CC	Ligação entre Citosina e Citosina
$d\theta$	Variação no ângulo teta
$DistP$	Distância entre os sítios doador (d) e aceptor(a)
DNA	Ácido desoxirribonucleico
E	Matriz de coordenadas homogêneas do átomo
$(\exp(-\triangle E/K_B T))$	fator de Boltzmann
G	Guanina
$GCWC$	Ligação entre Guanina e Citosina de Watson e Crick
$GCrWC$	Ligação entre Guanina e Citosina reverso de Watson e Crick
$GC(II)$	Ligação entre Guanina e Citosina (II)
$GG(I, II, III, IV)$	Ligação entre Guanina e Guanina
$mRNA$	RNA mensageiro
PDB	Banco de dados de Proteína
RNA	Ácido ribonucleico
rnd	Números pseudoaleatórios entre [0,1]
R_α	Matriz de rotação do ângulo alpha
R_θ	Matriz de rotação do ângulo teta
SRB	Sistema de referência da Base
SRO	Sistema de referência do Objeto
SRU	Sistema de referência do Universo (global)
T	Timina
$TT(I, II, III, IV)$	Timina e Timina
T_r	Matriz de translação

$1pa$	1 ponte de hidrogênio "a"
$1pb$	1 ponte de hidrogênio "b"
$1pc$	1 ponte de hidrogênio "c"
$1pd$	1 ponte de hidrogênio "d"
$1pe$	1 ponte de hidrogênio "e"
$1pf$	1 ponte de hidrogênio "f"
$1pg$	1 ponte de hidrogênio "g"

RESUMO

Neste trabalho, analisam-se os processos de formação de ligações de hidrogênio entre as bases Adenina, Timina, Guanina e Citosina usando o método Monte Carlo probabilístico. A possibilidade de formação de pares é inicialmente verificada considerando critério geométrico (distância e orientação das moléculas) seguida pela análise da probabilidade energética, que é proporcional ao fator de Boltzmann. Os resultados mostram que a probabilidade de ocorrência, para alguns modelos, não segue a estrutura mais provável segundo o fator de Boltzmann. Isto sugere que existe uma forte influência geométrica na formação dos pares (ligações simples e múltiplas). Tal análise fornece a base para a construção de modelos mais complexos bem como para o entendimento de alguns mecanismos que ocorrem em processos relacionados à mutações, visando compreender este tipo de fenômeno biológico.

ABSTRACT

The aim of this work is the formation of hydrogen bonds (H-bonds) between the bases Adenine, Thymine, Guanine and Cytosine for base pairing, using the probabilistic Monte Carlo method. The pair formation probability is verified firstly by geometrical criteria (molecular distance and orientation) followed by an energetic criteria, which is proportional to the Boltzmann factor.

Numerical results show that the pairs occurrence probability, of same models, does not follow only the Boltzmann factor probability. Such indicates that there is a strong geometric influence at pairs formation (simple and multiple connections). This analysis gives the base for the construction of more complex models as well as to understand some mechanisms which occur at processes related to mutation, in order to understand this kind of biological phenomena.

1 INTRODUÇÃO

O conhecimento científico sobre os organismos vivos tem sido ampliado com o estudo da natureza molecular dos fenômenos biológicos [55]. Hoje em dia, não basta apenas o conhecimento do organismo do ponto de vista histológico, morfológico e fisiológico, mas se faz cada vez mais necessário o conhecimento dos detalhes das estruturas moleculares envolvidas nas funções biológicas. Os novos desenvolvimentos em Biologia molecular, a sofisticação de métodos físicos e os avanços atuais em técnicas de microscopia, cristalografia de raio X e outras espectroscopias têm permitido a determinação de estruturas moleculares complexas, com um grande impacto no estudo das funções biológicas [10].

A aplicação dos métodos computacionais cada vez mais sofisticados tem permitido um avanço sem precedentes em várias áreas da Ciência [80], em particular as perspectivas de aplicações computacionais em Ciências Biológicas tem tido uma grande repercussão e provavelmente terá um desenvolvimento ainda maior num futuro próximo. Um dos campos que tem sido mais explorado é a modelagem computacional de estruturas de interesse biológico [42]. A aplicação destas técnicas abre a perspectiva de um grande avanço na compreensão de processos biológicos em nível atômico-molecular e na proposta de novas estruturas moleculares com elevada eficiência biológica [10].

Surgem, desta forma, duas áreas de pesquisa que estão caminhando juntas: a Bioinformática, que pode ser descrita como a aquisição, análise e estoque de informação biológica, especificamente de ácidos nucleicos e proteínas; e a Biologia Molecular Computacional, que estuda o desenvolvimento de algoritmos e programas computacionais para resolver problemas nesta área. Ambas tem crescido enormemente na última década, motivados pelos projetos Genomas em curso.

Um dos objetivos fundamentais destas áreas é converter a informação contida na seqüência de nucleotídeos ou de proteínas (linguagem biológica) em co-

nhecimento bioquímico e biofísico (funções estruturais, funcionais e evolutivas). Na realidade, o que se procura é decodificar uma linguagem desconhecida, extraindo o seu significado biológico [32]. Esta linguagem pode ser decomposta em frases (proteínas), palavras (motivos ou padrões de estrutura) e letras (nucleotídeos e /ou aminoácidos). Assim como na linguagem natural, a substituição de uma única letra em uma palavra pode alterar o seu significado (como por exemplo casa-capa); a substituição de um único aminoácido pode causar, por exemplo, uma mudança na função de uma proteína [102].

Podemos citar ainda como um dos objetivos destas áreas a localização e identificação de genes, exons (regiões que codificam proteínas), introns (regiões que não codificam proteínas) e seqüências reguladoras. Em relação a este último, requer-se comparações extensivas com seqüências de espécies relacionadas assim como com as de outras espécies.

Um outro objetivo é entender a estrutura da proteína, pois esta é essencial na determinação da função do genes. A função da proteína depende da sua estrutura tridimensional (3D), pois é a que ela assume no meio biológico. Dados os padrões de estrutura 3-D (motivos) que as proteínas assumem, que são mais conservados que os dos aminoácidos, este tipo de homologia (semelhança entre estruturas de diferentes organismos) é o mais adequado ao seu estudo. Certos motivos podem ter funções similares em diferentes proteínas e assim este tipo de informação pode ser importante para análise de genomas. Bancos de nucleotídeos, de proteínas e agora de genomas completos estão disponíveis para serem analisados. À medida que novas tecnologias de seqüenciamento são desenvolvidas, a atualização nos bancos de dados e a quantidade de "softwares"desenvolvidos, de domínio público ou privado, para análises dos resultados cresce enormemente [102].

Assim, a simples representação em um espaço tridimensional, concebido virtualmente no computador, de estruturas macromoleculares, é uma ferramenta útil para a identificação e análise dos sítios receptores destes complexos. O teste destes sítios receptores com pequenas moléculas, pré-existentes ou possíveis de serem sin-

tetizadas, pode ser realizado virtualmente em computadores através de simulações e da utilização de computação gráfica, possibilitando uma enorme economia de tempo de ensaios e síntese de possíveis fármacos [10].

Além das propriedades geométricas do encaixe molecular, outras propriedades físico-químicas são cruciais na função de reconhecimento molecular, tais como a distribuição de cargas elétricas e a natureza hidrofóbica ou hidrofílica dos grupos químicos envolvidos [10]. O desenvolvimento da Modelagem Molecular está fortemente relacionado com o extraordinário avanço da tecnologia computacional que tem tornado cada vez mais eficientes e acessíveis máquinas de todos os portes. Em especial, grandes avanços foram conseguidos em computação gráfica com o desenvolvimento de processadores específicos para o uso gráfico e programas especializados em simulações de imagens tridimensionais.

A Modelagem Molecular compreende pelo menos três etapas distintas:

1. A representação gráfica tridimensional de uma estrutura conhecida ou proposta, permitindo uma análise detalhada de regiões ou sítios de ligação correlacionados a atividade biológica e a comparação entre estruturas diferentes;

2. A simulação de um campo de forças (Mecânica Molecular) descrevendo as interações intra- e inter-moleculares, permitindo o estudo de conformações e configurações de energia mínima, possibilitando a previsão das estruturas mais prováveis e o estudo estatístico dos estados conformacionais;

3. Simulações de Dinâmica Molecular, podendo-se investigar a evolução temporal entre as várias conformações, incluindo-se os efeitos de temperatura e pressão sobre o sistema molecular.

A utilização de uma ou duas destas etapas de modelagem pode ser bastante útil; entretanto a combinação das três etapas fornece um método poderoso na

análise estrutural e dinâmica das moléculas biológicas. A aplicação destas técnicas no estudo de estrutura e função de biomoléculas tem despertado um grande interesse tanto em desenvolvimentos de pesquisas fundamentais como em aplicações nas áreas de Física, Química, Biologia, Ciência dos Materiais e Biotecnologia [21, 68].

A simulação computacional nestes últimos anos tem tido um papel de suma importância no desenvolvimento de estruturas em sistemas biomoleculares [42, 67]. As metodologias para o procedimento de simulação são oriundas dos princípios gerais das dinâmicas clássica e quântica e dos princípios de extremos envolvendo grandezas físicas. Isto dá origem aos métodos Gradientes, Dinâmica Molecular (DM), os métodos híbridos clássico-quânticos e aos métodos Estocásticos (ME)[68].

Geralmente, as técnicas envolvidas nestas metodologias são a da dinâmica clássica [12], usando campos de forças devidamente parametrizados, a partir de informações experimentais diversas acerca do sistema; e dos cálculos quânticos usando métodos *ab-initio* semi-empíricos para a determinação dos campos de forças com seus parâmetros para a determinação das funções potenciais a serem utilizadas nos esquemas estocásticos [19, 68].

Em particular, estas técnicas têm sido usadas como ferramenta na análise de estruturas moleculares, ligações químicas bem como nas interações moleculares que afetam muitos processos bioquímicos e biofísicos, tais como por exemplo o reconhecimento molecular. Um outro exemplo são as interações entre as bases dos ácidos nucleicos em DNA, que são responsáveis pela sua estrutura e de importância crucial para sua função.

Assim, muitos de nossos conhecimentos sobre a estrutura do DNA têm sido obtidos através de estudos em cristais de oligonucleotídeos. Outras fontes de informações são estudos de ressonância magnética nuclear de DNA em solução [29]. Recentemente, as técnicas experimentais têm sido completadas por estudos teóricos e computacionais. Isto refere-se à aplicação de mecânica-quântica bem como aos

métodos estatísticos. A principal idéia dos cálculos químico-quânticos é complementar os experimentos e proporcionar informação e predições que não são facilmente acessíveis por técnicas experimentais, de modo a elucidar a natureza dos processos estudados.

Sabe-se que o DNA é um sistema desafiador para estudos químico-quânticos, pois necessita-se uma descrição exata das interações intermoleculares fracas, manifestadas em ligações de hidrogênio bem como as interações de empilhamento dos pares de bases [99]. Na literatura, existem muitos trabalhos abordando vários métodos para o estudo de ligações de hidrogênio e empilhamento dos pares de bases no DNA, como por exemplo cálculos *ab initio* [29]. De outra forma, os métodos semi-empíricos, tais como AM1 e PM3, também são amplamente usados para o cálculo de estruturas eletrônicas, uma vez que estes são várias ordens de magnitude mais rápidos que os métodos *ab initio* [29].

Em particular, a Modelagem Molecular junto com o desenvolvimento de métodos de simulação e técnicas de computação gráfica tem permitido um avanço em estudos relacionados à análise e compreensão dos processos biológicos. Um exemplo é o estudo do mecanismo envolvido na formação de ligações de hidrogênio entre as bases (Adenina (A), Timina (T), Guanina (G) e Citosina (C)) no DNA, e a diversidade de modelos existentes resultante da interação entre estas bases, bem como as interações de ligações de hidrogênio em RNAs [10] .

As ligações de hidrogênio são a "chave" para muitos fenômenos químicos, incluindo a formação e a estabilização de estruturas secundárias, a flexibilidade e estabilidade de proteínas, reconhecimento molecular e reações enzimáticas que envolvem transferência de prótons [43]. Também, as ligações de hidrogênio entre as bases no DNA contribuem para a estabilidade da dupla hélice e proporcionam a especificidade para a transferência de informação genética. Entretanto, a geometria do par de bases também é um fator que influi nesta estabilidade [45]. No DNA encontramos somente dois pares de bases na estrutura padrão proposta por Watson e Crick [24]: Adenina com Timina (A,T) e Guanina com Citosina (G,C).

Mas, do crescente número de estudos sobre as estruturas das bases em DNA [4], sabe-se que além do modelo padrão proposto por Watson e Crick, que nos dá as formas complementares necessárias para permitir uma formação eficiente de ligações de hidrogênio, existem muitas outras estruturas, em que a relação espacial das bases nitrogenadas é diferente daquela encontrada no modelo padrão, mas satisfazem a evidência química relacionada à ligação de uma simples cadeia de nucleotídeos (tem estrutura com as bases perpendicular ao eixo da hélice) [24].

Desta forma, o método clássico do pareamento de bases de Watson e Crick é somente um dos vários tipos de pareamento existentes. Considerando que as ligações de hidrogênio entre as bases não podem ocorrer entre o N9-H das purinas (G,A) e o N1-H das pirimidinas (T,C), existem 29 diferentes hetero e homopares de bases, 28 deles descritos por Donohue [24] e o último descrito por Poltev e Shulypina [45, 77, 76]. Para estes diferentes modelos de estruturas propostos na literatura, existe um número significativo de trabalhos que abordam várias técnicas para a análise de propriedades e aspectos relacionados à formação dos pares de bases [17, 25, 38, 45, 73].

Dentre estes, cabe citar uma das primeiras pesquisas que foi realizada por Donohue [24], que investiga as estruturas dos pares de bases geometricamente aceitáveis em polinucleotídeos. Donohue desenvolveu um estudo sistemático, cuja idéia era investigar a possibilidade de formação de outras estruturas no processo de ligações de hidrogênio entre as bases (A,T,G e C), com uma relação espacial diferente daquela proposta no modelo padrão de Watson e Crick. Os resultados obtidos mostraram vinte e quatro modelos de pareamento possíveis entre as bases. Entretanto persistem dúvidas se todos são significativos na natureza. Donohue estabeleceu que a formação das ligações de hidrogênio nos modelos de pares de bases adicionais pode conferir um grau de estabilidade extra na estrutura do modelo de Watson e Crick e que os polinucleotídeos podem assumir outras estruturas de dupla cadeia, além daquela proposta por Watson e Crick.

Contribuições importantes na caracterização das estruturas dos pares de bases foram obtidas por Sponer et al. [91] com o estudo das propriedades (geometria mais favorável, energias de interação, entalpias e momentos dipolos) das ligações de hidrogênio entre os pares de bases de DNA. De acordo com os resultados obtidos, segundo os autores, as energias das estruturas dos pares de bases não são determinadas somente pelas ligações de hidrogênio, mas também são influenciadas pela polaridade dos monômeros e por uma ampla variedade de interações eletrostáticas secundárias, que envolvem também os átomos de hidrogênio ligados a átomos de carbono do anel aromático. Portanto, a estabilidade das ligações de hidrogênio nos pares de bases é um resultado de muitas contribuições que são acopladas e compensadas mutuamente e não podem ser completamente separadas [83].

Hobza e Sandorfy [45] investigaram teoricamente (usando cálculos *ab initio*) a estrutura e a energia de estabilização das ligações de hidrogênio dos 29 diferentes pares de bases formados por A,T,G e C. Eles verificaram, em seus resultados, que a energia de dispersão representa para todos os pares de bases uma contribuição importante da energia de estabilização. Ainda, segundo os autores, dos 29 diferentes pares de bases estudados, o par G-C (guanina-citosina) com três ligações de hidrogênio é o mais estável, o par G-G (guanina-guanina) com somente duas ligações de hidrogênio é comparavelmente estável. Além disso, diferentes pares com os mesmos tipos de ligações de hidrogênio (que são quase lineares) podem diferir consideravelmente em estabilidade.

Estes fatores sugerem que nem o número de ligações de hidrogênio, nem a linearidade, são unicamente responsáveis para a estabilidade dos pares de bases; é impossível explicar este fato usando somente os átomos que formam as ligações de hidrogênio; todos os átomos de ambos subsistemas devem ser considerados. Ainda concluíram [45] que os heteropares são mais estáveis que os homopares com pares de bases complementares. Os homopares, por outro lado, são mais estáveis que os heteropares com os pares de bases não complementares e a formação dos quatro

modelos de pares de bases entre A e T é igualmente provável, mesmo se a entropia é considerada [73].

Estudos realizados por Nir et al. [73], investigando os detalhes da estrutura química do DNA, através da técnica de laser, forneceram evidências do mecanismo envolvido no processo de ligações de hidrogênio entre as bases do DNA. Além dos vários estudos sobre a estrutura e função do DNA, estes autores investigaram se o caminho específico da ligação de hidrogênio entre as bases é característico da estrutura das próprias bases ou devido à influências externas. Eles usaram pulsos laser de 10 nm a 1062 nm para vaporizar guanina e citosina como moléculas simples, formando pares de bases isolados e analisaram as ligações de hidrogênio usando a técnica de espectroscopia. Quando as bases isoladas foram resfriadas, formaram pares de bases guanina e citosina em um caminho específico, o que significa que o pareamento entre as bases é mais provável devido à estrutura química que a influências externas.

Num trabalho posterior, Gonzalez et al. [38] utilizaram o método de Monte Carlo para simular a hidratação de moléculas separadas e de ligações de hidrogênio dos quatro modelos possíveis de pares de bases entre Adenina e Timina. Para os cálculos das características energéticas, estes autores encontraram um decréscimo no valor absoluto da energia de hidratação das bases separadas e concluíram que a hidratação de um par de bases é menos favorável que das bases separadas. Este decréscimo, segundo os autores, é maior que a energia de interação das bases; por isso os pares de bases não se formam espontaneamente em quantidades consideráveis em solução aquosa (e meio intracelular).

Da mesma forma, reconhece-se a importância das ligações de hidrogênio através do estudo realizado por Hobza e colaboradores [46]. Estes avaliaram as características termodinâmicas para a formação de ligações de hidrogênio em pares de bases no DNA, em fase gasosa usando dados *ab initio* (geometrias e freqüências vibracionais de bases e pares e energias de estabilização de pares). Segundo os autores, as energias de estabilização para vários pares de bases no DNA diferem

consideravelmente (mais de 200%), uma vez que os termos entrópicos são mais uniformes e variam menos que 40%. Ainda destacam que para várias estruturas de um par de bases, a entropia é quase constante (por exemplo para os 4 modelos de pares de bases entre A e T). Concluíram ainda que apesar disto, a inclusão do termo entrópico produz algumas mudanças na ordem da estabilidade dos pares de bases. A mudança mais importante é em relação aos dois pares mais estáveis, GCWC e GG1, onde o par GG1 apresentou ligeiro acréscimo na energia de estabilização em relação ao GCWC. A conclusão sobre a preferência do par GG1 é surpreendente e contradiz a visão tradicional da estabilidade no DNA.

Além de estudos sobre as características energéticas em modelos de pares de bases, são de grande importância os que abordam aspectos geométricos sobre as estruturas dos ácidos nucleicos, os quais propiciam a interpretação da complexidade existente nas configurações tridimensionais destas estruturas. Um número significativo de trabalhos tem sido publicado para enfatizar estas características; exemplos podem ser encontrados nos artigos de Babcok [6] e [5] e colaboradores, onde são apresentados os parâmetros matemáticos que descrevem as relações espaciais entre as estruturas dos ácidos nucleicos: empilhamento e pareamento de bases.

Outro trabalho que integra não só a análise das interações entre as bases dos ácidos nucleicos, mas também os efeitos e as influências que resultam da determinação do *ponto pivô* em relação ao par de bases, é habilmente apresentado por Gendron et al. [37]. Segundo os autores, o *pivô* é o ponto relativo ao sistema de coordenadas do par sobre o qual a base rotaciona. Neste artigo os autores indicam que muitas das dependências discutidas na literatura são resultado das influências entre os métodos matemáticos e a escolha do sistema de origem e eixo relacionado ao par. Esta influência é significativamente evidenciada se os pontos *pivôs* não são mantidos rigorosamente uniformes, quando os cálculos de todos os parâmetros de uma classe particular são realizados; como por exemplo, em cálculos dos ângulos e distâncias relacionando bases complementares.

Com esta hipótese, os referidos autores reforçam que a determinação do ponto *pivô* associado à geometria do par possibilita modelar os movimentos físicos das bases, bem como a habilidade para analisar comparavelmente as estruturas que envolvem as relações dos pares de bases, sejam eles Watson e Crick, Hoogsteen, etc.

Trabalhos semelhantes, que abordam conceitos matemáticos básicos para a análise tridimensional das estruturas de DNA, são apresentados por Suzuki et al. [96]. Estes indicam a necessidade de um ponto de referência comum para descrever o mecanismo dos arranjos tridimensionais de bases e seus pares em estruturas de ácidos nucleicos. Para tal são definidos os comprimentos virtuais entre os carbonos $1^{'}$ das respectivas bases por $d_{C1'...C1'}$ e os ângulos virtuais entre o $N9 - C1^{'}...C1^{'}$ da purina e $N1 - C1^{'}...C1^{'}$ da pirimidina dados por λ. Através de experimentos realizados, estes autores afirmaram que a localização da origem do sistema de referência em relação ao par é significativamente dependente do comprimento do par idealizado, ou seja, a distância entre os átomos $C1^{'}$ de cada base do par estabelecido e o pivotamento destas complementares, λ, no plano do par determinado.

Os resultados apresentados por estes autores evidenciam pequenas variações no comprimento das ligações de hidrogênio na estrutura do par analisado. Segundo estes, as variações são devido às pequenas deformações impostas na geometria do par para que este se ajuste ao modelo padrão do par de bases assumido. O sistema de coordenadas determinado é baseado no par padrão de Watson e Crick. Eles também enfatizam que pequenas mudanças impostas nas configurações das estruturas têm pouca influência relativa no modelo do par de bases padrão. Por exemplo, o aumento de λ de 54.5^0 para 55.5^0 afeta a distância $d_{C1'...C1'}$ em apenas 0.1 Å, atribuindo um desvio nos comprimentos das ligações de hidrogênio por menos de 0.05 Å. Portanto, é preciso uma escolha cuidadosa do sistema de referência para cada modelo de par.

Modelos de ligações de hidrogênio também são de suma importância, principalmente em estruturas de RNAs [70, 69]. Dada a sua importância, a investigação na formação dos pares de bases têm sido exaustivamente estudada para

mostrar a diversidade de modelos de pareamentos com ênfase particular, sobre os tipos não-canônicos e uma nomenclatura sistemática tem sido proposta [3, 57]. Do ponto de vista dos modeladores, as relações espaciais definidas por tais interações de ligações de hidrogênio podem ser usadas para definir a pesquisa do espaço conformacional (exploração dos arranjos espaciais (formas) energeticamente favorável de uma molécula (conformações)) em RNA.

O trabalho proposto por Lemieux e Major [57] aborda o estudo de identificação objetiva e sistemática de pares de bases canônicos e não-canônicos em estruturas tridimensionais de RNAs. Com uma aproximação probabilística e a implementação computacional que detecta e analisa todos os tipos de pares contidos em estruturas tridimensionais de RNAs, estes autores utilizaram um algoritmo para distinguir objetivamente todos os tipos de pareamentos de bases canônicos e não-canônicos formados por três, duas e uma ligação de hidrogênio, bem como modelos de pares com ligações bifurcadas.

O reconhecimento da importância das ligações de hidrogênio e sua extensiva ocorrência em macromoléculas biológicas aparece também no trabalho de Lindauer e colaboradores [59]. Estes criaram o programa *HBexplore*, uma ferramenta para identificar e analisar os padrões de ligações de hidrogênio em macromoléculas biológicas. Este programa possibilita a seleção de todos os potenciais de ligações de hidrogênio de acordo com um critério geométrico. Deste modo, contribui para a elucidação de princípios gerais sobre a arquitetura de macromoléculas biológicas e para a predição e refinamento de estruturas individuais. Os autores reforçam ainda que os resultados obtidos em *HBexplore* dependem da qualidade da estrutura de dados e da seleção do critério geométrico apropriado. Este programa pode ainda contribuir para o aperfeiçoamento das ligações de hidrogênio nos modelos construídos durante a determinação de estruturas experimentais.

Ainda enfatizando a importância da geometria em ligações de hidrogênio e pares de bases, Fabiola e colaboradores [31] desenvolveram uma função dependente da direção da ligação e um critério de seleção apropriado para o aper-

feiçoamento em refinamento cristalográfico. Neste sentido, os potenciais de ligações de hidrogênio, calibrado neste caminho pelos dados de cristalografia de proteínas, podem ser otimizados para o envolvimento com proteínas e, então, ser usado para melhor descrever as ligações em simulações moleculares ou refinamento de outras proteínas. Entretanto, segundo os autores, o aperfeiçoamento é efetivo somente quando se usa um critério severo na verificação da ligação.

Após essa breve descrição de trabalhos encontrados na literatura, podemos inferir que existe uma grande quantidade de trabalhos sobre as características energéticas e estruturais das ligações de hidrogênio dos vários modelos de pares de bases do DNA e RNA. Contudo, não se tem conhecimento de estudos numéricos que abordem aspectos quantitativos da probabilidade de formação para a variedade de modelos de pares existentes, considerando simultaneamente critérios geométrico e energético. Apesar de aqui realizar-se cálculos de otimização para os pares de bases, sabe-se que estudos que abordem este tópico são experimental e teoricamente bem conhecidos na literatura e portanto, não é o foco principal deste trabalho. Identificar a probabilidade de formação de cada par de bases em uma determinada classe de modelos através de critérios geométrico e energético é o principal interesse deste trabalho. Trata-se, portanto, de um tema original e com um amplo potencial de interesse aplicado.

A importância em analisar os processos de formação de pares de bases bem como a diversidade de modelos existentes, considerando os princípios citados, é fundamental, pois através destes pode-se prever quais modelos tendem a ocorrer em maior quantidade e os que são menos previsíveis, obter regras preditivas na formação particular destes modelos, entender os aspectos relacionados à imperfeições nestes pares e suas conseqüências nos sistemas biológicos [100].

Outro aspecto relevante é que, uma vez que os ácidos nucleicos constituem o material genético, admite-se que porções sucessivas destas moléculas sejam diferentes e, portanto, contenham diferentes partes de informação. A diferenciação na cadeia não pode ser devido à pentose ou ao fosfato, uma vez que cada um deles

está presente em todos os nucleotídeos. Portanto, todas as diferenças na informação *genética* ao longo do comprimento da cadeia de polinucleotídeos, devem ser devidas às bases orgânicas A,T,G e C presentes. Deste modo, o conteúdo de informação de um ácido nucleico está nas suas bases orgânicas [43].

Adicionalmente, a determinação da ocorrência (maior freqüência) de um determinado modelo de par de bases, frente aos vários tipos possíveis existentes, pode ser importante se correlacionadas com seu freqüente aparecimento em determinadas posições na dupla-hélice. Ou seja, as distorções locais observadas na dupla-hélice poderão estar relacionadas com o número (freqüência) e o tipo de pareamento (conformações) que são geométrica e energeticamente mais acessíveis dentre os possíveis pareamentos existentes. Assim, devido à variedade estrutural existente entre os modelos de ligações com pares de bases, um estudo dos que apresentam maior freqüência pode auxiliar na identificação dos modelos que são reconhecidos em processos associados à mutagênese.

Ainda destaca-se que a análise da probabilidade da ocorrência de cada modelo em uma determinada classe de par de bases (AT, AA, TT, GC, GG, CC) é importante para obter dados sobre quais modelos apresentam maior probabilidade e que tipo de estrutura contribui para a sua formação. Pois, uma vez sabendo-se o tipo de modelo que contribui para a elevada probabilidade de formação do par, sabe-se quais sítios (átomos) propiciam as ligações, bem como os tipos de ligações de hidrogênio envolvidos no modelo. Como uma aplicação prática deste tipo de estudo, pode-se destacar a análise de regiões com uma composição particular de seqüências repetitivas de bases, como regiões de introns. Por exemplo, num estudo particular de seqüências de pares de bases ricas em G e C [52], analisar quais tipos de estruturas tem a maior probabilidade de ocorrência, uma vez que, para interações entre G e C existem três modelos diferentes.

Outro exemplo de aplicação prática do estudo da probabilidade de formação destes modelos é em regiões com grandes *repetições de dímeros idênticos* [23], em seqüências codificadoras e não-codificadoras em DNA.

Neste sentido, o presente trabalho objetiva, por intermédio de simulação computacional via Método de Monte Carlo, avaliar os processos de formação de ligações de hidrogênio entre as bases biológicas Adenina, Timina, Guanina e Citosina, bem como investigar dentre todos os possíveis modelos de pares de bases existentes, a probabilidade de formação associada a cada modelo, considerando os critérios citados anteriormente. A simulação aqui proposta permitirá, dentro das limitações do presente modelo, isto é, bidimensional e sem a consideração explícita das moléculas de solvente (água), a previsão da probabilidade de ocorrência das estruturas para cada modelo de uma determinada classe de par de bases, através do estudo estocástico, a identificação de estruturas que envolvem diferentes modos de pareamento de bases biológicas, contribuindo para uma maior compreensão da base molecular química. Em suma, é importante investigar as ligações de hidrogênio entre as bases heterocíclicas (pirimidinas e purinas), pois elas são o fundamento da estrutura e correspondem a interações que ocorrem durante os processos de translocacão, transcrição e replicação do DNA [79].

1.1 *Objetivos gerais e específicos*

O objetivo do presente trabalho é investigar o processo de formação de uma estrutura resultante da associação de duas subunidades, ou seja, as bases biológicas Adenina, Timina, Guanina e Citosina, considerando os critérios geométrico e energético. Deste modo, abrange o estudo probabilístico dos aspectos quantitativos e qualitativos, nos processos de formação da variedade de pares de bases biológicas existentes, considerando os critérios citados anteriormente.

Neste sentido, o algoritmo desenvolvido e implementado, aqui usando a técnica de Monte Carlo, apresenta uma versão mais eficiente em relação aos apresentados por Netz e Samios [72] em 1992 e Inda e Samios [49] em 1995.

Em virtude da complexidade inerente ao sistema químico analisado, o modelo computacional desenvolvido apresenta algumas simplificações: as bases biológicas são modeladas em espaço bidimensional e não há consideração das moléculas do solvente. O estudo do pareamento de bases adicionais em DNA pode contribuir não somente para o acréscimo da capacidade em armazenar informação, mas também para a replicação do DNA contendo novos grupos funcionais.

Mais especificamente, este trabalho busca investigar as seguintes questões:

- a compreensão dos mecanismos envolvidos no processo de formação de ligações de hidrogênio entre os pares de bases;
- a análise da previsibilidade dos possíveis tipos de pareamentos de bases existentes em regiões específicas;
- a estimativa da freqüência de ocorrência de diversos pares (possíveis) entre as bases através da acessibilidade geométrica;
- a análise quantitativa nos processos de formação dos vários modelos de pares de bases utilizando inicialmente o critério geométrico;

- a habilidade para analisar a quantidade de cada uma das estruturas nas relações dos pares de bases, sejam eles de Watson e Crick, Hoogsteen, etc;

- o estudo probabilístico da análise quantitativa na formação dos vários modelos de pares de bases, considerando simultaneamente critérios geométricos e energéticos.

Neste sentido, as principais tarefas a serem desenvolvidas referem-se ao desenvolvimento e à implementação numérica do algoritmo para a simulação do processo estocástico do sistema estudado, bem como a execução do mesmo para o conjunto de experimentos estabelecidos para a análise do sistema proposto. A seguir apresenta-se uma revisão bibliográfica direcionada ao tema da tese.

2 UMA BREVE INTRODUÇÃO À BIOLOGIA MOLECULAR

Neste capítulo, é feita uma pequena revisão de tópicos fundamentais sobre a química de compostos biológicos. Discute-se a relevância dos compostos orgânicos em sistemas biológicos e apresenta-se uma abordagem específica sobre o tema tratado neste trabalho.

Ao conjunto de todos os genes existentes em qualquer organismo - humano ou não, denomina-se Genoma, uma vez que este contém a informação para todas as estruturas e atividades celulares durante todo o tempo de vida das células e dos organismos. Encontrado em todo núcleo de cada célula (trilhões de células no homem), o genoma consiste numa cadeia de ácido desoxirribonucleico (DNA), associado a moléculas de proteínas e organizado na forma de uma estrutura denominada cromossoma. Para cada organismo, desde uma simples bactéria até o complexo genoma humano [32], os componentes desta cadeia contêm a informação necessária para a construção e manutenção da vida. Entender como o DNA desempenha este papel requer conhecimento de sua estrutura e de sua organização [60, 102].

Assim, destaca-se como uma das principais realizações da ciência, no século XX, a descoberta da dupla hélice do DNA, feita por Watson e Crick em 1953 [104]. Graças ao trabalho destes dois cientistas, e também às importantes pesquisas que antecederam as suas descobertas, sabemos hoje que o DNA (ácido desoxirribonucleico) é a molécula que contém todas as informações necessárias para a formação das cadeias polipeptídicas, a estrutura primária das proteínas. Sabemos, também, que os genes, os "fatores" hereditários abstratos, primeiramente descritos por Mendel [39], são segmentos de DNA e, portanto, formados por pares de nucleotídeos.

Antes de falarmos sobre a estrutura em dupla hélice do DNA e de sua importância na transmissão das características hereditárias, vamos apresentar um pequeno histórico das descobertas que antecederam e, de certa maneira, permitiram,

aquelas de Watson e Crick. O DNA foi descoberto por Friedrich Miescher [39], em 1869. Já em 1944, Avery, McLeod e McCarty [39] propuseram que o DNA contem as funções principais e transmite as informações gênicas; isto significa que ele serve de molde para a síntese de RNAs e alguns destes serão traduzidos em proteínas correspondentes. Estas funções foram comprovadas por Hershey em 1953. Baseado neste conjunto de informações, em 1953 Watson e Crick publicaram suas conclusões sobre a investigação da estrutura química da molécula de DNA [74, 104]. Esse modelo explicava as regularidades da composição das bases, especialmente sua duplicação na célula. O mérito do modelo proposto é que, a partir dele, tornou-se possível explicar as propriedades (químicas e físicas) do DNA e também a capacidade de autoduplicação do material genético.

O DNA é um componente dos ácidos nucleicos. Junto com o RNA (ácido ribonucleico), estas duas moléculas são as maiores e mais importantes moléculas orgânicas e são encontradas em toda e qualquer forma de vida, desde um pequeno vírus ou mesmo uma bactéria, até as mais especializadas formas de vida como os vegetais e os animais. Eles foram observados pela primeira vez no núcleo da célula, o que determinou seus nomes. Após pesquisas, descobriu-se que os ácidos nucleicos são encontrados também em outras partes da célula fora o núcleo [26, 74, 75]. O DNA é considerado a molécula fundamental da maioria dos seres vivos. É ele que contém todas as informações genéticas de cada indivíduo, e que tem a capacidade de transmiti-las à sua descendência. O RNA é uma molécula intermediária na síntese de proteínas; faz a intermediação entre o DNA e as proteínas. As principais diferenças entre o RNA e o DNA são sutis, mas essas diferenças fazem com que o DNA seja mais estável.

Em suma, a principal importância dos ácidos nucleicos para os seres vivos corresponde ao fato de que eles comandam todo o funcionamento das células e do organismo, pois formam os genes onde estão codificadas as instruções que serão utilizadas na síntese de proteína [39].

2.1 Estrutura do DNA e do RNA

Após ficar claro o papel central do DNA na hereditariedade, muitos cientistas começaram a determinar sua estrutura. [39]. Embora a estrutura do DNA não fosse conhecida, os blocos constituintes do DNA já eram conhecidos há muitos anos. Sabia-se que o DNA era formado por apenas quatro moléculas fundamentais chamadas nucleotídeos, que são as unidades estruturais dos ácidos nucleicos [30]. Estes eram formados por três unidades básicas: uma base orgânica nitrogenada ligada ao carbono 1 de uma pentose (açúcar de 5 carbonos), que por sua vez se liga, pelo carbono 5, a um grupo fosfato, como se verifica na figura 2.1.

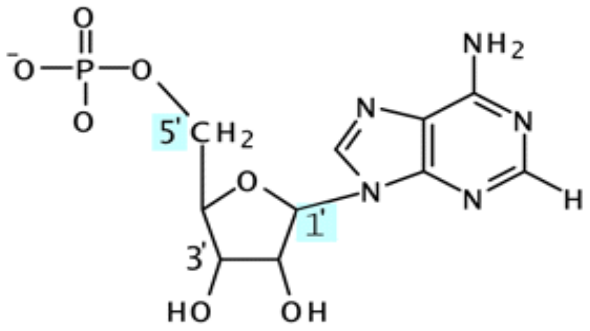

Figura 2.1: Estrutura do Nucleotídeo

O açúcar era uma pentose, com sua estrutura formada por cinco átomos de carbono, sendo numerados em ordem, e denotados por um número seguido de uma marca (’). O carbono de número 1, denotado por 1’ como o primeiro carbono da molécula da direita para a esquerda. Num movimento circular, no sentido horário, tem-se os carbonos 1’, 2’, 3’, 4’ e 5’, caracterizando assim todos os carbonos da molécula do açúcar, como pode ser verificado na figura 2.2.

Como na seqüência de nucleotídeos o açúcar presente tem um átomo de hidrogênio no carbono 2’, tem-se assim um átomo de oxigênio “em falta”junto a este carbono e, por isso, denomina-se o açúcar de desoxirribose. O fosfato pode

ligar-se ao açúcar na forma de um, dois, ou três grupos fosfato e o nucleotídeo denominar-se-á mono, di e trifosfato, respectivamente.

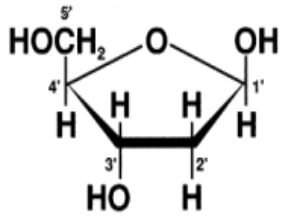

Figura 2.2: Estrutura da Desoxirribose

A base nitrogenada se liga à molécula de açúcar pelo carbono 1' e o fosfato ao carbono 5'. Esta é a estrutura denominada de nucleotídeo, como mostrado na figura 2.1. Os nucleotídeos são as unidades básicas dos ácidos nucleicos. Os quatro nucleotídeos encontrados no DNA são combinações de Adenina, Guanina, Citosina, ou Timina com desoxirribose e fosfato [94].

As bases nitrogenadas são compostos orgânicos heterocíclicos de carbono, nitrogênio, hidrogênio e, em alguns casos, oxigênio. A Adenina e a Guanina, por terem estruturas similares, formadas por 2 anéis conjugados, um anel de pirimidina ligado a um anel pentagonal (chamado anel imidazol), foram denominadas de purinas; a Citosina e a Timina, formadas por um anel com estrutura hexagonal, foram denominadas de pirimidinas. As quatro bases nitrogenadas, por apresentarem nomes de certa maneira *complexos*, como vistos acima, são referidos normalmente pelas iniciais de suas bases: A (Adenina), C (Citosina), G (Guanina) e T (Timina). Pela composição dos átomos, as bases são rígidas e seus átomos se encontram aproximadamente no plano [43].

Apesar das bases possuírem dois tamanhos; onde as pirimidinas (T,C) são menores que as purinas (G,A), como podemos verificar nas figuras 2.3 e 2.4, os pares de bases A...T e G...C têm aproximadamente o mesmo tamanho [105]. Dessa maneira, os dois pares ocupam o mesmo espaço, permitindo uma dimensão

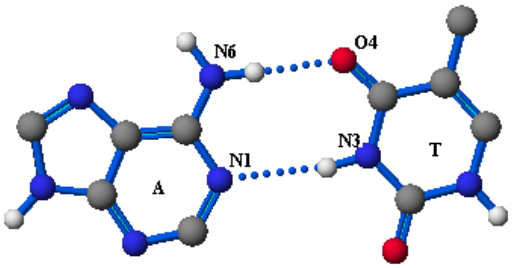

Figura 2.3: Par de bases: Adenina e Timina

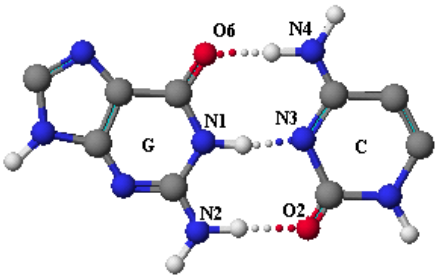

Figura 2.4: Par de bases: Guanina e Citosina

uniforme ao longo da molécula de DNA. Assim, não existe nenhuma restrição quanto à seqüência de nucleotídeos ao longo do DNA [105].

Watson e Crick [104] trabalharam com difração de raios X da estrutura de DNA, o qual pode ser resumido da seguinte maneira: os raios X são disparados sobre as fibras de DNA e a difusão dos raios pelas fibras é observada em filmes fotográficos. Quando os raios passam pelas fibras eles produzem pontos nos filmes fotográficos; esses pontos podem ser interpretados como ângulos de difusão, que nos informam a posição de cada átomo ou conjuntos de átomos na molécula de

DNA. Estes primeiros dados sugeriam que a molécula de DNA seria longa e fina, composta por dois filamentos paralelos e ligados um ao outro, e que seria uma molécula helicoidal.

Outros dados utilizados por Watson e Crick vieram do trabalho de Erwin Chargaff [103], que determinou, a partir do DNA de muitos organismos, que a quantidade total de nucleotídeos pirimídicos (T+C) é sempre igual a quantidade dos púricos (A+G). Com todas estas informações, Watson e Crick deduziram uma estrutura tridimensional para a molécula de DNA. Esta estrutura seria a de uma dupla hélice enrolada ao longo de um mesmo eixo com sentido rotacional à direita, onde cada hélice é uma cadeia de nucleotídeos, mantidos juntos por ligações fosfodiéster (um grupamento fosfato formando uma ligação com os grupamentos OH) [40].

A seqüência de bases púricas e pirimídicas na molécula de DNA constitui o código genético, enquanto os grupos pentoses fornecem a estrutura de sustentação da molécula. Ainda com base nestes estudos, concluiu-se que na dupla hélice as duas fitas de DNA (nucleotídeos) estão em direções opostas; isto significa que são anti-paralelas. Diz-se, portanto, que elas têm polaridade inversa. Com base na estrutura de dupla hélice e nas características de hidrofobicidade das moléculas, a estrutura do DNA fica da seguinte forma, conforme mostra a figura 2.5 onde:

1. Duas cadeias de polinucleotídeos orientadas em sentidos opostos se enrolam em torno de um eixo comum para formar uma hélice dupla.

2. As bases purinas e pirimidinas encontram-se no interior da hélice, enquanto os fosfatos e as unidades de desoxirriboses estão no exterior da hélice.

3. A Adenina (A) liga-se à Timina (T) e a Guanina (G) à Citosina (C). Os pares A-T são unidos por duas ligações de hidrogênio, e os pares G-C por três destas ligações.

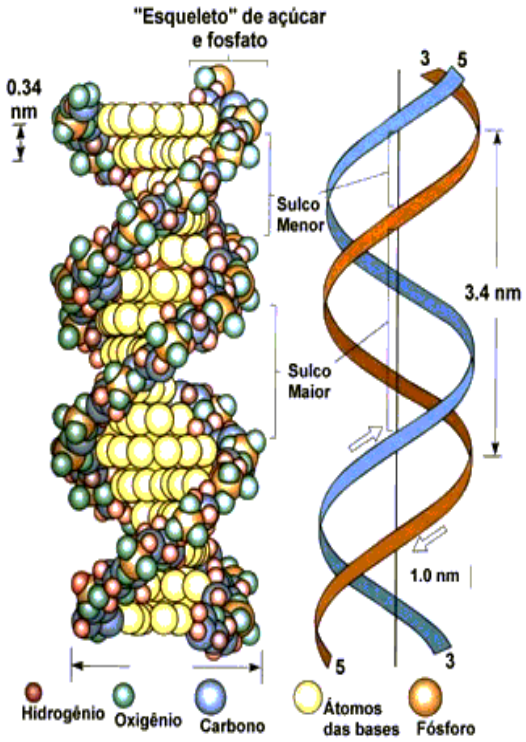

Figura 2.5: Estrutura da dupla-hélice [39]

O pareamento (ligação) das bases de cada fita se dá de maneira padronizada, sempre uma purina com uma pirimidina, ou mais especificamente: Adenina com Timina e Citosina com Guanina; portanto, é preciso respeitar uma certa afinidade molecular. Essas características de pareamento explicam o fato que, em qualquer seqüência de DNA, a relação molar entre A/T é igual a 1.0, o mesmo ocorrendo com a relação G/C, embora as concentrações molares entre A...T e G...C variem com a seqüência de DNA analisada. Essa característica de pareamento tem grande significância fisiológica e, devido a ela, as duas fitas de DNA são ditas complementares [105].

As cadeias complementares da dupla hélice têm polaridade invertida, antiparalelas: um filamento está no sentido 5'-3', já o outro progride no sentido 3'-5'. Na cadeia de DNA os nucleotídeos possuem a mesma orientação, estando o carbono 5' da cadeia da pentose, um grupo fosfato e, na extremidade 3', um grupo hidroxila [81].

Portanto, a ligação entre Guanina e Citosina é mais forte que a existente entre a Adenina e Timina [45]. As ligações ocorrem entre átomos de hidrogênio com uma pequena carga positiva e átomos aceptores com uma pequena carga negativa. Desta maneira, o nitrogênio da base tende a atrair para si os elétrons (carga negativa) que participam da ligação N-H, tornando o hidrogênio do grupamento $N-H_2$ da base nitrogenada levemente positivo. Por outro lado, o oxigênio da base complementar possui seis elétrons que formam uma nuvem em volta daquele átomo, tornando-o levemente negativo [14].

Assim, conclui-se que as pontes de hidrogênio são determinadas pela presença de grupos contendo um hidrogênio ligado a um elemento fortemente eletronegativo, por exemplo, O-H, N-H. O hidrogênio ligado a este tipo de átomo interage fortemente com átomos também fortemente eletronegativos presentes na mesma ou em outra molécula. Deste modo, a ligação de hidrogênio é a interação entre dois átomos por "intermédio"do hidrogênio. O átomo "de onde"a ligação provém é denominado *doador* e o que "aceita"a ligação, *aceptor* [56, 75].

Uma característica importante das pontes de hidrogênio é o seu caráter profundamente direcional, ou seja, a ponte de hidrogênio é mais forte se os átomos participantes estiverem "frente a frente"nas orientações ideais [71]. A presença de ligações de hidrogênio pode ser identificada com facilidade a partir do momento que analisamos a estrutura química de uma substância. Se os átomos doador e aceptor da ligação de hidrogênio estiverem em moléculas diferentes, forma-se uma ponte intermolecular. Se, ao contrário, os átomos fortemente eletronegativos estiverem presentes na mesma molécula, a ligação é intramolecular e não terá o mesmo efeito de interação que a intermolecular [71].

As ligações de hidrogênio são muito fracas quando comparadas com outros tipos de ligações químicas; mas esta característica, entretanto, é de suma importância para as funções do DNA relacionadas à hereditariedade. Se as ligações de hidrogênio não fossem do tipo fracas, a dupla hélice não teria tendência à ruptura, o que dificultaria o DNA de passar para as sucessivas gerações, além de comprometer

todos os outros processos relacionados à divisão celular e à formação das proteínas. Sendo mais fracas que as ligações covalentes, elas são contudo as interações mais significantes na flexibilidade e estabilização das moléculas de DNA e RNA [98].

Além da ponte de hidrogênio, outras forças agem em conjunto para estabilizar a estrutura da dupla hélice do DNA. É fundamental que essas forças sejam fortes o suficiente para manterem sua integridade, mas devem permitir uma flexibilidade conformacional, que é essencial para sua atividade [79]. Além das ligações covalentes, que unem os átomos nas colunas, outras forças mais fracas atuam no DNA. Efeitos hidrofóbicos (moléculas ou grupos funcionais de moléculas que são pobremente solúveis em água) estabilizam o pareamento - os anéis purínicos e pirimídicos das bases são forçados para o interior da dupla-hélice, por coesão interna de moléculas de água, e os sítios hidrofílicos (relativo à molécula ou grupo de moléculas que facilmente se associam à água) das bases ficam expostos ao solvente nas cavidades. A força das ligações entre as fitas complementares depende também das bases vizinhas mais próximas. Os vizinhos mais próximos são as bases que formam cada fita. No DNA natural, por exemplo, existe regiões ricas em "string"de A, que se ligam a "string"de T da fita complementar. Por exemplo, $5'\ AAAAAA\ 3'\ \ 3'\ TTTTTT\ 5'$ Esses locais favorecem estruturas geometricamente diferenciadas das demais regiões da molécula, por exemplo, formação de cunhas (DNA Bent), que podem alterar a probabilidade de ligação entre os pares. O empilhamento das bases no interior da hélice permite o estabelecimento de forças de Van der Walls entre os anéis aromáticos das bases adjacentes que são fracas, mas aditivas na manutenção da estrutura. A ligação fosfodiéster, como apresentada na figura 2.6, une a hidroxila 5' da desoxirribose à hidroxila 3' de um desoxirribonucleotideo adjacente, para formar uma estrutura repetida.

A dupla hélice corresponde muito bem aos dados de raios X e também aos dados de Chargaff [39]. Analisando os modelos destas estruturas, Watson e Crick concluíram que o raio observado na dupla-hélice (conhecido pelos dados de raios X) seriam explicados se uma base purínica sempre se parear (por pontes de

hidrogênio) com uma base pirimidínica. Tal pareamento contribuiria para a regularidade (A+G)=(T+C) observada por Chargaff [101], mas previa quatro possíveis pareamentos: T...A, T...G, C...A e C...G. Os dados de Chargaff, entretanto, indicavam que T só se pareia com A, e C só com G. Watson e Crick mostraram que apenas estes dois pareamentos tinham a complementaridade tipo "chave e fechadura"necessária para permitir um eficiente pareamento por pontes de hidrogênio [90].

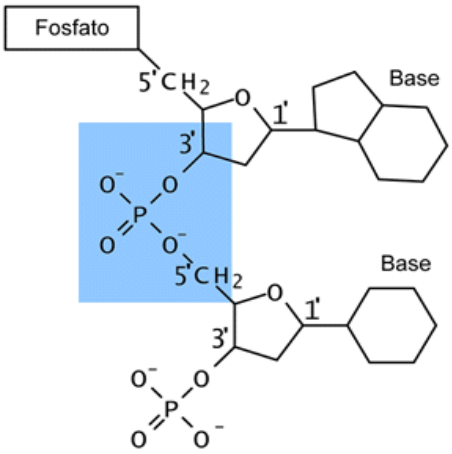

Figura 2.6: Ligações fosfodiéster

De maneira resumida, podemos dizer que o conhecimento da estrutura da molécula de DNA elucidou muitas questões em vários campos da Biologia, principalmente na Genética. Isto deveu-se a dois motivos fundamentais. O primeiro estaria relacionado ao modo pelo qual o DNA pode ser duplicado e replicado graças à natureza das pontes de hidrogênio, que fazem a ligação entre as bases nitrogenadas. O segundo seria que esta estrutura sugere, talvez, que a seqüência dos pares de nucleotídeos esteja ditando a seqüência dos aminoácidos durante o processo da síntese de proteínas, ou seja, um tipo de código genético transformaria as informações do DNA; no caso a sequência de nucleotídeos, numa seqüência de aminoácidos, a qual caracterizaria determinada proteína [105].

O modelo da estrutura do DNA foi fundamental para se compreender, posteriormente, que as informações genéticas estão organizadas nessa molécula na forma de um código de leitura e que esse código é comum aos diferentes organismos.

O RNA, por outro lado, é uma molécula de ácido nucléico formada, geralmente, por uma só cadeia, como pode-se observar na figura 2.7. A seqüencia de bases (estrutura primária) é similar à do DNA, exceto pela substituição da desoxirribose por ribose e de Timina (T) por Uracil (U).

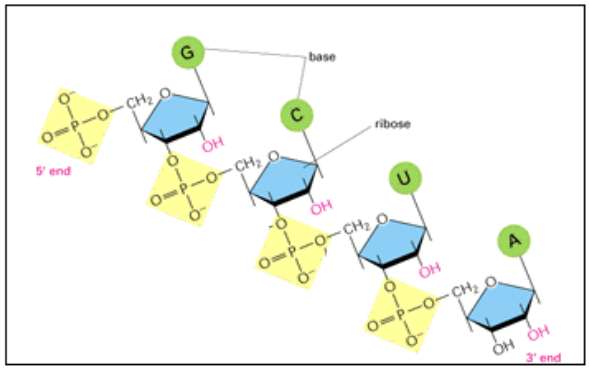

Figura 2.7: Estrutura do RNA

As outras três bases Adenina, Citosina e Guanina também estão presentes. O RNA encontra-se normalmente na forma de fita simples, embora os pareamentos entre C e G e entre A e U possam ocorrer entre regiões da própria cadeia, formando estruturas secundárias que são importantes na função dos RNAs e no reconhecimento de proteínas-RNAs [105].

Os principais tipos de RNA são os RNAs mensageiros (mRNAs), os transportadores (tRNAs) e os ribossomais (rRNA). Os RNAs mensageiros são aqueles que codificam as proteínas. Os RNAs ribossomais fazem parte da estrutura do ribossomo, junto com diversas proteínas e são eles que permitem a ligação entre dois aminoácidos na síntese de proteínas. Os RNAs transportadores carregam o aminoácido específico para complementar a seqüência de nucleotídeos do mRNA, quando este está ligado (unido) ao ribossomo.

2.2 Tipos de DNA e suas propriedades físicas e químicas

O DNA pode assumir diferentes conformações, dependendo da sua composição de bases e do meio em que se encontra. Estudos sobre as estruturas mostram que existem duas formas de DNA dextrosa (dupla hélice com giro para a direita), chamadas A-DNA e B-DNA, e uma forma sinistrosa (dupla hélice com giro para a esquerda) chamada Z-DNA [14]. A diferença entre as duas formas que giram para a direita está na distância necessária para fazer uma volta completa da hélice devido ao ângulo que as bases fazem com o eixo da hélice.

A forma B-DNA tem a dupla-hélice longa e estreita: para completar uma volta na hélice são necessários 10 pares de bases. A forma A-DNA é mais curta e grossa, sendo necessários 11 pares de bases para completar uma volta na hélice. Em solução, geralmente o DNA assume a conformação B. Quando há pouca água disponível para interagir com a dupla hélice, o DNA assume a conformação A-DNA.

A terceira forma de DNA difere das duas anteriores; pode ocorrer em filamentos de DNA que possuam purinas e pirimidinas alternadas, pois seu sentido de rotação é para a esquerda [105]; este tipo de DNA é chamado de Z-DNA. Esta conformação é mais alongada e mais fina que o B-DNA; para completar uma volta na hélice são necessários 12 pares de bases [14].

Soluções de DNA, em pH = 7,0 e à temperatura ambiente, são altamente viscosas; mas, a altas temperaturas ou pH extremos o DNA sofre desnaturação [98], pois ocorre ruptura das pontes de hidrogênio entre os pares de bases. Esta desnaturação faz com que diminua a viscosidade da solução de DNA. Durante a desnaturação nenhuma ligação covalente é desfeita ficando, portanto, as duas fitas separadas. Quando o pH e a temperatura voltam ao normal, as duas fitas de DNA espontaneamente se enrolam, formando novamente o DNA dupla fita. Este processo, como veremos a seguir, envolve duas etapas: a primeira é mais lenta pois envolve o encontro casual das fitas complementares de DNA, formando um curto segmento de

dupla hélice; a segunda é mais rápida e envolve a formação das pontes de hidrogênio entre as bases complementares reconstruindo a conformação tridimensional.

2.3 Replicação do DNA

Replicação do DNA é o processo de auto-duplicação do material genético, mantendo assim o padrão de herança ao longo das gerações [105]. A replicação do conteúdo informacional inicia com a separação local da dupla hélice, que está interligada pelas pontes de hidrogênio formadas entre as bases complementares. Cada fita atua como um molde para a formação de uma nova molécula de DNA. Desta forma, as informações genéticas permanecem fiéis às contidas na molécula que as originaram.

Duas teorias [39] tentaram explicar a replicação do DNA. A teoria conservativa, onde cada fita do DNA sofre duplicação, e as fitas formadas sofrem pareamento, resultando num novo DNA dupla fita, sem a participação das fitas "parentais"(fita nova com fita nova formam uma dupla hélice e fita velha com fita velha formam a outra fita dupla). Por outro lado, na teoria semi-conservativa, cada fita do DNA é duplicada formando uma fita híbrida, isto é, a fita velha pareia com a nova formando um novo DNA; de uma molécula de DNA formam-se duas outras iguais a ela. Cada DNA recém formado possui uma das cadeias da molécula mãe; por isso o nome semi-conservativa.

Falamos, anteriormente, em replicação do DNA, ou seja, o processo que duplicaria esta molécula durante os eventos de divisão celular. Este processo consiste na quebra de uma molécula parental e na subseqüente formação de duas novas dupla hélices. Esta quebra, segundo o modelo de Watson e Crick, dar-se-ia analogamente a um "zíper"que se abre a partir de uma de suas pontas; assim, os dois filamentos desenrolados iriam expor as suas bases isoladas, como podemos verificar na figura 2.8.

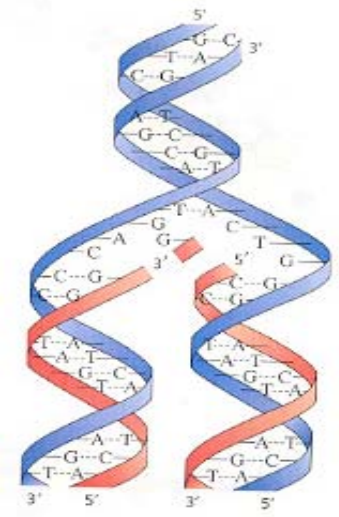

Figura 2.8: Duplicação do DNA

Cada um destes filamentos agiria como molde, onde cada base exposta (A,T,G e C) iria se parear com a sua base complementar (T,A,C e G), reconstruindo as duas dupla hélices. Acredita-se que as novas bases (A,T,G e C) que irão compor as novas hélices venham de um *"pool"* (reservatório) de nucleotídeos presentes na célula. Este tipo de replicação é denominado de replicação semi-conservativa, pois cada dupléx filho irá conter um filamento parental e outro recém sintetizado. O modelo de replicação semi-conservativa foi testado e corroborado por vários experimentos [98] realizados posteriormente.

Mas, para que possa se dar a replicação, a dupla hélice precisa girar e, subseqüentemente, desenrolar-se, já que os dois filamentos encontram-se entrelaçados [74]. Nesta etapa atuam duas enzimas: uma delas é a DNA-girase do grupo das DNA-topoisomerases, que convertem os anéis de DNA de uma forma topológica para outra, fazendo-os girar. Esta enzima provoca uma superhelicoidização do DNA, o que facilita o desenrolamento da dupla hélice. Este desenrolamento será provavelmente realizado por uma enzima chamada helicase. Com o desenrolar da dupla hélice os dois filamentos expostos estariam sujeitos a degra-

dação, se não fosse a ação de uma outra proteína, a SSB (proteína ligante de DNA), que previne este problema [74].

Os dois filamentos da dupla hélice que estão separados vão sofrer a ação da DNA-polimerase I e III, que vai sintetizar um novo filamento complementar para cada filamento parental que se separou. Um dos filamentos será sintetizado de forma contínua, enquanto o outro será sintetizado de maneira descontínua, permanecendo alguns trechos deste filamento incompletos [86]. Estas falhas serão, posteriormente, reconstituídas pela ação da DNA-ligase. Durante este processo pode ocorrer, também, alguns erros de inserção; tais erros podem ser corrigidos pela atuação da DNA-polimerase I e III, que removem as bases mal pareadas. Devemos saber que a ação da DNA-polimerase, a enzima que refaz o novo filamento, só vai se dar se existir, pelo menos, uma curta região da dupla hélice que serve como iniciador ou "primer". Nas bactérias existe uma enzima, chamada primase, que sintetiza este "primer"[105].

Durante a replicação, tem-se cerca de um erro em cada 10^9 ou 10^{10} nucleotídeos [105]. Como este número é muito pequeno; pensou-se que não seria possível tanta fidelidade de replicação dada pelo pareamento de bases, mesmo porque estudos relataram que se os erros derivassem única e exclusivamente do pareamento, a frequência de erros seria muito maior. Estes dados levaram pesquisadores [105] a desconfiar da existência de outro fator, ou fatores, que estaria agindo para diminuir os erros da replicação. A resposta a esta dúvida veio a ser esclarecida através da observação da existência de uma das ações das enzimas DNA polimerases I e III, como citado anteriormente, na sua *ação exonucleásica* de 3' para 5', retirando nucleotídeos em direção oposta àquela em que funciona a "polimerase".

Se um nucleotídeo errado é inserido na cadeia, a enzima polimerase reconhece, pois os nucleotídeos não irão formar pontes de hidrogênio, e retorna ao ponto onde ocorreu o erro "hidrolisando"o nucleotídeo errado a partir da extremidade 3'. Depois de removido o nucleotídeo, a enzima polimerase continua agindo, agora com atividade "polimerásica". Esta revisão e a capacidade de correção é muito

importante pois erros na replicação comprometem toda a espécie, enquanto que erros na transcrição (processo onde o RNA é sintetizado a partir de uma cópia de DNA) ou na tradução comprometem apenas uma proteína de determinada célula.

Em síntese, o DNA pode replicar e dar origem a novas moléculas de DNA; pode ser transcrito em RNA e este, por sua vez, traduz o código genético em proteínas. Isso é conhecido como o Dogma central da Biologia e definido por Crick, em 1956.

2.4 Código Genético

Teoricamente, toda a informação genética contida no DNA pode ser representada por quatro letras: A, T, G e C. A relação entre a seqüência de bases no DNA e a seqüência correspondente de aminoácidos, na proteína, é chamada de código genético [105]. O código genético encontra-se na forma de "*triplets*" (trinucleotídeos) que são chamados de códons: um códon é uma seqüência de três nucleotídeos que correspondem a um determinado aminoácido. Esta especificação ocorre em toda forma de vida no planeta.

Entretanto, a descoberta de que cada códon é constituído por três nucleotídeos e de como cada um deles especifica um determinado aminoácido, levou muitos anos de estudos. Esses estudos se baseavam em alguns princípios já estabelecidos, na época [105]: nas células eucarióticas, a informação genética está contida no DNA, dentro do núcleo, e a síntese de proteínas ocorre no citoplasma, dirigida pelo mRNA (RNA mensageiro). A dúvida era como a seqüência de nucleotídeos, presente no DNA e transmitida para o mRNA, era decodificada no citoplasma sob a forma de aminoácidos.

Por volta de 1955 haviam duas hipóteses:

1) George Gamow [105] propunha que a informação no DNA estaria sob a forma de código;

2) Francis Crick [105] sugeria que haveria moléculas adaptadoras que interagiriam com o mRNA e com os aminoácidos, sendo, no mínimo, 20 adaptadores diferentes, um para cada aminoácido.

Ambas estavam corretas; descobriu-se, mais tarde, que a seqüência de nucleotídeos do DNA determina a seqüência de aminoácidos de uma proteína, de acordo com um código genético universal entre os organismos vivos. Descobriu-se, também, que a molécula adaptadora proposta por Crick era o tRNA e que esse mediava a tradução do código genético em seqüências de aminoácidos [105].

O código genético parece ter surgido muito cedo e permanecido altamente conservado durante a evolução. Essa afirmativa é baseada na universalidade do código genético, isto é, com raras exceções, ele é o mesmo nos mais diversos organismos, desde as bactérias até o homem.

2.5 Mutações

O DNA de um organismo não é uma molécula estática. Freqüentemente suas bases estão expostas a agentes, naturais ou artificiais, que provocam modificações na sua estrutura e composição química. Modificações súbitas e hereditárias no material genético são denominadas mutações [105].

Todos os seres vivos sofrem um certo número de mutações como resultado de funções celulares normais ou interações aleatórias com o ambiente. Tais mutações são denominadas espontâneas; a taxa de ocorrência das mesmas é característica para um determinado organismo e constitui o chamado nível basal [105]. A ocorrência de mutações pode ser aumentada pelo tratamento com determinados compostos. Tais compostos são denominados agentes mutagênicos e as modificações que eles causam mutações induzidas. Muitos mutagênicos atuam diretamente no DNA devido a sua capacidade de atuar como uma determinada base ou de se incorporar à cadeia polinucleotídica.

Deste modo, mutações podem também ser definidas como alterações aleatórias no DNA celular [62]. Elas mudam o código genético que determina a seqüência dos aminoácidos nas proteínas, introduzindo assim erros bioquímicos de graus de severidade variáveis. As mutações têm sido classificadas como deleções (perda de bases do DNA), inserções (ganho de bases do DNA), e translocações (substituição de uma base do DNA). Dentre estes tipos de mutações destacam-se [62]:

- *desaminação*: a base nitrogenada perde um radical amino (-NH2) e liga-se erradamente com seu par.
- *perda de base*: a base é perdida durante uma duplicação do DNA; a perda mais comum é a de purinas: 5000 por célula por dia!
- *dimerização de pirimidinas*: a radiação UV (ultravioleta) provoca ligações erradas entre duas pirimidinas (geralmente T) vizinhas. Essa mutação é corrigida por uma enzima de reparo que é ativada pela luz normal, visível. Esse tipo de mutação é a causa da luz UV provocar câncer de pele; afinal, o câncer sempre origina-se de um defeito no material genético de uma célula.
- *intercalação de agentes mutagênicos*: cancerígenos no DNA
- *agentes físicos*: além da UV, também os raios X, alfa, beta, gama, raios cósmicos e feixes de neutrons são perigosos em doses grandes; os efeitos dependem do total de radiação recebida, não importando o tempo; são doses aditivas.

Qualquer base do DNA pode ser mutada. Uma mutação de ponto envolve modificação em um único par de base (substituição, adição ou deleção) e pode ser o resultado de um mau funcionamento do sistema celular que replica ou repara o DNA, inserindo uma base errada na cadeia polinucleotídica que está sendo sintetizada, ou de uma interferência química diretamente sobre uma das bases do DNA [39, 105] .

As mutações devem normalmente causar alguma modificação detectável para que a sua presença seja reconhecida. Tais modificações podem ser tão peque-

nas que são identificadas apenas por técnicas genéticas e bioquímicas especiais, ou modificações grosseiras na morfologia ou ainda modificações letais [62].

Entretanto, algumas mutações naturais destroem completamente a função de um gene. Essas mudanças mais drásticas, chamadas *mutações nulas*, incluem não somente mudanças de bases, inserções ou deleções de uma base, mas também extensas inserções e deleções e mesmo rearranjos grosseiros na estrutura cromossômica [66].

Tais mudanças podem ser causadas, por exemplo, pela inserção de um transposon (são seqüências de DNA que se movem (pulam) de um lugar para outro no genoma), o qual tipicamente transfere (insere) algumas centenas de pares de bases de um DNA estranho numa seqüência codificadora de um gene, ou por ações aberrantes do processo de recombinação celular [88].

Mesmo sob as melhores condições, haverá mutações, pois os agentes mutagênicos não podem ser completamente eliminados [62]. Existem os raios solares, banhando constantemente os seres vivos com a luz ultravioleta. Existem as radiações que emanam substâncias radioativas presentes em minúsculas quantidades, no solo, no mar e no ar. Em outras palavras, os acidentes continuarão a acontecer e as mutações a surgir. Uma ligeira, porém definida e contínua taxa de mutação, acompanhada de uma taxa nula de mudanças genéticas positivas irá, eventualmente, transformar o código genético humano em uma mensagem ilegível. O problema é como um grande "livro", escrito com uma gramática perfeita a princípio, mas com substituições aleatórias das letras introduzidas num ritmo contínuo. O "livro" ainda permanecerá legível por algum tempo, mas finalmente irá perder todo o sentido [11].

Mas as mutações não consistem apenas uma fonte de destruição. Algumas alterações podem, por pura sorte, melhor ajustar o organismo ao seu meio ambiente. Sem a mutação, todos os genes existiriam apenas em uma forma, e os organismos não seriam capazes de evoluir e de se adaptar às condições ambientais. Obviamente, uma alta freqüência de mutações desestabilizará totalmente a trans-

missão da informação genética de uma geração para outra. É desse fato que depende o curso da evolução, através da seleção natural. Assim, um século depois de Darwin ter elaborado a sua teoria da evolução, com base em laboriosas observações sobre os organismos, os cientistas dão substância a essa teoria com base em observações sobre as moléculas.

2.6 Tipos de pareamento de bases em DNA

Sabe-se que as ligações de hidrogênio são elementos fundamentais da estrutura e reatividade química, auxiliando no entendimento das estruturas e propriedades de compostos como a água, as proteínas e ácidos nucléicos como, por exemplo, o DNA. Assim, a natureza das interações físicas que contribuem para as ligações de hidrogênio tem sido objeto de inúmeras discussões na literatura [59].

As ligações de hidrogênio são a chave para muitos fenômenos, incluindo a formação e a estabilização de estruturas secundárias, flexibilidade e estabilidade de proteínas e reconhecimento molecular. Da mesma forma, reconhece-se a importância da ligação de hidrogênio para explicar uma série de propriedades físicas e químicas de vários compostos, além de efeitos conformacionais em proteínas, enzimas e ácidos nucleicos [27].

Através de estudos realizados [104], sabe-se que biologicamente as bases (A,T, G e C) em DNA não podem ser dispostas de qualquer maneira. Como vimos, o modelo padrão derivado por Watson e Crick sugere um pareamento específico. É importante notar que o pareamento específico das bases é o resultado direto da hipótese que ambas as cadeias fosfato-açúcar são helicoidais. Eles propuseram que somente certas estruturas proporcionam a combinação de ligações de hidrogênio dos átomos *aceptor (a)* e *doador (d)*, que permitem as bases Citosina (C) parear com Guanina (G) e a Timina (T) com a Adenina (A), o que assegura os dois filamentos juntos [101].

Teoricamente, as bases podem existir em várias formas, dependendo de onde os átomos de hidrogênio estão ligados. Mas presume-se que, para cada base, uma forma é muito mais provável que todas as outras [14].

Portanto, o método clássico do pareamento das bases de Watson e Crick é somente um dos vários modelos de pareamento de bases. Existem muitos outras estruturas [44, 93, 95, 97] em que as duas bases podem ser mantidas juntas por ligações de hidrogênio [38, 92]. Como por exemplo, o par de bases reverso de Watson e Crick entre Adenina e Timina (ATrWC) é formado quando um nucleotídeo rotaciona 180 graus em relação ao nucleotídeo complementar. Neste caso, as ligações glicosídicas (e colunas açúcar-fosfato) estão na orientação *trans* (a base se une ao açúcar e fosfato em configurações diferentes) em vez de *cis* (a base se une ao açúcar e fosfato na mesma configuração). Devido à simetria no potencial da ligação de hidrogênio da Timina nas posições C2-N3-C4 esta pode rotacionar no eixo N3-C6 para formar um par de base reverso ao par de base de Watson e Crick (ATrWC) [54]. Este tipo de pareamento de bases, conforme mostra a figura 2.9 é encontrado em DNA paralelo.

Outro tipo de pareamento é o par de bases de Hoogsteen entre Adenina e Timina que utiliza a face C6-N7 da purina (A) para formar a ligação de hidrogênio com a face N3-C4 da pirimidina (T) de Watson e Crick [14]. Uma característica principal do pareamento de bases de Hoogsteen [48] é que a posição N7 da purina é a base pareada, alterando a reatividade química desta posição. Um par de Hoogsteen

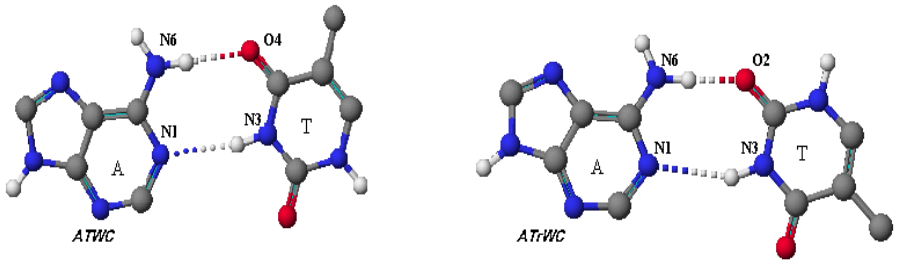

Figura 2.9: Pares de bases ATWC e ATrWC

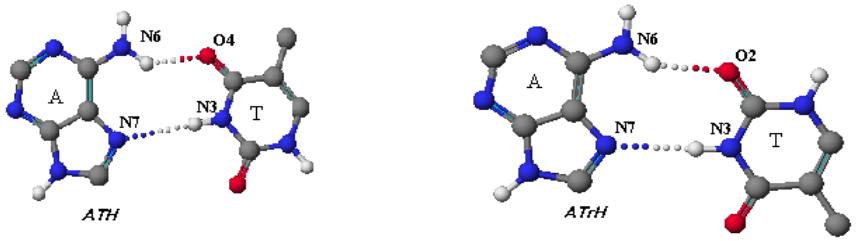

Figura 2.10: Pares de bases ATH e ATrH

reverso envolve uma volta de uma das bases 180 graus em relação a outra, como podemos observar na figura 2.10.

Como citado anteriormente, dentre os diferentes modelos de pares de bases existentes, os que ocorrem em DNA *A-T* e *G-C* são os mais importantes, existindo três estruturas para o par de bases *G-C* e quatro estruturas para o par de bases *A-T*. Além das possíveis estruturas existentes entre estas bases e definidas na literatura [47], considera-se neste trabalho os modelos que envolvem o pareamento de bases heterogêneas *A-T*, *G-C* e homogêneas *A-A*, *T-T*, *G-G* e *C-C* segundo dados da literatura [45, 47], bem como os modelos formados por estas bases, mas por apenas uma ligação de hidrogênio, representados respectivamente nas tabelas 2.1,2.2, 2.3 ,2.4, 2.5 e 2.6.

Tabela 2.1: Características geométricas de diferentes pares entre T e T

Modelos de pares de bases	*ligações simples*	*codificação*			
TT(1pa)	O4...H-N3	1	0	0	0
TT(1pb)	N3-H...O2	0	1	0	0
TT(1pc)	O2...H-N3	0	0	1	0
TT(1pd)	N3-H...O4	0	0	0	1
	ligações duplas	*codificação*			
TT(I)	O4...H-N3 e N3-H...O2	1	1	0	0
TT(II)	O4...H-N3 e N3-H...O4	1	0	0	1
TT(III)	N3-H...O2 e O2...H-N3	0	1	1	0
TT(IV)	O2...H-N3 e N3-H...O4	0	0	1	1

Tabela 2.2: Características geométricas de diferentes pares entre G e C

Modelos de pares de bases	*ligações-H simples*	*codificação*						
GC(1pa)	O6...H-N4	1	0	0	0	0	0	0
GC(1pb)	N1-H...N3	0	1	0	0	0	0	0
GC(1pc)	N2-H...O2	0	0	1	0	0	0	0
GC(1pd)	N1-H...O2	0	0	0	1	0	0	0
GC(1pe)	N2-H...N3	0	0	0	0	1	0	0
GC(1pf)	N3...H-N4	0	0	0	0	0	1	0
	ligações-H duplas e tripla	*codificação*						
GC(WC)	O6... H-N4, N1...H-N3 e N2-H...O2	1	1	1	0	0	0	0
GC(rWC)	N1-H...O2 e N2-H...N3	0	0	0	1	1	0	0
GC(II)	N3... H-N4 e N2-H...N3	0	0	0	0	1	1	0

Tabela 2.3: Características geométricas de diferentes pares entre C e C

Modelos de pares de bases	*ligações-H simples*	*codificação*			
CC (1pa)	N4-H...N3	1	0	0	0
CC(1pb)	N3...H-N4	0	1	0	0
	ligações-H duplas	*codificação*			
CC	N4-H...H-N3 e N3...H-N4	1	1	0	0

A organização destas tabelas indica os modelos dos pares de bases, os tipos de ligações de hidrogênio entre os átomos doador (d) e aceptor (a) das bases

Tabela 2.4: Características geométricas de diferentes pares entre A e T

Modelos de pares de bases	*ligações-H simples*	*codificação*			
AT(1pa)	N6-H...O4	1	0	0	0
AT(1pb)	N1...H-N3	0	1	0	0
AT(1pc)	N6-H...O2	0	0	1	0
AT(1pd)	N7...H-N3	0	0	0	1
	ligações-H duplas	*codificação*			
AT(WC)	N6-H...O4 e N1...H-N3	1	1	0	0
AT(rWC)	N1...H-N3 e N6-H...O2	0	1	1	0
AT(H)	N6-H...O4 e N7...H-N3	1	0	0	1
AT(rH)	N6-H...O2 e N7...H-N3	0	0	1	1

Tabela 2.5: Características geométricas de diferentes pares entre G e G

Modelos de pares de bases	*ligações-H simples*	*codificação*						
GG(1pa)	O6...H-N1	1	0	0	0	0	0	0
GG(1pb)	N1-H...O6	0	1	0	0	0	0	0
GG(1pc)	N2-H...N7	0	0	1	0	0	0	0
GG(1pd)	N1-H...N7	0	0	0	1	0	0	0
GG(1pe)	N2-H...O6	0	0	0	0	1	0	0
GG(1pf)	N3...H-N2	0	0	0	0	0	1	0
GG(1pg)	N2-H...N3	0	0	0	0	0	0	1
	ligações-H duplas	*codificação*						
GG(I)	O6...H-N1 e N1-H...O6	1	1	0	0	0	0	0
GG(II)	N1-H...O6 e N2-H...N7	0	1	1	0	0	0	0
GG(III)	N1-H... N7 e N2-H...O6	0	0	0	1	1	0	0
GG(IV)	N2-H... N3 e N3...H-N2	0	0	0	0	1	1	0

nitrogenadas envolvidas na ligação (onde, por exemplo na tabela 2.1, tem-se que O4...H-N3 corresponde à ligação de hidrogênio entre o oxigênio 4 da Timina 1 com o nitrogênio 3 da Timina 2) e a codificação numérica implementada no algoritmo é identificada por "0"e "1", onde o programa lê 1 para identificar uma ligação de hidrogênio existente no modelo e 0 para casos onde o tipo de ligação não ocorre no referido modelo do par de bases.

Tabela 2.6: Características geométricas de diferentes pares entre A e A

Modelos de pares de bases	*ligações-H simples*	*codificação*			
AA(1pa)	N6-H...N1	1	0	0	0
AA(1pb)	N1...H-N6	0	1	0	0
AA(1pc)	N6-H...N7	0	0	1	0
AA(1pd)	N7...H-N6	0	0	0	1
	ligações-H duplas	*codificação*			
AA(I)	N6-H...N1 e N1...H-N6	1	1	0	0
AA(II)	N1...H-N6 e N6-H...N7	0	1	1	0
AA(III)	N6-H...N7 e N7...H-N6	0	0	1	1
AA(IV)	N6-H...N1 e N7...H-N6	1	0	0	1

A implementação deste mecanismo no código computacional é bem definida. Na verificação da formação de ligação de hidrogênio entre as bases, o algoritmo "varre"uma lista dos possíveis sítios ativos entre as moléculas que permitem a formação de pontes de hidrogênio no modelo e determina, através dos critérios estabelecidos, se a ligação será aceita.

No próximo capítulo será apresentada uma abordagem rápida e objetiva sobre a técnica de simulação utilizada. São descritos os fundamentos do método e seus conceitos principais, bem como os procedimentos pertinentes à implementação numérica do algoritmo para a simulação do sistema em estudo.

3 IMPLEMENTAÇÃO DO MODELO: O ALGORITMO

3.1 Fluxograma do algoritmo

Este capítulo apresenta a implementação do algoritmo desenvolvido, descrevendo suas características, particularidades, inovações e o conjunto de estratégias acopladas para a sua formação. Todos os procedimentos aplicados são detalhados e discutidos. Para facilitar a explicação da implementação do algoritmo utilizado na simulação, apresenta-se na figura 3.1 um fluxograma deste.

3.2 Definição e evolução do sistema

O sistema utilizado para a simulação computacional de processos de formação de pares de bases biológicas, por implementação numérica do método de Monte Carlo [41], considerando princípios probabilísticos geométricos e energéticos, constitui um conjunto de programas escritos em linguagem FORTRAN e rodado em ambiente LINUX e UNIX. O programa tem como parâmetros de entrada o domínio computacional (representação do domínio físico), o número de moléculas a considerar, a natureza química de cada molécula, o iniciador do gerador de números aleatórios, a freqüência com que se deseja o cálculo dos parâmetros, o número de etapas computacionais, os tipos de ligações de hidrogênio para os vários modelos de pares de bases, os ângulos e as distâncias entre os sítios ativos e os valores dos fatores de Boltzmann de cada modelo de ligação de hidrogênio para a formação dos pares de bases, em concordância com os dados da literatura [22, 36, 91].

Em virtude da complexidade inerente ao sistema químico estudado, o modelo computacional desenvolvido apresenta algumas simplificações: as bases nitrogenadas Adenina, Timina, Guanina e Citosina são modeladas num espaço bidi-

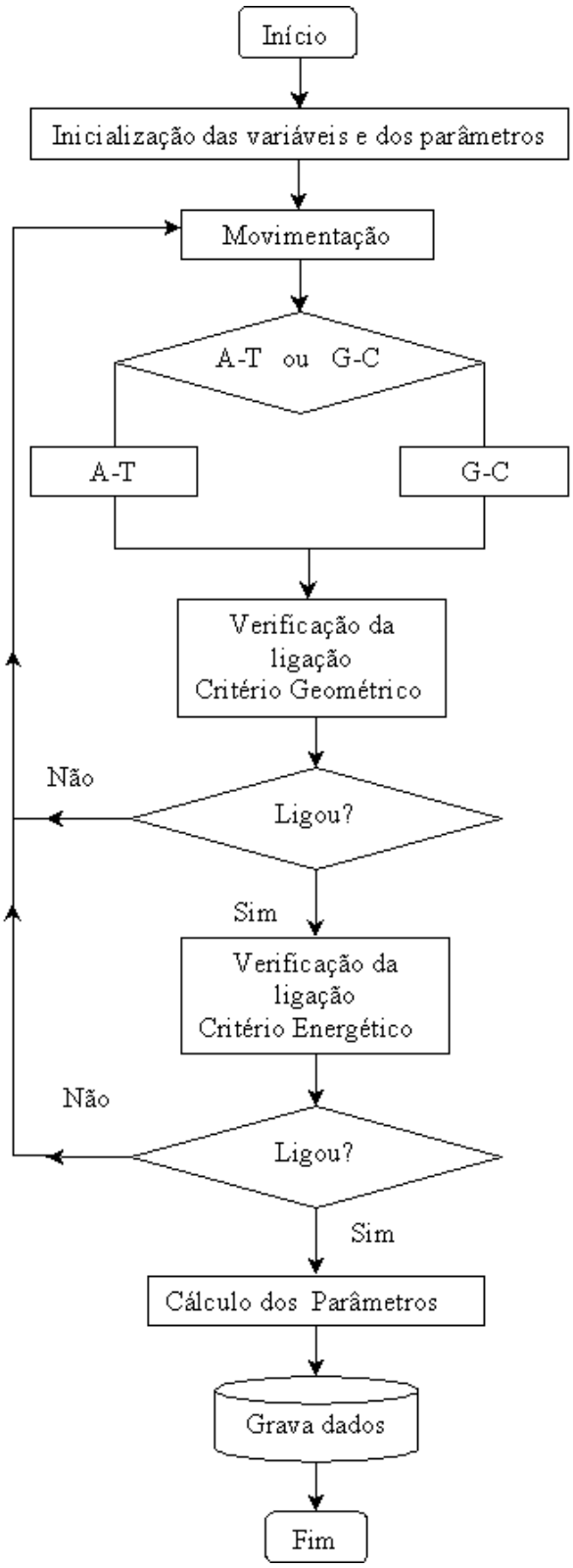

Figura 3.1: Fluxograma do algoritmo

mensional no qual cada elemento estrutural (átomo) guarda correspondência ao SRB (sistema de referência da base) [51] e não há consideração das moléculas do solvente.

A próxima etapa consiste na colocação das bases nitrogenadas (A,T,G e C) no espaço de simulação bidimensional. Nesta colocação é permitida a escolha de modo consecutivo, onde todas as moléculas de uma determinada espécie são colocadas em posições aleatórias. Este processo é feito através de um iniciador de números aleatórios que gera números pseudoaleatórios uniformemente distribuídos no intervalo [0,1]. O gerador de números pseudoaleatórios utilizado na presente simulação é uma subrotina intrínsica ao FORTRAN, *Random-number*, que pode ser inicializada por uma semente (*Random-seed*) [28]. Deste modo, é possível controlar a simulação pois, atribuindo-se um valor constante à semente, obtém-se a mesma seqüência de números pseudoaleatórios o que permite, por exemplo, repetir uma ação mantendo umas condições e alterando outras.

Durante a colocação das bases nitrogenadas no espaço de simulação, o gerador de números pseudoaleatórios fornece tanto o vetor posição de coordenadas de cada molécula, bem como sua orientação no espaço de simulação bidimensional. A colocação, bem como a posterior movimentação das moléculas, devem obedecer ao critério da não sobreposição, isto é, a proibição da dupla ocupação no espaço de simulação e respeitar a conectividade das moléculas.

Os modelos geométricos da estrutura química de cada base podem ser visualizados utilizando o software RASMOL [84]. Para simplificar a interface com programas de visualização, as coordenadas atômicas de cada base são transformadas em um arquivo no formato PDB (do inglês *Protein Data Bank, ou banco de dados de proteínas*) [82], através de subrotinas implementadas no algoritmo.

São apresentados a seguir os principais fundamentos da técnica de Monte Carlo, além da origem, definições, procedimentos, e outros aspectos importantes relacionados à aplicação do método.

3.3 O método de Monte Carlo

Nestes últimos anos, pesquisadores das áreas biológicas começaram a introduzir ferramentas computacionais preditivas em pequisas genéticas [57]. O atual grau de desenvolvimento alcançado pelas técnicas de modelagem computacional, junto ao rápido crescimento da capacidade de cálculo dos computadores, tem permitido o estudo, desenvolvimento e solução de modelos biológicos altamente sofisticados capazes de antecipar os resultados de importantes pesquisas de interesse biológico [53].

Assim, numa simulação em computador, inicia-se pelo desenvolvimento do modelo do sistema físico a estudar. Em seguida, deve-se especificar o procedimento, isto é, o algoritmo para a implementação do modelo no computador. O programa simula o sistema físico e define a experiência computacional que, para situações simples, serve como ponte entre as experiências laboratoriais e a teoria, mas para casos complicados pode ser a única forma de obter alguns resultados. É óbvio que as simulações em computador não devem ser vistas como substituto do pensar, nem do experimentar! Contribuem, no entanto, para compreender situações complexas, desde que os resultados sejam encarados com o necessário espírito crítico [7, 8].

Os métodos teóricos são utilizados para o cálculo de propriedades moleculares e são importantes na elucidação e compreensão de sistemas químicos. Dentre os métodos teóricos utilizados para descrever esses fenômenos, citaremos o método de Monte Carlo, uma vez que se trata da técnica aplicada em nossa simulação.

O método de Monte Carlo [8, 9, 34] consiste em simular um experimento com a finalidade de determinar propriedades probabilísticas de uma população, a partir de uma nova amostra aleatória dos componentes dessa população. O método, também chamado de amostragem estocástica, possui aplicações práticas bastante amplas e divide-se, basicamente, em dois tipos: os determinísticos e os probabilísticos. Nos determinísticos, [41] é desenvolvida uma teoria que tenta as-

segurar ao mesmo convergência ao resultado desejado. A maioria dos métodos de Monte Carlo determinísticos envolvem o cálculo de uma integral múltipla.

Nos métodos de Monte Carlo probabilísticos simula-se uma situação com a ajuda de geradores aleatórios e tenta-se inferir algo sobre o comportamento do sistema através destas simulações. É obvio que é possível desenvolver teorias que tentem explicar analiticamente estes modelos, mas normalmente a sua complexidade é muito grande e os modelos teóricos deixam muito a desejar [49].

O uso de métodos de Monte Carlo para modelar problemas físicos nos permite examinar sistemas mais complexos; por exemplo, resolver equações que descrevem as interações entre dois átomos é simples, mas resolver as mesmas equações para centenas e milhares de átomos é quase impossível. Com métodos de Monte Carlo, um sistema grande pode ser modelado num número de configurações aleatórias, e os dados podem ser usados para descrever o sistema como um todo [51].

Este método é considerado simples e flexível para ser aplicado em problemas de qualquer nível de complexidade [1]. Entretanto, a maior inconveniência do método é o número de simulações necessário para se reduzir o erro da solução procurada, o que tende, na prática, a torná-lo muito lento. Mas, graças ao crescente desempenho dos computadores e da sofisticação das rotinas para a geração de números pseudoaleatórios, esses métodos têm sido cada vez mais utilizados.

O termo método de Monte Carlo deve ser aqui entendido como sinônimo de um método genérico de geração aleatória do sistema, às semelhanças de alguns estudos de agregação [33, 72], não implicando necessariamente uma amostragem do espaço de fases.

Quando estamos usando números aleatórios é interessante perguntar se estes são realmente aleatórios. Isso porque quando os usamos para representar acontecimentos (eventos), precisamos estar certos de que esses também se comportam globalmente segundo a mesma distribuição de probabilidade do fenômeno físico em questão. Para responder a esta pergunta são feitos testes estatísticos com os números

gerados, de modo que se garanta o seu caráter aleatório. Contudo, tais testes são também sujeitos a dificuldades pois, estritamente falando, seriam necessários infinitos números aleatórios gerados por um mesmo processo e efetuar um número infinito de testes estatísticos [1].

Neste sentido, jamais poderemos ter números aleatórios "genuínos", mas números pseudoaleatórios ou quase-aleatórios. Os números gerados por um computador constituem uma seqüência de números calculados matematicamente por uma regra prefixada e que são aprovados em testes estatísticos de aleatoriedade. Isto porque são gerados em quantidades praticamente infinitas, antes de ser iniciada a geração da mesma seqüência. Tais seqüências de números podem ser chamadas de pseudoaleatórias e não deixam de possuir uma grande vantagem que é a de poder ser repetida desde o início e assim possibilitar uma repetição do processo computacional da simulação (quando desejado).

Deste modo, as propriedades essenciais de um gerador de números pseudoaleatórios desejado são:
- *Repetitividade*- a mesma seqüência pode ser produzida com os mesmos valores iniciais (ou sementes);
- *Randomicidade*- produzir variáveis aleatórias independentes e uniformemente distribuídas, que passam por todos os testes estatísticos de aleatoriedade;
- *Periodicidade*- uma seqüência de números pseudoaleatórios usa aritmética de precisão finita, portanto a seqüência deve ser repetida após um longo e finito período;
- *Insensibilidade a sementes*- as propriedades aleatoriedade e periodicidade não devem depender dos valores iniciais (sementes).

A seguir apresentam-se as etapas de verificação para os processos de formação de pares de bases, considerando os critérios geométrico e energético.

3.4 Critério geométrico

Um conjunto de instruções e transformações geométricas utilizadas na consideração e análise da formação das ligações de hidrogênio em pares de bases será apresentada de forma sistemática, contribuindo assim para a elucidação das técnicas e estratégias utilizadas.

3.4.1 Representação geométrica das bases (A,T,G e C) - SRB

Geralmente, a representação geométrica de objetos (figuras), consiste de um conjunto de pontos que os descrevem e a elaboração de um algoritmo específico para representá-los. Existem dois caminhos para especificar a posição de um ponto do objeto: coordenadas absolutas ou coordenadas relativas. Os sistemas de coordenadas permitem quantificar distâncias entre os referenciais dos objetos, ou seja, fornecem métricas para descrever as distâncias entre os pontos. Deste modo, a representação do objeto é realizada aplicando-se uma seqüência de transformações fundamentais: rotações, translações e escalas, sobre o conjunto de pontos deste objeto [64].

Estas operações são geralmente realizadas usando uma matriz de transformação geral que opera sobre o conjunto de dados (pontos) do objeto representado em coordenadas homogêneas. Uma matriz de transformação homogênea codifica, na forma de uma matriz 4x4, as operações geométricas necessárias para relacionar objetos de um sistema de referência à outro, no espaço tridimensional. Deste modo, nas interações entre as bases biológicas, neste trabalho, uma matriz de transformação homogênea descreverá a relação espacial via composição de matrizes de rotação e translação entre os dois referenciais locais das bases nitrogenadas envolvidas [37].

Usando composição de matrizes, obtém-se uma matriz de transformação geral que codifica as coordenadas de cada base nitrogenada em relação ao sistema de referência do universo (global). Deste modo, uma matriz de transformação ho-

mogênea possibilita relacionar os dois sistemas de referência para a verificação da formação dos pares de bases [78].

Ainda, para facilitar a interpretação física e a representação gráfica do objeto no espaço de simulação, ou seja, exibir a ocupação espacial dos elementos constituintes (geometria), opta-se por geralmente descrever os seus diversos componentes em função de um referencial local, ou seja, estabelecer a origem do sistema de coordenadas através de um ponto localizado dentro ou perto do próprio objeto. Deste modo, os demais componentes do objeto, são descritos em função do sistema de referência local. Em outras palavras, o sistema de referência do objeto, também chamado SRO, é o sistema de coordenadas local, onde se define o modelo. Este procedimento pode ser realizado através da técnica de transformações geométricas, como mencionado anteriormente [6].

As transformações geométricas são a base de inúmeras aplicações gráficas usadas para manipular um modelo, isto é, através delas é possível mover um objeto, rotacioná-lo, ou alterar o seu tamanho. Se descrevermos as transformações de pontos, então descrevemos a transformação dos objetos. Deste modo, as transformações geométricas, descrevem a ocupação espacial e a forma dos elementos constituintes (geometria). Todas estas transformações podem ser realizadas usando técnicas matemáticas que serão descritas a seguir [78].

Estas transformações são realizadas diretamente sobre os valores de coordenadas dos objetos, através da mudança de escala, rotação e translação. O conjunto de transformações é definido em forma de uma Matriz Geral, pois este tipo de armazenamento é de melhor implementação (otimização). O uso de matrizes possibilita manipulação através de índices, tornando o código fonte mais legível com a redução do número de variáveis necessárias para manipular a mesma estrutura, permitindo a construção de modelos complexos, a partir de primitivas básicas.

A principal motivação do uso da representação matricial das transformações geométricas, conforme foi mencionado, é a possibilidade de combinar as

matrizes de translação, rotação e escala em uma única matriz e posteriormente multiplicá-la pela matriz de coordenadas. A seguir, algumas das transformações usadas na representação das bases nitrogenadas serão apresentadas através de implementações práticas.

Deste modo, para a representação geométrica da estrutura química de cada uma das bases nitrogenadas (A,T,G e C) no espaço de simulação, utiliza-se a técnica de transformação geométrica. Semelhante ao mecanismo estabelecido por Donohue [24] considera-se, na representação das bases nitrogenadas, uma geometria adaptada com pentágonos e hexágonos regulares, como se mostra na figura 3.2; com lados iguais a 1.36 Å, e as ligações externas C-NH2 e C=O de 1.36 Å e 1.21 Å respectivamente, seguindo dados da literatura [20, 24].

Para as pirimidinas (C,T), as quais têm estrutura hexagonal, o SRB é definido por um sistema de coordenadas cartesiano, no centro do hexágono. A

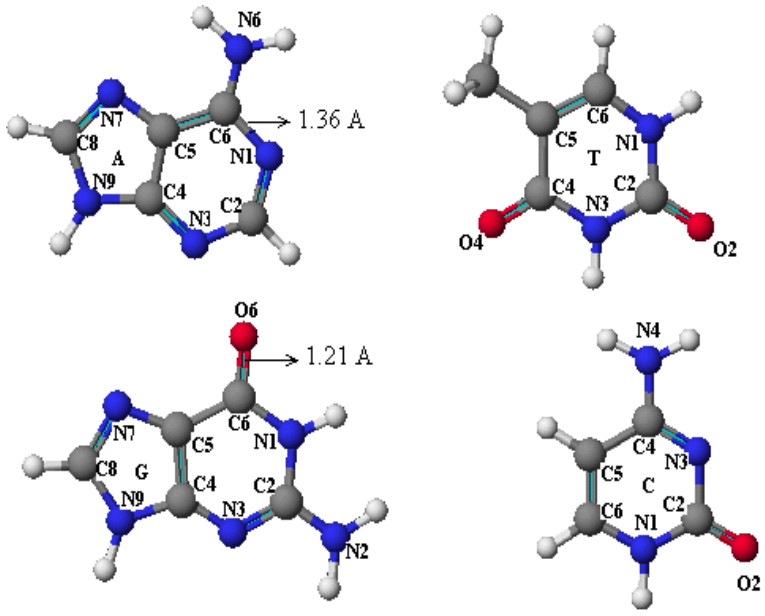

Figura 3.2: Pirimidinas: Citosina(C) e Timina(T); Purinas: Adenina(A) e Guanina (G)

partir do SRB, são descritas as posições das coordenadas dos átomos (C,N,H,O) constituintes na estrutura química da pirimidina [5]. De forma semelhante, as puri-

nas (A,G), que têm a estrutura composta por dois anéis conjugados, um pentágono e um hexágono regular, tem o SRB, onde se define o modelo, na posição do carbono 4 desta base. A partir do C4 são descritas as posições das coordenadas dos átomos (C,N,H,O) que compõem a estrutura química da purina. As coordenadas cartesianas dos átomos (C,N,H,O), em relação ao SRB de cada base (A,T,G e C), são apresentadas, respectivamente, nas tabelas 3.1 e 3.2. É importante salientar que o sistema de referência de cada base pode ser definido arbitrariamente, mas deve ser o mesmo para cada base [96].

A conectividade interna da estrutura química de cada molécula é gerenciada pelas matrizes espécies, nas quais se armazenam as coordenadas de cada um dos átomos que compõem a base, bem como a sua natureza química, conforme apresentado, respectivamente, nas tabelas 3.1 e 3.2.

Como as bases (A,T,G e C) apresentam uma estrutura química bem definida, o algoritmo usa as configurações pré-definidas, na colocação dos átomos (C,N,H e O) em torno do sistema de referência da base [18, 89].

Tabela 3.1: Coordenadas atômicas das bases Adenina e Timina no plano cartesiano

Adenina	
AN1= (2.8,0.0,0.0)	AN6= (3.1, 2.2,0.0)
AN7= (-0.2, 2.3,0.0)	AN3= (0.7,-1.2,0.0)
AC2= (2.1,-1.2,0.0)	AC4= (0.0, 0.0,0.0)
AN9= (-1.5,0.5,0.0)	AC8= (-1.6, 1.9,0.0)
AC5= (0.7,1.2,0.0)	AC6= (2.1, 1.2,0.0)
Timina	
TO4= (-1.3 , 2.3,0.0)	TN3= (-1.4, 0.0,0.0)
TO2= (-1.3 ,-2.3,0.0)	TC4= (-0.7 , 1.2,0.0)
TN1= (0.7 ,-1.2,0.0)	TC2= (-0.7 , 1.2,0.0)
TC5= (0.7 , 1.2,0.0)	TC6= (1.4, 0.0,0.0)

Especificamente, a representação geométrica da estrutura química de cada base, no espaço bidimensional, é determinada por vetores posições no sistema de referência e sua orientação é definida a partir de dois ângulos, através das pro-

priedades:

• um vetor posição das coordenadas de cada base no SRU (sistema de referência do universo ou global);

• um vetor posição das coordenadas atômicas (C,N,H,O) em relação ao sistema de referência de cada base;

• um ângulo θ (*theta*), que determina a orientação de cada base no plano;

• um ângulo $\alpha(alpha)$, que define a orientação de cada base biológica em relação à ligação glicosídica (posição do açúcar) e as faces das duas bases que estão interagindo no par de bases no plano.

Tabela 3.2: Coordenadas atômicas das bases Guanina e Citosina no plano cartesiano

Guanina	
GN1= (2.8, 0.0,0.0)	GO6= (2.8, 2.2,0.0)
GN7= (-0.2, 2.3,0.0)	GN3= (0.7,-1.2,0.0)
GC2= (2.1,-1.2,0.0)	GC4= (0.0, 0.0,0.0)
GN9= (-1.5, 0.5,0.0)	GC8= (-1.6, 1.9,0.0)
GC5= (0.7, 1.2,0.0)	GC6= (2.1, 1.2,0.0)
GN2= (3.1,-2.2,0.0)	
Citosina	
CN4= (-1.3 , 2.3,0.0)	CN3= (-1.4, 0.0,0.0)
CO2= (-1.3 ,-2.3,0.0)	CN1= (0.7 , -1.2,0.0)
CC4= (-0.7 ,1.2,0.0)	CC2= (-0.7 ,-1.2,0.0)
CC5= (0.7 , 1.2,0.0)	CC6= (1.4, 0.0,0.0)

As bases biológicas se diferenciam pelas suas composições químicas, fazendo com que a notação para cada base seja relacionada com sua própria estrutura estequiométrica [35].

3.4.2 Representação geométrica das bases (A,T,G e C) - SRU

Aqui, discute-se as transformações geométricas aplicadas ao sistema de coordenadas de definição do objeto. Como vimos anteriormente, cada base nitrogenada representada graficamente no espaço de simulação tem o seu sistema de

referência local (SRB). Mas é preciso especificar a posição de cada base em função de um referencial único, o que permite especificar o seu posicionamento relativo. Este referencial comum é denominado sistema de referência do Universo (SRU), que em nossa simulação refere-se ao espaço de simulação bidimensional, ou seja, onde se define a posição do modelo (base) no universo.

Dadas as coordenadas atômicas (C,N,H e O) em relação ao sistema de referência de cada base (SRB), a obtenção das coordenadas no SRU é bastante simples. Para tal, aplica-se a técnica de transformação de matrizes homogêneas, utilizando um conjunto de matrizes combinadas produzindo, assim, uma matriz genérica. Considerando as coordenadas atômicas em relação ao sistema de referência da base, sua representação em relação ao SRU é determinada pelo seguinte conjunto de matrizes [78]:

1. Matriz de rotação em relação ao ângulo θ de cada molécula no plano bidimensional.

$$R_\theta = \begin{bmatrix} cos\,\theta & -sen\,\theta & 0 & 0 \\ sen\,\theta & cos\,\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

2. Matriz de rotação em relação ao ângulo α que define a orientação da molécula em relação à posição do açúcar no plano;

$$R_\alpha = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\,\alpha & -sen\,\alpha & 0 \\ 0 & sen\,\alpha & cos\,\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

3. Matriz de translação que dá a posição das coordenadas de cada base em relação ao SRU.

$$T_r = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ x_0 & y_0 & z_0 & 1 \end{bmatrix} \tag{3.3}$$

onde x_0, y_0 e z_0 são as coordenadas de cada base em relação ao sistema de referência do universo (SRU).

4. Coordenadas homogêneas do átomo.

$$E = \begin{bmatrix} x_R & y_R & z_R & 1 \end{bmatrix} \tag{3.4}$$

onde x_R, y_R e z_R são as coordenadas atômicas absolutas que compõem a estrutura química de cada molécula na configuração pré-definida, ou seja, representam a posição ideal de cada átomo quando visto em relação ao sistema de referência da base (SRB).

Para criar uma matriz única que represente as coordenadas atômicas da estrutura química de cada base no SRU, basta multiplicar as matrizes correspondentes pela matriz de coordenadas homogêneas do átomo, da respectiva base, onde os componentes de cada átomo da base no SRU, podem ser apresentados na seguinte forma:

$$\begin{aligned} x_R(i) &= x_R\, cos\theta + (y_R\, cos\alpha + z_R\, sen\alpha)\, sen\theta + x_0 \\ y_R(i) &= -x_R\, sen\theta + (y_R\, cos\alpha + z_R\, sen\alpha)\, cos\theta + y_0 \\ z_R(i) &= -y_R\, sen\alpha + z_R\, cos\alpha + z_0 \end{aligned} \tag{3.5}$$

Ou ainda, de modo simplificado, pode-se expressar matematicamente o mesmo cálculo destas matrizes através de:

$$E^* = E\, R_\alpha\, R_\theta\, T_r \tag{3.6}$$

onde

E^* representa o objeto transformado, ou seja as coordenadas atômicas de cada base em relação ao SRU.

E representa o objeto não transformado, ou seja, as coordenadas atômicas absolutas de cada base em relação ao SRB.

R_α é a matriz de rotação do ângulo alpha α.

R_θ é a matriz de rotação do ângulo theta θ.

T_r a matriz de translação das coordenadas da molécula no SRU.

Para facilitar a explicação da representação geométrica das coordenadas das bases em relação ao sistema de referência do universo (SRU), a figura 3.3 ilustra as coordenadas de um átomo qualquer de uma base "i" quando medido no sistema de referência do universo SRU dadas por (P_x, P_y) e no sistema de coordenada de referência da base (SRB), dadas por (P_{xi}, P_{yi}), onde o_i é a origem do sistema da base i.

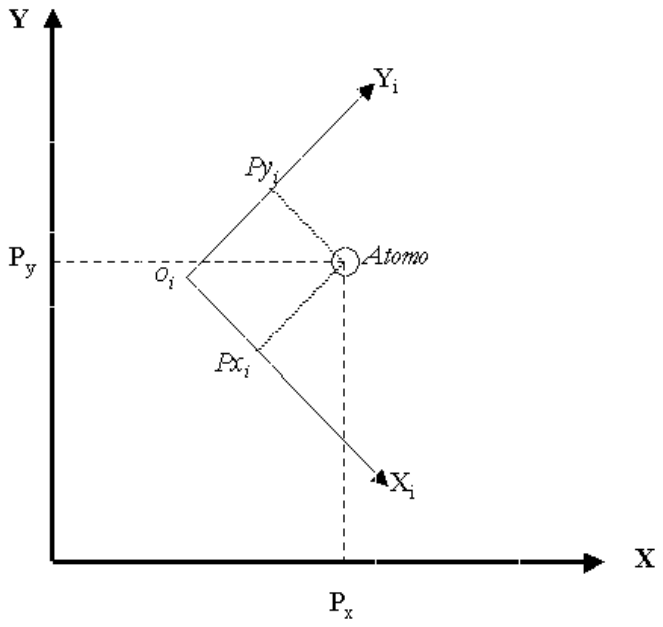

Figura 3.3: Representação das coordenadas dos átomos da base i

O uso de matrizes possibilita uma manipulação através de índices, tornando o código fonte mais legível com a redução do número de variáveis para manipular a mesma estrutura [78]. A principal motivação do uso da representação matricial

das transformações geométricas, conforme foi mencionado, é a possibilidade de combinar as matrizes numa única e posteriormente multiplicar a matriz resultante pela matriz de coordenadas. A composição de matrizes é extremamente importante para a consistência de comparações estruturais, pois proporcionam toda a informação necessária para a construção de modelos.

Tendo definido o modelo para a representação geométrica das bases e a colocação aleatória das mesmas no espaço de simulação, procede-se a evolução do sistema, ou seja, as bases são movimentadas estocasticamente. A movimentação é feita mediante um algoritmo de translação seguida de rotação das mesmas em torno do sistema de referência de cada base (SRB). Os movimentos de translação e rotação, como indicado na figura 3.4, são efetuados através de incrementos aleatórios (por soma de vetores) às coordenadas e ângulos das bases, a partir de sua posição anterior, para reposicionar a base no espaço de simulação, com o uso da rotina de gerador de números aleatórios, ou seja:

$$\begin{aligned} x_i^* &= x_i + ds_x \\ y_i^* &= y_i + ds_y \\ z_i^* &= z_i + ds_z \end{aligned} \tag{3.7}$$

sendo ds, o incremento às coordenadas da molécula (vetor de translações), definido por

$$ds = [\eta(dr) - 1, \xi(dr) - 1, 0.0]$$

onde η e ξ são números aleatórios uniformemente gerados entre [0,1] e dr é o valor máximo permitido para cada translação.

Para os incrementos no ângulo das bases, tem-se:

$$\Theta_i^* = \Theta_i + d\Theta \tag{3.8}$$

onde $d\Theta$ é a variação do ângulo da molécula no intervalo determinado.

A figura 3.4 ilustra o mecanismo de movimentação das bases no espaço bidimensional, onde T_1 e θ_1 são, respectivamente, o vetor posição (translação) e o

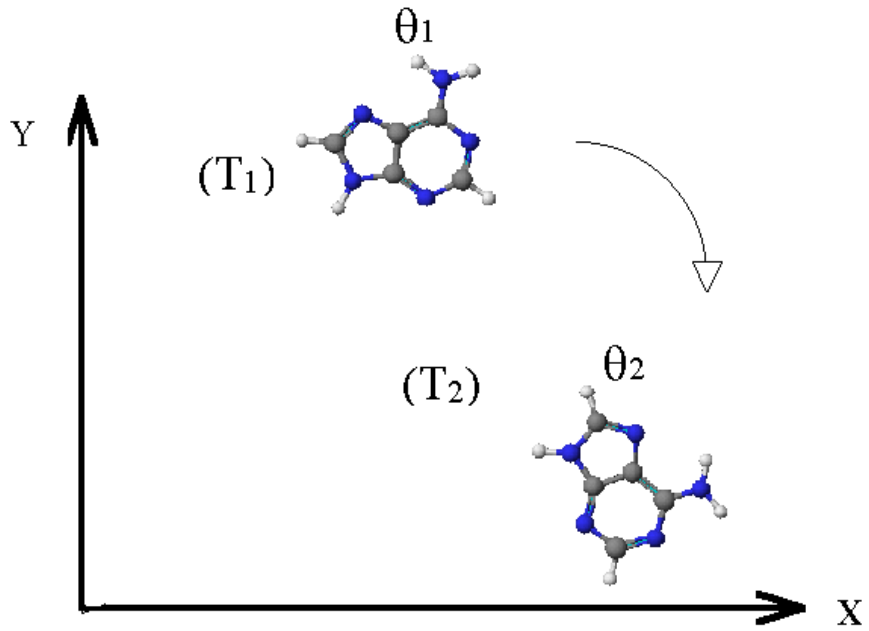

Figura 3.4: Movimentação das bases

ângulo da molécula em relação (orientação) ao sistema de referência do universo (SRU) antes da movimentação. Ainda, na mesma figura, tem-se T_2 e θ_2 que representam a nova posição e ângulo após a aplicação do algoritmo de movimentação.

As coordenadas e o ângulo de cada base são atualizados a cada movimento de translação e rotação da mesma, após o que se investiga a possibilidade de ligações de hidrogênio entre as bases movimentadas e que possibilitem a formação de um determinado modelo de par de bases. Uma vez que um par de bases se forme através das ligações de hidrogênio, o algoritmo "considera"a estrutura (par de bases) como única transladando as moléculas ligadas como uma só. De acordo com mecanismos pré-estabelecidos (distância, orientação e natureza das bases), verifica-se a qual modelo de par pertencem as ligações formadas.

Após os movimentos de translação e rotação aleatórios das moléculas no espaço de simulação, verifica-se o processo de agregação, ou seja, a formação das ligações de hidrogênio entre os sítios ativos das bases envolvidas. Este processo é realizado através da análise de critérios geométrico e energético. Inicialmente,

investiga-se a formação de ligações de hidrogênio para os pares de bases, considerando o critério geométrico. Este procedimento é feito calculando-se a distância entre os sítios ativos das moléculas que determinam as pontes de hidrogênio do modelo considerado e a orientação das bases, segundo regras previamente definidas na literatura. A análise do critério energético será discutida posteriormente.

Durante a verificação da formação das ligações de hidrogênio entre os sítios ativos das bases envolvidas, considera-se a possibilidade de duas bases estarem próximas de modo a favorecer geometricamente a ligação de apenas uma ponte de hidrogênio, como mostrado na figura 3.5.

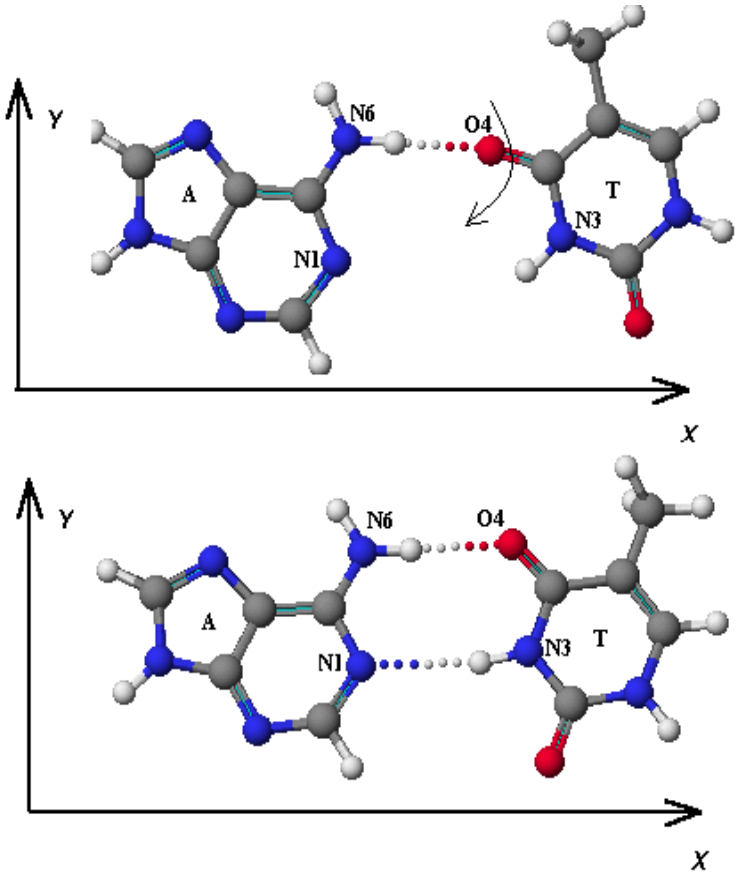

Figura 3.5: Movimentação das bases fora do sistema de referência

Deste modo, para que o algoritmo investigue, na medida do possível, todas (ou quase todas) as possibilidades de formação de ligações de hidrogênio nos

modelos de pares de bases, considera-se, além das ligações de hidrogênio duplas, ligações simples (uma ponte).

Na hipótese de formação de ligação de hidrogênio simples, a ligação é aceita, e o modelo é considerado como um par de bases com uma ligação simples, conforme nomenclatura para cada tipo de ligação simples formada entre as bases. Assim, no próximo processo iterativo, além da verificação de ligações de hidrogênio entre as bases não ligadas, o algoritmo também seleciona os pares formados por ligação simples e investiga as possibilidades de formação de ligação dupla para um determinado modelo. Este procedimento é realizado utilizando-se as técnicas de transformações geométricas, aplicando-se um algoritmo de rotação com centro específico.

Enquanto o movimento de translação das bases pode ser tratado como uma operação simples, o movimento de rotação destas em ponto arbitrário (fora do SRB), deve ser considerado em três passos. A razão é que quando as coordenadas de um átomo são multiplicadas por uma matriz de rotação esta executa uma rotação na origem do sistema de referência da base. Deste modo, devemos primeiro transladar as coordenadas do átomo da base nitrogenada para a origem do sistema de referência desta, realizar a rotação apropriada e, então, transladar o átomo rotacionado para a sua posição original.

Assim, a matriz que imprime o movimento de rotação a uma das bases do par ligado é a mencionada anteriormente; a rotação através de um ponto arbitrário sendo dada por:

1. Matriz de translação das coordenadas do átomo da base para a origem do sistema de referência.

$$T_r^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_R & -y_R & -z_R & 1 \end{bmatrix} \tag{3.9}$$

2. Matriz de rotação apropriada em torno da origem.

$$R_\gamma = \begin{bmatrix} cos\,\gamma & -sen\,\gamma & 0 & 0 \\ sen\,\gamma & cos\,\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.10}$$

3. Matriz de translação das coordenadas do átomo da base para sua posição original.

$$T_r = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ x_R & y_R & z_R & 1 \end{bmatrix} \tag{3.11}$$

Uma única matriz realizando estes três passos pode ser obtida através da multiplicação das matrizes das três transformações, com o uso da técnica de transformações homogêneas, obtendo-se as coordenadas correspondentes de cada base no SRU, cujos componentes podem ser apresentados nas formas:

$$\begin{aligned} x_0^* &= (x_0 - x_R)\ cos\gamma + (y_0 - y_R)\ sen\gamma + x_R \\ y_0^* &= -(x_0 - x_R)\ sen\gamma + (y_0 - y_R)\ cos\gamma + y_R \\ z_0^* &= z_0 \end{aligned} \tag{3.12}$$

onde esta notação matemática significa que os vetores posições x_0, y_0 e z_0 (coordenadas do centro de referência de cada molécula) no SRU são transformados para x_0^*, y_0^* e z_0^*, após a aplicação da matriz de rotação e x_R, y_R e z_R são as coordenadas absolutas de cada átomo (C,H,N,O) que compõem a estrutura química da base.

Ou ainda, de modo simplificado, pode-se expressar matematicamente o cálculo destas matrizes através de:

$$X^* = X \;\; T_r^{-1} \; R_\gamma \; T_r \tag{3.13}$$

onde

X^* e X representam as coordenadas do centro de referência da molécula após e antes de aplicar o movimento de rotação respectivamente.

T_r^{-1} é a matriz de translação das coordenadas do átomo da base para a origem do sistema de referência (SRB).

R_γ é a matriz de rotação do ângulo γ

T_r é a matriz de translação das coordenadas do átomo da base para a posição original.

Efetuada a implementação do algoritmo para o movimento de rotação com centro específico (fora do SRB), verifica-se a possibilidade de ocorrência de uma nova ligação de hidrogênio nos pares de bases formados por uma ligação de hidrogênio simples. O processo intermediário na verificação do par de bases, isto é, a possibilidade de formação de *uma* ligação de hidrogênio, é importante para se calcular a acessibilidade relativa das diferentes geometrias de formação de pares de bases, ou seja, o quão facilmente um par de moléculas consegue se posicionar formando uma dada geometria. Em seguida, prossegue-se com os mecanismos estabelecidos no desenvolvimento do algoritmo de movimentação das bases e pares no espaço de simulação. Após um número determinado de etapas computacionais, são calculados os parâmetros que descrevem a situação atual do sistema. A seguir discute-se os procedimentos usados na implementação das condições de contorno consideradas na simulação.

3.4.3 Condições de contorno periódicas

Os sistemas macroscópicos possuem um número de moléculas da ordem do número de Avogrado (10^{23}); tais números seriam impossíveis de serem representados ou simulados em nossos computadores atuais. Entretanto, é possível trabalhar

com um grande número de átomos e moléculas que pode representar razoavelmente um sistema macroscópico através de condições de contorno periódicas [10].

O uso de condições de contorno periódicas é um artifício que visa minimizar os chamados "efeitos de superfície" ou "efeitos de tamanho finito", decorrentes do fato do número de partículas utilizado nas simulações ser muito pequeno comparado ao número de partículas existentes em sistemas macroscópicos. Em sistemas macroscópicos, o número de partículas próximas à superfície é insignificante frente ao número total de partículas. Já nas simulações, este número é bastante significativo. Considerando-se o fato que as partículas próximas às superfícies possuem menor mobilidade, chega-se à conclusão que os resultados das simulações podem ser bastante diferentes do comportamento do sistema macroscópico [49].

Quando se utilizam condições de contorno periódicas, tudo se passa como se o espaço de simulação E fosse parte de um sistema infinito, formado por infinitas cópias do mesmo. Deste modo, é como se não existissem fronteiras; quando uma partícula "sai" por um dos lados do espaço E, ela emerge do lado oposto do mesmo, uma vez que é suposta a conservação da massa. Isto faz com que não haja partículas próximas às paredes, diminuindo bastante os "efeitos de superfície". É lógico que os resultados continuam sendo aproximados, uma vez que o sistema físico simulado não é infinito nem tampouco periódico [72].

Para implementar as condições de contorno periódicas é necessário, após calcular a posição (x, y, z) de cada molécula, verificar se a molécula continua nos limites do espaço de simulação ou se ela saiu do mesmo; neste último caso, deve-se fazer com que a mesma retorne ao espaço de simulação [34]. Para isto, basta executar a seguinte operação em cada uma das componentes da molécula:

$$Rx(i) = Rx(i) - BOXL\lfloor Rx(i)/BOXL \rfloor \tag{3.14}$$

onde, BOXL é uma variável contendo o comprimento L da caixa. Declarações semelhantes são aplicadas para as coordenadas y e z. A operação na equação 3.14, é feita através da função *anint(x)* implementada no FORTRAN, que retorna o inteiro mais

próximo de "x"convertendo o resultado para o tipo REAL; desta maneira anint(-0.49) tem o valor 0, uma vez que anint(-0.51) é -1. Usando estes métodos, sempre estamos utilizando as coordenadas das N moléculas que se encontram corretamente na caixa central.

Além disso, sempre que for necessário calcular a distância entre duas moléculas $p_1 = (x_1, y_1, z_1)$ e $p_2 = (x_2, y_2, z_2)$, é necessário calcular a *distância mínima* entre as duas, ou seja, é necessário saber qual das cópias de p_1 está mais próxima de p_2 [1]. Para isto, após calcular as componentes $dx = x_1 - x_2$, $dy = y_1 - y_2$ e $dz = z_1 - z_2$, executa-se a seguinte operação [1]:

$$\begin{aligned} Rx(i,j) &= Rx(i,j) - BOXL\lfloor Rx(i,j)/BOXL \rfloor \\ Ry(i,j) &= Ry(i,j) - BOXL\lfloor Ry(i,j)/BOXL \rfloor \\ Rz(i,j) &= Rz(i,j) - BOXL\lfloor Rz(i,j)/BOXL \rfloor \end{aligned} \tag{3.15}$$

O código produz o vetor imagem mínima, não importando o tamanho da caixa.

Em suma, a idéia central é trabalhar com um grande número de réplicas que estão umas em contato com as outras através de suas fronteiras. Em cada compartimento réplica é simulada exatamente a mesma situação; desta forma, apenas um cálculo na caixa central (principal) é necessário, mas a presença das caixas vizinhas é levada em conta através das interações com todas as caixas, simulando-se de fato um grande contínuo de sistemas idênticos ao mesmo tempo. Se uma molécula sai de uma das caixas, automaticamente entra na caixa vizinha, o que corresponde a sua entrada no lado oposto da caixa central (principal), mantendo-se o número de moléculas constante.

Após a análise dos processos de formação de ligações de hidrogênio em pares de bases com o critério geométrico, procede-se a etapa da verificação considerando o critério energético, que é de suma importância nas propriedades moleculares.

[1] onde a operação $\lfloor\ \rfloor$ é feita através da intrínsica anint.

3.5 Critério energético

Sabe-se que a geometria molecular, no limite a baixas temperaturas, é governada pela energia potencial, ou seja, uma molécula se ajusta rapidamente à geometria que dá a energia potencial mais baixa [87]. Neste sentido, indica-se a importância das considerações energéticas como segundo critério na verificação da formação das ligações de hidrogênio entre as bases nitrogenadas, aborda-se a metodologia utilizada na modelagem molecular, bem como o método e programa utilizados na implementação dos cálculos energéticos considerados neste trabalho.

A modelagem molecular é uma ferramenta que pode ser utilizada para propor diferentes estruturas e métodos de otimização de geometria para encontrar a estrutura mais estável, como por exemplo, as interações entre anéis aromáticos e calcular as diferentes energias para cada estrutura proposta [80]. Inicialmente utiliza-se um programa computacional para gerar estruturas químicas. Após este processo, deve-se proceder a otimização da geometria, que geralmente é feita por métodos acoplados ao próprio programa computacional.

Deste modo, partindo-se de modelos de ligações de hidrogênio entre as bases nitrogenadas A,T,G e C definidos e conhecidos na literatura, bem como das diferentes estruturas que estes podem formar, realiza-se a otimização e cálculo das energias destes diferentes modelos de pares de bases através da aproximação semi-empírica AM1 (Austin Model 1) implementada no programa CAChe [13]. Os resultados energéticos obtidos serão utilizados na determinação do fator de probabilidade de Boltzmann, que será aplicado na análise do critério energético. Os valores de energia para as moléculas otimizadas pelo método semi-empírico AM1 podem ser comparados com dados encontrados na literatura [46, 76].

É importante salientar que existem inúmeros trabalhos que apresentam estudos relativos às energias de estabilização para os vários modelos de par de bases bem como vários métodos de otimização destas geometrias. Podia-se selecionar dados destes trabalhos e utilizá-los no cálculo do fator de probabilidade de Boltzmann.

O inconveniente é que todos estes apresentam os valores de energia para os pares de bases considerando o número máximo de ligações de hidrogênio que cada modelo pode formar; não há dados energéticos para modelos de par de bases considerando apenas uma ligação de hidrogênio. Desta maneira, para que se utilize valores energéticos oriundos de um mesmo procedimento, nos cálculos do fator de Boltzmann, computa-se os valores de energia de estabilização para ambas as classes de modelos: par de bases com ligações de hidrogênio simples e múltiplas.

Em suma, utilizando-se o programa computacional CAChe, calculamos os valores de energia para a formação de cada modelo de par de bases com o número máximo de ligações de hidrogênio e também para aqueles modelos formados por apenas uma ligação de hidrogênio. O procedimento para a implementação e análise do critério energético para a formação do par de bases bem como os conceitos principais de otimização são descritos a seguir.

3.5.1 Otimização da geometria molecular

Antes de descrever o procedimento para a otimização dos pares de bases, veremos algumas definições e fundamentos de otimização. Literalmente, otimização corresponde a tornar algo "tão perfeito, efetivo ou funcional quanto possível". Desta forma, podemos definir otimização como sendo um processo baseado em instruções que permitam obter o melhor resultado de uma dada situação.

Segundo Mundim [68], a otimização da geometria é uma técnica que visa encontrar um conjunto de coordenadas que minimizam a energia potencial do sistema de interesse. O procedimento básico consiste em caminhar sobre a superfície potencial na direção em que a energia decresce, de maneira que o sistema é levado a um mínimo de energia local próximo. Geralmente a configuração final, após este processo, não difere em muito da inicial. A minimização da energia cobre, portanto, somente pequena parte do espaço de configurações. Porém, pelos ajustes nas posições atômicas, ela relaxa as distorções nas ligações químicas, nos ângulos en-

tre ligações e nas interações de van der Walls [10]. A minimização de energia, ou otimização da geometria, é um processo iterativo.

Para encontrar as conformações estáveis (mínimos de energia) de uma molécula, pode-se usar a mecânica molecular clássica e a mecânica quântica semi-empírica ou pode-se combinar ambos os métodos de otimização. Os métodos mecânico-clássicos são mais flexíveis que os métodos mecânico-quânticos [82]. Entretanto, se a busca envolve a formação e a quebra de ligações, com exceção das ligações de hidrogênio que são tratadas automaticamente por mecânica molecular, então um método quântico pode ser necessário [13].

A análise conformacional consiste na exploração de arranjos espaciais (formas) energeticamente favoráveis de uma molécula (conformações). Na análise utiliza-se mecânica molecular, dinâmica molecular, cálculos químico-quânticos ou análise de dados estruturais determinados experimentalmente por RMN ou cristalografia de raios-X, por exemplo [68]. Os métodos de mecânica molecular e químico-quânticos são empregados para calcular as energias conformacionais, enquanto os métodos de busca sistemática e aleatória, de Monte Carlo, de dinâmica molecular e de geometria de distâncias (freqüentemente combinados com procedimentos de minimização de energia) são usados para explorar o espaço conformacional [61].

Todos os métodos gradiente de otimização de geometria clássicos e mecânico-quânticos encontram uma conformação de energia mínima próximo da geometria inicial [10]. Este mínimo não é necessariamente o verdadeiro mínimo de energia global da estrutura. Portanto, às vezes, é necessário comparar muitas conformações possíveis de uma molécula para encontrar um mínimo global verdadeiro, ou vários mínimos mais baixos. Este processo é conhecido como pesquisa conformacional e tem sido desenvolvido uma variedade de métodos para tratar situações diferentes.

Neste sentido, para otimizar uma molécula, o programa CAChe ajusta sistematicamente as coordenadas dos átomos na molécula, até que a conformação

de menor energia seja encontrada. Deste modo, os procedimentos disponíveis para otimização com CAChe são:

- uso da Mecânica computacional, que através da geometria de uma dada molécula, aplica cálculos da mecânica clássica para reduzir os desvios da estrutura, em relação aos valores em mecânica clássica ideal;
- uso da aplicação computacional MOPAC, que usa a presença e as posições dos elétrons entre os átomos para calcular a energia mínima de uma estrutura baseada na equação de Schrödinger;
- sobrepondo uma molécula otimizada sobre a outra para ver as diferenças estruturais entre elas.

3.5.2 Modelagem molecular usando o software CAChe

CAChe é uma ferramenta de modelagem molecular auxiliado por computador, que opera sobre sistemas Microsoft Windows ME, Microsoft Windows 98 ou Microsoft Windows NT 4.0 [13]. O planejamento molecular auxiliado por computador (CAMD, do Inglês *Computer Assisted molecular Design*), consiste na investigação das estruturas e propriedades moleculares usando a química computacional e técnicas de visualização gráfica. Este programa é suficientemente flexível e fornece subsídios para desenhar, modelar e realizar cálculos de uma molécula. Além disso, apresenta uma interface gráfica para construção e manipulação de estruturas moleculares, bem como de módulos de cálculo de propriedades moleculares em diversos níveis de aproximação. São disponíveis métodos clássicos como Mecânica Molecular e Simulações por Dinâmica Molecular e Monte Carlo, além de métodos quânticos *ab initio* e semi-empíricos.

Deste modo, usando a química auxiliada por computador (CAChe), pode-se medir "precisamente" a geometria de uma estrutura bem como locar as distâncias dos átomos e ângulos de ligação para cada valor desejado. Neste, os cálculos exploram as características do exemplo químico, predizendo propriedades

tais como a distribuição de elétrons e gerando conformações que possuem energia potencial ou valores de calor de formação baixos. Os cálculos sobre a molécula são realizados por aplicações computacionais, as quais usam equações da mecânica clássica e quântica. Os resultados experimentais obtidos em CAChe podem ser exibidos por vários caminhos, tais como:

• Movimento dos átomos da molécula e ligações produzindo uma estrutura otimizada ou de baixa energia;

• Amostra das propriedades eletrônicas como superfícies sobrepostas sobre uma molécula;

• Construção de gráficos de energia tridimensionais vistas ao longo de uma série de conformações de baixa energia;

• Análise dos dados experimentais como uma extensão de valores contidos num arquivo de dados ou arquivo de saída, gerados automaticamente por cada experimento.

Além disso, CAChe oferece muitas opções diferentes de exibir os átomos e ligações para que se possa ver facilmente a aparência da molécula inteira, ou porções selecionadas, sem alterar a estrutura molecular. Deste modo, pode-se exibir a molécula como o desenho em linha simples, ou esferas e cilindros tridimensionais, ou como uma combinação de muitos estilos diferentes de modelagem. Uma outra propriedade que o CAChe contempla é a possibilidade de usar modelos de estrutura molecular dos exemplos contidos no Fragment Library (Biblioteca) que acompanham o programa. Também inclui, por exemplo, o planejamento de sínteses, a pesquisa de banco de dados e a manipulação de bibliotecas combinatoriais.

Para estudar sistemas de um grande número de moléculas ou átomos, como em estruturas de macromoléculas biológicas, a química computacional pode fazer uso da chamada mecânica molecular, MM. Esses casos, em geral apresentam

conformações favorecidas energeticamente, e os resultados destes cálculos podem auxiliar na compreensão de problemas complexos [87].

Segundo Allinger [2], é importante notar que os cálculos de MM são feitos em um "modelo molecular". Esse modelo possui propriedades, as quais reproduzem fatos experimentais, que não correspondem a uma reprodução fiel da molécula em estudo. Somente significa que a informação, em particular, que foi usada para desenvolver o modelo, é reproduzida pelo modelo. Um campo de força molecular descreve uma energia potencial de uma molécula em relação à energia de uma dada geometria de referência. O campo de força contém parâmetros de constantes de força que são obtidos indutivamente de uma comparação sistemática entre propriedades moleculares observadas e calculadas.

Ao realizarmos cálculos de MM estaremos tentando predizer propriedades de moléculas e de sistemas moleculares, tais como, calor de formação, energias de confôrmeros, barreiras de rotação, geometria de sistemas no estado fundamental, geometria de moléculas em cristais e geometria do estado de transição, entre outros. Deste modo, a energia mais baixa ou estrutura ótima da conformação de uma molécula exibe características e propriedades que mais parecem refletir o verdadeiro comportamento de um exemplo químico.

Neste sentido, utiliza-se aqui o programa computacional CAChe [85], para gerar estruturas dos modelos de ligações de hidrogênio entre as bases nitrogenadas (purinas e pirimidinas) que são a base da estrutura do DNA. Após este processo, utiliza-se o método especificado para o cálculo de otimização da geometria.

3.5.3 Procedimentos para Modelagem usando CAChe

Para fazer modelagem molecular precisa-se gerar uma estrutura de um exemplo químico, isto pode ser feito no CAChe Editor [13]. Após desenhar a estrutura de cada uma das bases biológicas Adenina, Timina, Guanina e Citosina,

usando as ferramentas do CAChe, muitos atributos sobre as propriedades destas estruturas podem ser analisados. Por exemplo, ângulos e comprimentos de ligação, características espectrais e atributos termodinâmicos, os quais podem ser obtidos usando diferentes aplicações.

Para que se tenha uma estrutura correta para cada base é preciso fazer uso de um programa que "*conheça*"as regras de ligações. O Programa CAChe possui ferramentas práticas que fornecem uma estrutura final correta. Assim, ele aperfeiçoa a estrutura da molécula com o uso do menu *Beautify*. Basicamente, o que *Beautify* faz é dar ângulos e comprimentos de ligações corretos para a estrutura desenhada.

Selecionando *Beautify* e a opção *Comprehensive* tem-se a estrutura em ligação convencional satisfazendo as regras de valência e hibridização dos objetos selecionados e corrigindo a estrutura do anel e a geometria. Ainda usando as ferramentas do CAChe é possível rotacionar cada uma das estruturas desenhadas para se ter uma idéia das diferentes orientações que elas podem ter no espaço e das partes da molécula que estão interagindo através das ligações de hidrogênio.

Com o programa CAChe, pode-se ver uma animação de ligações de hidrogênio; o caminho que a animação ilustra é a formação e a quebra das ligações de hidrogênio na formação do par de bases, mas não se deve esquecer que esta é uma interação fraca, ela vai e volta e só se torna estável considerando um grande número de ligações de hidrogênio na estrutura, como por exemplo em DNA.

Para criar as ligações de hidrogênio nos pares de bases precisamos inicialmente aproximar as duas bases que formarão o par. Após alinhar as moléculas adequadamente, ajusta-se as distâncias e as orientações das ligações de hidrogênio do respectivo modelo, usando as ferramentas do CAChe. Em suma, os passos utilizados na construção dos pares de bases são:

- Seleciona-se os tipos de bases (A,T,G,C) que formarão o modelo;

- Posiciona-se cada base na orientação adequada para a formação do par de base desejado;
- Calcula-se a distância e orientação dos átomos doador e receptor participantes da ligação de hidrogênio;
- Conecta-se os átomos que estarão envolvidos na ligação de hidrogênio;
- Cria-se as combinações dos pares de bases através da conexão dos átomos envolvidos na ligação de hidrogênio;

Após fazer cada combinação dos modelos de pares de bases, ajusta-se a geometria de cada modelo através do comando *Beautify* do CAChe, corrigindo a valência, hibridização, geometria e as estruturas de cada anel aromático. Assim, CAChe satisfaz as regras de valência e hibridização das moléculas em cada passo, através de alterações nos ângulos e comprimentos das ligações de hidrogênio do par de bases para valores quimicamente confiáveis (de acordo com valores ideais da mecânica clássica). Além disso, CAChe força as estruturas cíclicas dentro do anel, analisando a estrutura planar ou outra conformação, ajustando comprimentos de ligação para valores apropriados.

Através da Mecânica molecular pode-se determinar o calor de formação das bases e pares de bases nitrogenadas ou a estabilidade relativa de diferentes moléculas. Dado um certo conjunto de parâmetros existentes no programa computacional, pode-se efetuar um cálculo para determinar diferentes atributos termodinâmicos das moléculas.

Deste modo, usando a ferramenta computacional *Project Leader*, pode-se determinar o calor relativo das bases e seus pares. Este componente do CAChe simplifica a aplicação computacional, necessitando especificar primeiro a propriedade do experimento químico que se deseja avaliar e, então, escolhendo o procedimento para análise daquela propriedade. Além disso, o *Project Leader* oferece a habilidade para realizar cálculos sobre vários exemplos químicos ao mesmo tempo. Os procedimentos no cálculo do calor de formação das bases e pares de bases usando CAChe

são apresentados em uma tabela que exibe em cada coluna: os modelos dos exemplos químicos, as propriedades moleculares investigadas, os procedimentos selecionados e os resultados obtidos. Assumindo que o calor de formação de uma reação é a diferença entre o calor de formação dos produtos e o dos reagentes, a energia de estabilização para cada modelo de par de bases pode ser obtida através da seguinte equação:

$$\Delta E^{A...B} = E^{A...B} - (E^A + E^B) \tag{3.16}$$

Assim, verifica-se, na equação (3.16), que a energia de estabilização de cada par de bases expressa por $A...B$ ($\Delta E^{A...B}$) é determinada pela diferença na energia do par ($E^{A...B}$) e a soma das energias das bases isoladas (E^A, E^B). As energias $E^{A...B}$, E^A, e E^B foram calculadas com o método semi-empírico AM1 disponível no programa computacional CAChe. Deste modo, a "*reação*" é a *formação das ligações de hidrogênio*. Se a formação da ligação de hidrogênio é energeticamente favorável, o ΔE de estabilização será negativo, senão será positivo.

É importante salientar que, para os modelos com ligações de hidrogênio simples, a energia de estabilização é calculada de acordo com cada geometria que esta têm no par. Assim, verifica-se que alguns modelos de ligações de hidrogênio simples apresentam pequenas variações nos valores de energia de estabilização dependendo do tipo de par de bases em que se encontra. Esta pequena variação é devido à mudança na geometria quando o par de bases é formado. Como não existe uma regra ditando se a acessibilidade geométrica favorecerá um modelo de ligação dupla ou outro, compara-se os valores da energia de estabilização de cada ligação simples comum a dois modelos e seleciona-se qual tem menor energia (maior fator de probabilidade), uma vez que é o mais provável. Uma vez definida a forma de determinação da energia de estabilização para a formação do par de bases, dada pela equação 3.16, pode-se empregá-la no cálculo das energias de estabilização para todos os modelos de par de bases existentes e possíveis entre as bases nitrogenadas. Posteriormente, estes valores de energia serão usados na análise do fator de probabilidade de Boltzmann, os quais são de suma importância na análise da formação do par de bases com o critério energético.

3.6 Verificação das ligações de hidrogênio considerando o critério geométrico

Após a implementação do processo estocástico na evolução do sistema e estando as moléculas na nova posição, investiga-se a presença de sítios ativos (átomos doador e aceptor) entre as vizinhanças destas moléculas, que possibilitem a formação de ligações de hidrogênio para os modelos de pares considerados.

Desta maneira, como utilizado por Ippolito et al. [50], para o critério geométrico, a formação das ligações de hidrogênio é definida pelo cálculo da distância entre os átomos aceptor e doador (participantes da ligação), a orientação das moléculas (posição geométrica adequada dos sítios ativos) e a natureza das bases nitrogenadas.

Cabe salientar aqui que, num primeiro passo, o algoritmo faz a verificação da formação das ligações entre as bases nitrogenadas pelo critério geométrico. Isto significa que ele não diferencia entre ligações de hidrogênio fortes e fracas. Felizmente, existem relações simples entre as propriedades geométricas e energéticas nas ligações de hidrogênio. Assim, uma ligação linear com uma distância curta entre os átomos doador e aceptor é certamente mais forte que uma não-linear para um valor maior desta distância. No entanto, com este critério é possível realizar uma seleção energética "razoável", através da variação dos limites geométricos como, por exemplo, a variação da distância de ligação de hidrogênio (Varmax no caso).

Especificamente para identificar as ligações, o algoritmo investiga a distância entre os pares de átomos doador (d) e aceptor(a) das moléculas, que satisfazem o critério geométrico para a formação de ligação de hidrogênio.

Se for constatada a presença de sítios ativos que, de acordo com os critérios estabelecidos, formam ligações de hidrogênio, é definida a reação (ligação) respeitando as seguintes considerações:

- o número de bases nitrogenadas (A, T, G e C) envolvidas é modificado;

• os sítios ativos do par formado são "congelados", isto é, sem envolver um átomo aceptor ou doador em ligações mais que uma vez;
• a conectividade do par formado deve levar em consideração o modelo;
• o par formado por uma ligação simples é investigado para a possibilidade de formação de uma ligação dupla (ou tripla no caso de GCWC) .

As reações (formações de ligações) que podem ocorrer, de acordo com os modelos considerados, segundo dados da literatura [24] são:
a)Adenina/Adenina
b)Adenina/Timina
c)Timina/Timina
d)Guanina/Citosina
e)Guanina/Guanina
f)Citosina/Citosina

Considera-se aqui, como inovação, e etapa intermediária aos processos de formação de pares, a análise das características geométricas e energéticas para modelos formados apenas por ligações de hidrogênio simples (1 ponte de hidrogênio). Esta consideração é importante para se calcular a acessibilidade relativa das diferentes geometrias de formação dos pares, ou seja, o quão facilmente um par de moléculas consegue se posicionar formando uma dada geometria. Por exemplo, se entre as bases Adenina e Timina existe uma ligação de hidrogênio simples N6-H....O4, este par pode vir a formar um modelo de ligação de hidrogênio de Watson e Crick ou o modelo de Hoogsteen, pois ambos possuem este tipo de ligação de hidrogênio em sua formação.

3.7 Verificação das ligações de hidrogênio considerando o critério energético

Por volta de 1953, Metropolis e colaboradores [65] introduziram um algoritmo simples para simular a evolução de um sólido em um banho em equilíbrio

térmico. O algoritmo introduzido por estes autores é baseado em técnicas de Monte Carlo e gera uma seqüência de estados do sólido como descrito a seguir.

Considera-se um sólido que se encontra num estado i, com energia Ei. Então o próximo estado j é gerado aplicando-se um mecanismo de pertubação, que transforma o estado corrente num próximo por uma pequena distorção. Deste modo, a energia do próximo estado é Ej. Assim, se a diferença de energia $Ej - Ei$ for menor ou igual a zero, o estado j é aceito como estado corrente. Se a diferença de energia for maior que zero, o estado j será aceito com uma certa probabilidade que é dado por

$$exp\left(\frac{E_i - E_j}{K_B T}\right) \tag{3.17}$$

onde T denota a temperatura do banho quente e k_B é uma constante física conhecida como *constante de Boltzmann.* A regra de aceitação descrita é conhecida como *critério de Metropolis* e o algoritmo que o usa como *Algoritmo de Metropolis* [65]. O equilíbrio térmico é caracterizado pela distribuição de Boltzmann. Esta distribuição dá a probabilidade de sólidos estando no estado i com energia Ei na temperatura T, e é dado por

$$P_T\{X = i\} = \frac{1}{Z(T)}\, exp\left(\frac{-E_i}{K_B T}\right) \tag{3.18}$$

onde X é uma variável estocástica denotando o estado corrente do sólido. $Z(T)$ é a função de partição que é definida por

$$Z(T) = \sum_j exp\,(-E(j)/k_B T) \tag{3.19}$$

onde o somatório abrange todos os estados possíveis. Na equação 3.19 o termo exponencial $(exp\,(-E(j)/K_B T))$ é conhecido como o fator de Boltzmann. Note que a soma é feita para todos os estados possíveis do sistema.

Para tornar mais clara esta explicação, podemos observar, na figura 3.6, o que foi exemplificado acima, onde a probabilidade associada a um estado decresce com a energia, ou seja, a distribuição de Boltzmann mostra que a população de estados decresce com a energia. Deste modo, a probabilidade *p(j)* que um sistema esteja em um estado *j* depende exponencialmente da energia do estado.

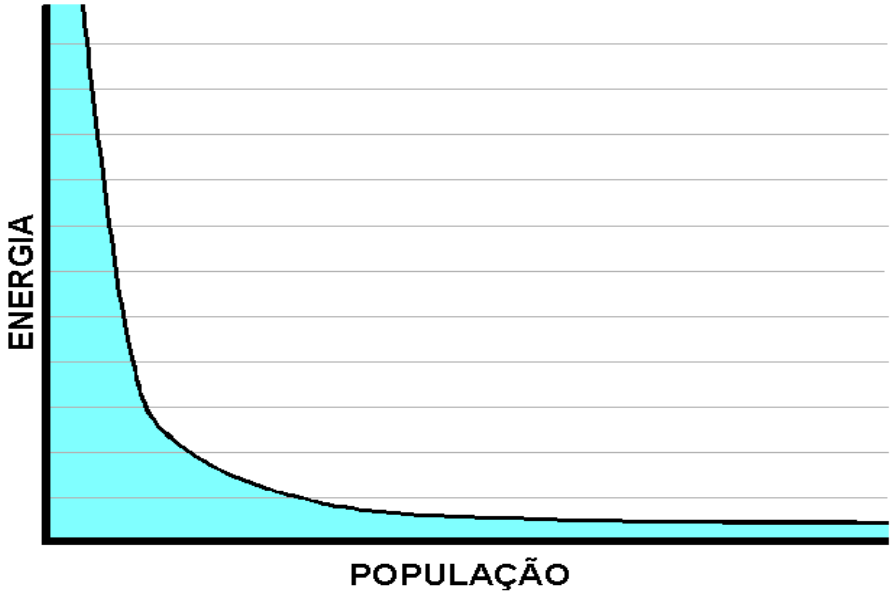

Figura 3.6: Distribuição de Boltzmann [63]

Assim, a verificação da formação de ligações de hidrogênio nos pares de bases, segundo o critério energético, é baseada no fator de probabilidade de Boltzmann. Neste sentido, de posse dos valores das energias de interações para todos os modelos de pares de bases, calculados com o programa computacional CAChe e o método AM1, utiliza-se a função distribuição de Boltzmann, como descrita anteriormente, para calcular o fator de probabilidade de ligação para cada modelo de par de bases.

É importante salientar que, para fins de interpretação, assumimos que o sistema (par de bases) está em equilíbrio térmico com uma temperatura T e uma dada energia *Ei*, ou seja, já se encontra num estado de energia apropriado. Também

destaca-se que inicialmente considera-se somente a probabilidade energética na formação dos pares de bases, sem levar em conta os níveis energéticos.

Neste sentido, supomos que o algoritmo encontre-se num estado onde a análise do critério geométrico, para a verificação da formação do par de bases, tenha sido efetuada. Isto significa que todos os detalhes envolvidos no processo de formação do par de bases considerando o critério geométrico são especificados (e conhecidos dentro do programa computacional). Uma vez considerando um estado de menor energia, calcula-se o fator de probabilidade energético para cada modelo de par de bases. Assim, o algoritmo deve agora decidir entre duas alternativas para a consideração do critério energético. O procedimento é feito através de um processo de decisão, que é ilustrado na figura 3.7.

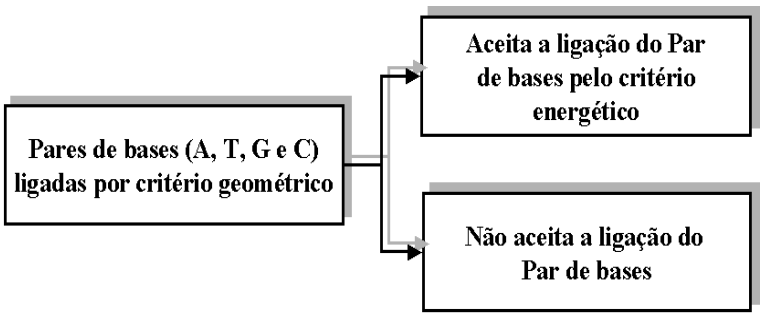

Figura 3.7: Processo de decisão

Assim, de acordo com a figura 3.7, considera-se um par de bases (A...B) formado apenas pelo critério geométrico. A probabilidade p_{AB} deste par de bases (A...B) formado pelo critério geométrico ser formado também pelo critério energético é então avaliada (isto pode ser feito uma vez que todos os detalhes que permitiram a formação do par de bases pelo critério geométrico são conhecidos). Deste modo, o processo de decisão consiste nos seguintes passos:

- Um número aleatório (R) é selecionado (usando subrotinas padrões do computador: *random-number*) com R no intervalo [0,1].

- Se $R < \ p_{AB}$, o algoritmo aceita a ligação de hidrogênio do par de bases também pelo critério energético;

- Se $R > \ p_{AB}$, o algoritmo não aceita a ligação do par de bases pelo critério energético, ou seja, energeticamente a probabilidade deste par de bases ocorrer é nula e as bases são "soltas"caminhando para um novo processo de verificação de ligação através dos critérios estabelecidos.

Aqui p_{AB} é o fator de probabilidade de ligação associado a cada modelo de par de bases A-B. Deste modo, a probabilidade p_{AB} de que um determinado par de bases esteja em um estado AB depende exponencialmente da energia do estado do par de bases, ou seja,

$$p_{AB} \propto \exp\ (-E(AB)/k_B T)$$

Os dois critérios, anteriormente descritos, geométrico e energético, são importantes para analisar os fatores de natureza essencialmente física que intervém nos processos relacionados à formação do par de bases. Em relação ao primeiro critério, deduz-se que um bom par de bases é aquele que satisfaz a natureza dos efeitos de distância e de orientação adequada para ocorrência das ligações de hidrogênio. O segundo critério salienta que deve existir sempre alguma afinidade química, como proximidade de grupos reativos, na formação das ligações, caso contrário não ocorrerá formação de uma reação significativa. Há, portanto, necessidade da consideração de ambos os critérios na análise da formação dos pares de bases analisados.

A seguir, apresenta-se os resultados obtidos na simulação computacional dos processos de formação de ligações de hidrogênio dos pares de bases, bem como as características e aspectos relevantes encontradas nestas simulações.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os resultados obtidos via simulação computacional dos processos de formação dos pares de bases considerando os critérios geométrico e energético. Para facilitar a apresentação, optou-se por subdividí-la em seções, de acordo com as propriedades analisadas e os parâmetros observados para os modelos de ligações de hidrogênio, ao mesmo tempo que serão discutidos os resultados.

Primeiramente é apresentada a sistemática utilizada, tanto do ponto de vista da coleta de dados quanto da apresentação dos mesmos. Com referência a primeira, esta foi realizada para cada modelo tomando como base uma planilha de experimentos expressos na tabela 4.1, na qual variam-se as dimensões do espaço de simulação; os iniciadores de cada experimento (*seed*) e desvios no comprimento das ligações de hidrogênio (*varmax*).

Tabela 4.1: Planilha de experimentos

Experimentos	*Dimensões*	*Seed*	*Varmax*
A 1	340 x 340	4109	0.20
A 2	370 x 370	4109	0.20
A 3	360 x 360	4109	0.20
B 1	340 x 340	3837	0.20
B 2	340 x 340	4109	0.20
B 3	340 x 340	13	0.20
C1	370 x 370	13	0.28
C2	370 x 370	13	0.15
C3	370 x 370	13	0.19

4.1 Parâmetros de simulação: domínio, iniciadores e varmax

Com o objetivo de analisar a dependência da evolução de cada um dos modelos de ligações de hidrogênio entre as bases em função do número de etapas computacionais, foram realizados experimentos com três parâmetros de controle, os quais são especificados na tabela 4.1. A determinação dos valores das dimensões levaram em consideração o número de moléculas (bases), seus respectivos raios e a sua mobilidade no espaço de simulação.

O modelo de simulação aqui descrito apresenta alguma dependência em função de como se inicia o processo, desde a etapa de colocação bem como da movimentação das moléculas. Com isso, cada experimento é "pré-estabelecido" quando da escolha do parâmetro inicial do gerador de números pseudoaleatórios (seed ou semente [1]). Seria necessário um número muito elevado de experimentos para simular todas as possíveis evoluções de configuração para um determinado sistema (qualquer linha da tabela 4.1). Por outro lado, um único experimento (um único iniciador) pode vir a ser demasiadamente dependente do caminho pelo qual se procede a evolução, podendo não ser representativo [33].

Para contornar esta problemática de probabilidades na evolução do sistema, consideramos a combinação de duas estratégias de amostragem. A primeira consiste em ao invés de avaliar cada evento (movimentação ou ligacão), contabilizam-se um conjunto deles, após ter sido satisfeita uma condição temporal (NPAR-número de etapas computacionais para a verificação de parâmetros). Em segundo lugar, utiliza-se, para a avaliação de cada parâmetro do sistema, a média simples ou específica para o parâmetro em questão de um conjunto de experimentos (diferentes iniciadores) para cada uma das linhas da tabela 4.1.

[1] isto porque não se tem um gerador de números aleatórios ideal e sim um gerador de números pseudoaleatórios

A realização de mais de um experimento (mais de uma semente) visa o tratamento estatístico dos dados na obtenção dos resultados. Um estudo clássico para o desenvolvimento de números pseudoaleatórios é apresentado na literatura [28], sendo a rotina geradora aquela que apresenta uma distribuição uniforme para o iniciador 13. Um estudo preliminar, dividindo o intervalo que corresponde a esta série em oito subdivisões, apresentou alguns desses valores igualmente espaçados, dos quais foram selecionados três deles (INI=13, 4109 e 3837) para serem experimentados [33].

O tratamento estatístico simplificado pode então ser ilustrado na construção dos gráficos do "*Número de ligações* x *Número de iterações*" para os experimentos realizados. Como aqui se faz necessário um grande número de experimentos, para contemplar o número de modelos a serem avaliados, optou-se pela utilização de três iniciadores, os quais são especificados na tabela 4.1, nos experimentos das séries B1, B2 e B3.

Com a finalidade de observar a dependência da evolução da formação de cada modelo de par de bases em relação ao comprimento das ligações de hidrogênio entre os sítios ativos (doador (d) e aceptor (a)), foi considerada uma variável varmax, a qual define os desvios aceitos no comprimento da ligação de hidrogênio. Para tanto, na verificação das ligações de hidrogênio nos pares de bases, utiliza-se o critério geométrico, onde uma ligação de hidrogênio se forma se a distância entre os átomos aceptor (a) e doador(d) segue a relação $(\bar{\times} - varmax) \leq DistP \leq (\bar{\times} + varmax)$, onde $DistP$ é a distância verificada entre os sítios ativos das bases envolvidas na ligação de hidrogênio e $\bar{\times}$ é o valor médio ou aceito da distância da ligação de hidrogênio considerada, a qual no caso é 3.0 $\mathring{A}$.

Os desvios padrões considerados, aqui designados pela variável varmax, e apresentados na tabela 4.1 por C1, C2 e C3, estão em concordância com os dados da literatura [58, 91].

A avaliação da evolução do próprio processo de ligações de hidrogênio, através do gerenciamento de parâmetros de saída, permite destacar os resultados para cada modelo considerado, tais como:

- o número de ligações em função do número de etapas computacionais;
- o número médio final de formação de pares de bases (ligações de hidrogênio) para cada modelo;
- o processo intermediário de formação de ligações de hidrogênio simples em relação as ligações múltiplas (duplas) tanto no critério geométrico quanto no energético;
- o comportamento quantitativo e qualitativo das curvas de formação de ligações com critérios geométrico e energético no início do processo iterativo;
- a probabilidade de ocorrência de cada modelo em relação a cada classe de par de bases.

Para o estudo dos resultados e avaliação dos parâmetros acima relacionados, o programa computacional gera 3 tipos de relatórios. O primeiro apresenta a evolução do sistema passo a passo, isto é, a cada ligação é computada a etapa computacional, as moléculas ligadas formando os pares de bases e o tipo de modelo formado. O segundo, gerenciado a cada NPAR (número de etapas computacionais) ou múltiplo dele, apresenta informações sobre o número de ligações ocorridas, a etapa computacional, os tipos de moléculas ligadas e o modelo de ligações de hidrogênio formado. E um terceiro, o mais significativo, que apresenta um resumo geral do estado em que se encontra o sistema, após um número estabelecido de etapas computacionais.

Cada um dos experimentos gera estes relatórios que, posteriormente, são tratados, tabelados e/ou graficados, conforme o caso. No prosseguimento deste

trabalho são apresentados os principais resultados obtidos do conjunto de simulações (tabela 4.1) realizado para cada um dos modelos computacionais analisados com os critérios geométrico e energético. Para facilitar a análise e compreensão dos resultados, inicialmente apresenta-se aqueles obtidos com o critério geométrico. Os resultados com o critério energético são descritos comparativamente aqueles anteriormente obtidos.

4.2 Simulações com o critério geométrico

A verificação do desempenho de cada modelo pode ser realizada através da análise da evolução do sistema em relação ao "tempo virtual" (número de etapas computacionais). A avaliação preliminar pode ser estabelecida de duas formas: a primeira interna a cada modelo, com a análise do efeito da variação dos parâmetros de simulação (dimensão, varmax e semente), a outra por comparação entre os diferentes modelos estudados.

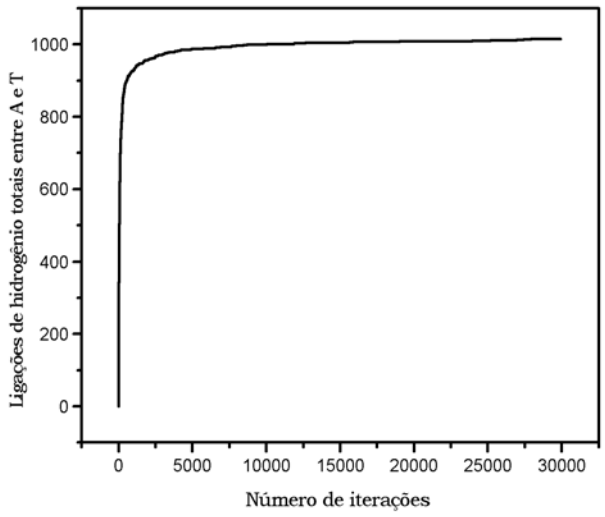

Figura 4.1: Análise da conversão na formação dos pares de bases: experimento A1

Numa visão panorâmica dos resultados dos modelos de ligações de hidrogênio envolvendo as bases A e T ou G e C, conforme exemplificado na figura 4.1, observou-se que, com a variação dos parâmetros a conversão (razão entre o número de ligações e o número máximo de ligações possível) alcançou um valor entre 97% e 98% para quase todos os experimentos simulados com os dados da tabela 4.1. A figura 4.1 mostra a curva de conversão em função do número de etapas computacionais para o experimento da série A1 (conforme tabela 4.1). É importante

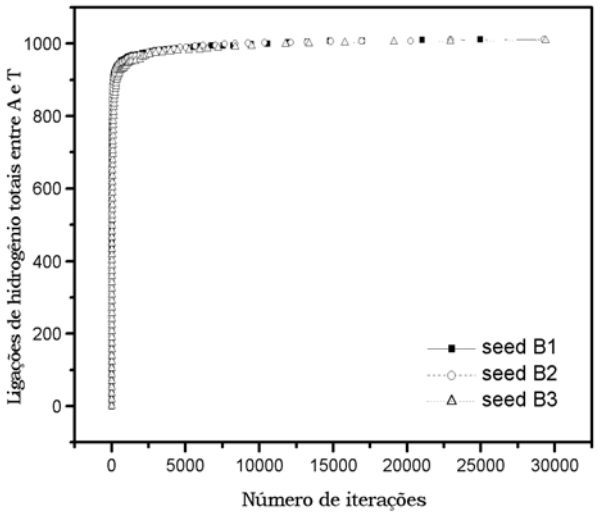

Figura 4.2: Ligações de hidrogênio totais entre A e T com iniciadores: B1, B2 e B3.

salientar que, após 1.9×10^4 etapas computacionais, quase todas as simulações podem ser consideradas como tendo atingido o grau máximo de conversão compatível com as condições da simulação, isto é, chegaram a um ponto estacionário. O mesmo comportamento foi observado nas simulações envolvendo as bases nitrogenadas Guanina e Citosina.

Os resultados analisados, com respeito a variação dos parâmetros (dimensão, seed, varmax), mostram que estes pouco afetam o comportamento quantitativo dos pares de bases. Como por exemplo, verifica-se na figura 4.2, o efeito dos três iniciadores (seed) para os experimentos das séries B1, B2 e B3; que não existem

diferenças significativas quanto ao número total de ligações obtidas na formação dos pares de bases [15].

Também há pouca diferença comparando-se o número de ligações entre os experimentos das séries A2 e B2. Assim, com as mesmas sementes e varmax, mas com dimensões diferentes, estes apresentaram 1007 e 1017 ligações totais, respectivamente, como se verifica na figura 4.3. O número total de bases contidas no domínio é de 2048.

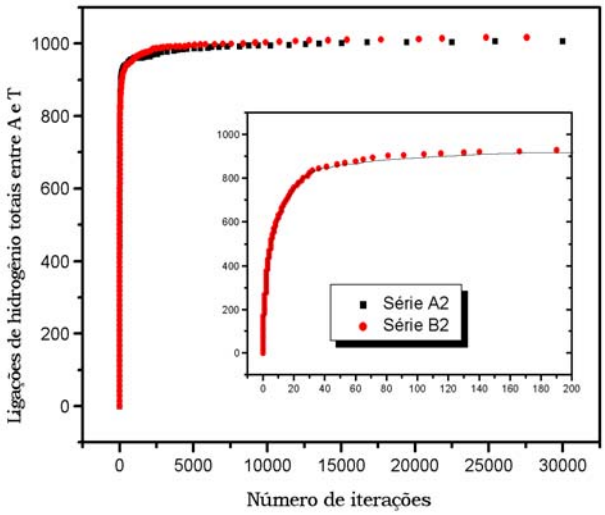

Figura 4.3: Ligações de hidrogênio totais entre A e T: experimentos A2 e B2.

A ordem de ocorrência para o número de ligações de uma série para outra, é pouco afetada, exceto para as séries C2 e C3 onde observa-se uma pequena variação. Este fato se deve ao valor do varmax (variação no comprimento das ligações de hidrogênio), que estes apresentam.

4.2.1 Ligações simples e duplas entre as bases Adenina e Timina

Os modelos de ligações de hidrogênio simples entre as bases A e T desempenham um papel importante para todos os modelos simulados. Isto porque,

elas participam das etapas de iniciação na formação dos modelos de ligações múltipla (dupla), ou seja, são um processo intermediário na formação dos pares de bases.

Uma análise das ligações de hidrogênio simples (uma ligação de hidrogênio), entre as bases nitrogenadas A e T, indica que existem 4 modelos diferentes, que são reconhecidos como 1pa (AN6,TO4), 1pb (AN1,TN3), 1pc (AN6,TO2) e 1pd (AN7,TN3) onde, por exemplo, a notação 1pa (AN6,TO4) significa uma ligação de hidrogênio simples entre o nitrogênio 6 da Adenina e o oxigênio 4 da Timina (ver tabela 2.3 na seção 2.6.)

Relativamente às características dos modelos com ligações de hidrogênio duplas, conforme dados da literatura [45], os 4 tipos existentes são ATWC (Adenina e Timina de Watson e Crick) que envolvem em sua formação as ligações de hidrogênio 1pa e 1pb; ATrWC (Adenina e Timina reverso de Watson e Crick) para 1pb e 1pc; ATH (Adenina e Timina de Hoogsteen) para 1pa e 1pd e ATrH (Adenina e Timina reverso de Hoogsteen) para 1pd e 1pc. Cada tipo de ligação simples pode contribuir para a formação de 2 modelos de ligações duplas; isto depende de sua acessibilidade geométrica, ou seja, o quão facilmente um par de moléculas consegue se posicionar formando uma dada geometria.

Quanto aos resultados da formação de ligações de hidrogênio simples, com o critério geométrico, observa-se para quase a totalidade dos experimentos a existência de diferenças significativas entre os modelos. Esta constatação pode ser confirmada escolhendo-se, como por exemplo, o experimento da série C2 apresentado na figura 4.4, onde são apresentadas as curvas de formação para os 4 modelos de ligações simples. Nesta pode-se concluir que os modelos 1pa na relação 40/79, e 1pc na relação 31/79, ocorrem em maior quantidade relativamente aos modelos 1pb na relação 4/79, e 1pd na relação 4/79, onde 79 (40+31+4+4=79) corresponde ao número total de ligações simples ocorridas neste experimento.

Assim, observa-se ainda na figura 4.4 um crescimento elevado no número de ligações no início do processo iterativo (até aproximadamente 75 iterações), con-

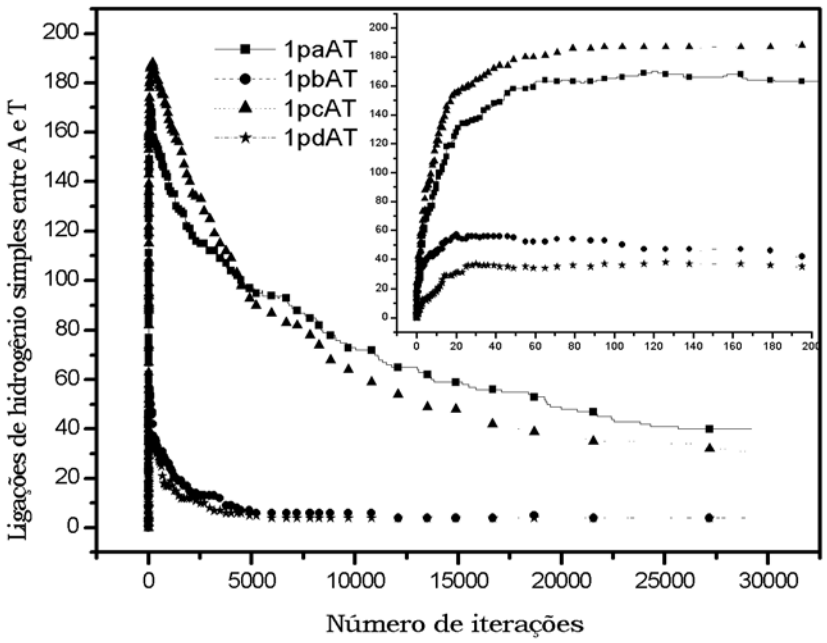

Figura 4.4: Ligações simples entre A e T: experimento C2

forme ampliação apresentada nesta figura, sofrendo, em seguida, um decréscimo e tendendo à estabilidade no final do processo iterativo. Há ainda pequenas flutuações na evolução das curvas de ligações simples entre as bases A e T. Estas são características da formação de ligações simples e os conseqüentes decréscimos contribuem para a formação de modelos com ligações duplas. Ou seja, se existe uma ligação de hidrogênio simples e num próximo passo existe a possibilidade para a formação de uma segunda ligação no modelo, este se transformará num com ligações múltiplas, não sendo mais contabilizado como ligação simples.

A figura 4.5, apresenta a evolução das curvas de ligações de hidrogênio, em relação ao número de etapas computacionais, para cada um dos modelos de ligações múltiplas comparando os experimentos das séries A1 e B3. Verifica-se que a ordem encontrada quanto ao número de ligações é a mesma: $ATWC > ATrWC > ATrH > ATH$, independente do iniciador.

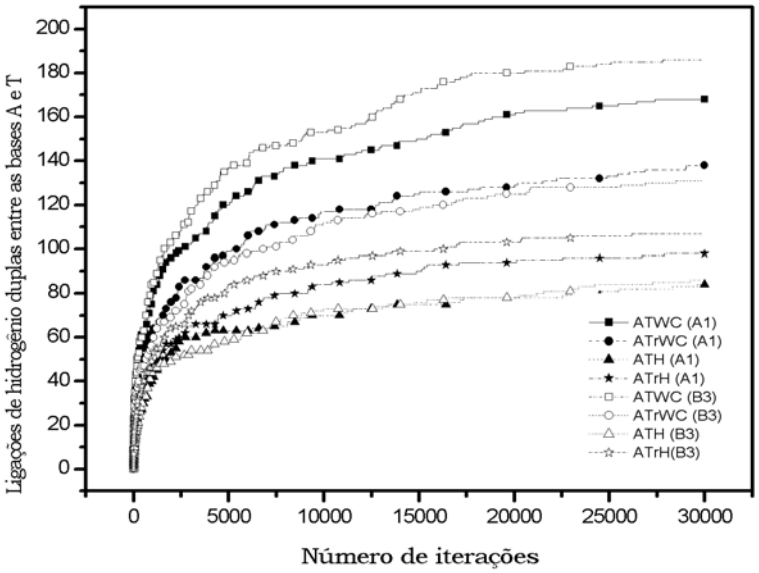

Figura 4.5: Ligações duplas entre A e T: experimentos A1 e B3

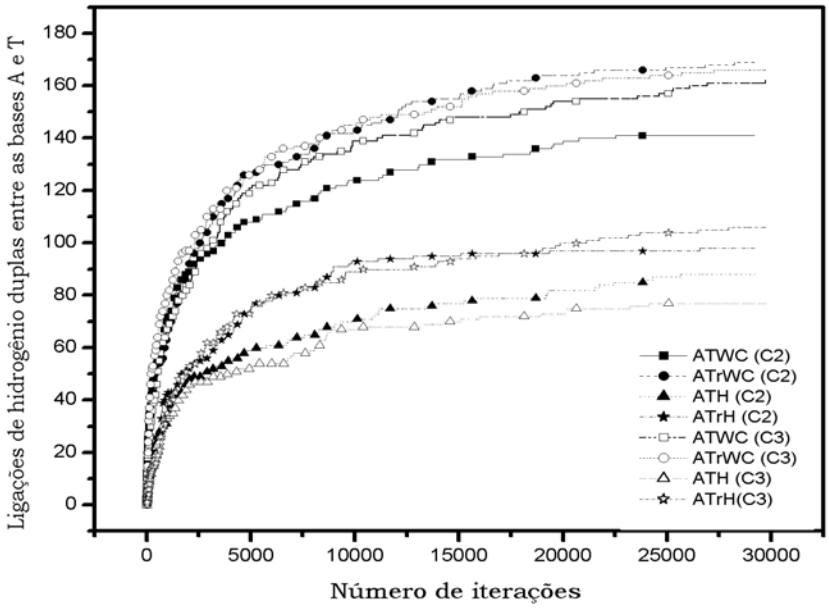

Figura 4.6: Ligações duplas entre A e T: experimentos C2 e C3

Analisando-se o comportamento dos modelos de ligações duplas para as séries C2 e C3, ilustrados na figura 4.6, observa-se que a ordem de ocorrência do número de ligações é dada por $ATrWC > ATWC > ATrH > ATH$ que difere da encontrada para as demais séries da planilha de experimentos. Como a única diferença entre os parâmetros destas séries encontra-se no varmax, este pode ter influenciado a ordem de ocorrência dos modelos. Repare ainda o valor do *varmax* para as séries A, B e C3 (tabela 41.). Assim, pode-se especular que no limite quando t (tempo computacional) tende para o infinito verifica-se que para a série C3 tem-se $ATWC \approx ATrWC$ quanto ao número de ligações, conforme figura 4.6.

Considerando-se a ordem encontrada nos resultados para a totalidade dos experimentos, mesmo com a variação dos parâmetros, pode-se concluir que a maior probabilidade de ocorrência de ligações duplas, com o critério geométrico, é dada por $ATWC > ATrWC > ATrH > ATH$.

Na figura 4.7 é mostrado o comportamento geral do somatório destes modelos, para o experimento da série C2.

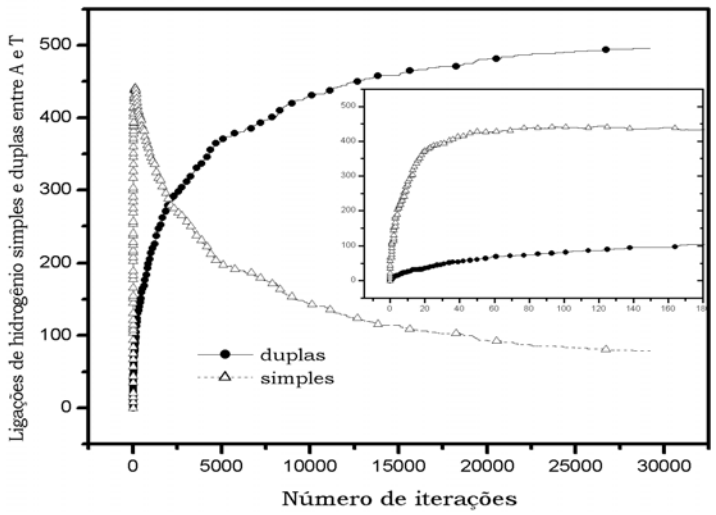

Figura 4.7: Ligações simples e duplas entre A e T: experimento C2

Assim, observa-se que existe uma relação no processo evolutivo entre as curvas de ligações, pois para um mesmo intervalo de (≈ 75) iterações, verifica-se que o decréscimo no número de ligações simples equivale ao acréscimo nas ligações duplas.

4.2.2 Ligações simples e duplas entre as bases Adenina e Adenina

Além da análise de formação de pares de bases para os modelos heterogêneos (Adenina e Timina), foram analisados o comportamento das ligações para os modelos homogêneos (Adenina com Adenina e Timina com Timina). Assim, existem quatro tipos de ligações de hidrogênio simples entre A e A, que são reconhecidos como 1pa (AN6,AN1), 1pb (AN1,AN6), 1pc (AN6,AN7) e 1pd (AN7,AN6) onde, por exemplo, a notação 1pa (AN6,AN1) significa uma ligação de hidrogênio simples entre o nitrogênio 6 da Adenina e o nitrogênio 1 da outra Adenina. Os resultados

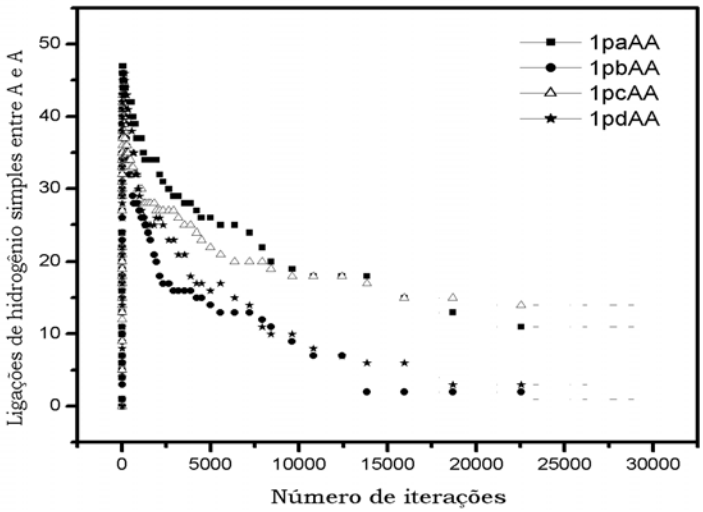

Figura 4.8: Ligações simples entre A e A: experimento C2

das simulações com as ligações simples entre as bases Adenina e Adenina fornecem variações significativas quanto ao número de ligações entre os modelos. Esta cons-

tatação pode ser confirmada escolhendo-se, como por exemplo, o experimento da série C2 apresentado na figura 4.8.

Nesta, pode-se concluir que os modelos 1pa, na relação 10/26, e 1pc na relação 14/26, ocorrem em maior quantidade relativamente aos modelos 1pd, na relação 1/26, e 1pb na relação 1/26, onde 26 corresponde ao número total de ligações simples ocorridas neste experimento. Nas ligações de hidrogênio duplas,

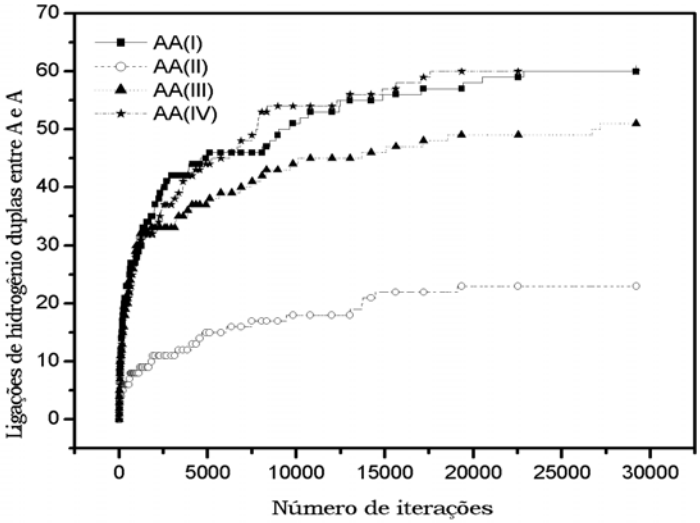

Figura 4.9: Ligações duplas entre A e A: experimento C2

os 4 modelos considerados são AA(I), que envolve em sua formação as ligações de hidrogênio 1pa e 1pb; AA(II) para 1pb e 1pc; AA(III) para 1pc e 1pd e AA(IV) para 1pa e 1pd.

Uma representação do comportamento destes modelos, dentre os vários experimentos realizados, é apresentada na figura 4.9, para o experimento da série C2. Pode-se verificar que os tipos de pares de bases que apresentam o maior número de ligações são AA(I) e AA(IV) com 61/195 e 60/195 ligações, respectivamente, onde 195 corresponde ao número total de ligações duplas ocorridas neste experimento.

4.2.3 Ligações simples e duplas entre as bases Timina e Timina

Existem 4 tipos de ligações simples, entre T e T que são nomeados como 1pa (TO4,TN3), 1pb (TN3,TO2), 1pc (TO2,TN3) e 1pd (TN3,TO4). Por exemplo, 1pa (TO4,TN3) significa uma ligação de hidrogênio entre o oxigênio 4 da Timina e o nitrogênio 3 da outra Timina. Uma vez que o comportamento destes é semelhante ao obtido nos modelos A com A e A com T simples, optamos por não representá-las graficamente. Destaca-se apenas que, para a maioria dos experimentos realizados entre T e T, os modelos 1pa e 1pc ocorrem em maior quantidade relativamente aos modelos 1pb e 1pd.

Finalmente, quanto à formação das ligações duplas, os 4 modelos considerados são TT(I), que envolve em sua formação as ligações de hidrogênio 1pa e 1pb; TT(II) para 1pa e 1pd; TT(III) para 1pb e 1pc e TT(IV) para 1pc e 1pd. Os resultados obtidos indicaram quantidades de ligações semelhantes entre os modelos para a maioria dos experimentos realizados. Uma exceção foi a diferença observada

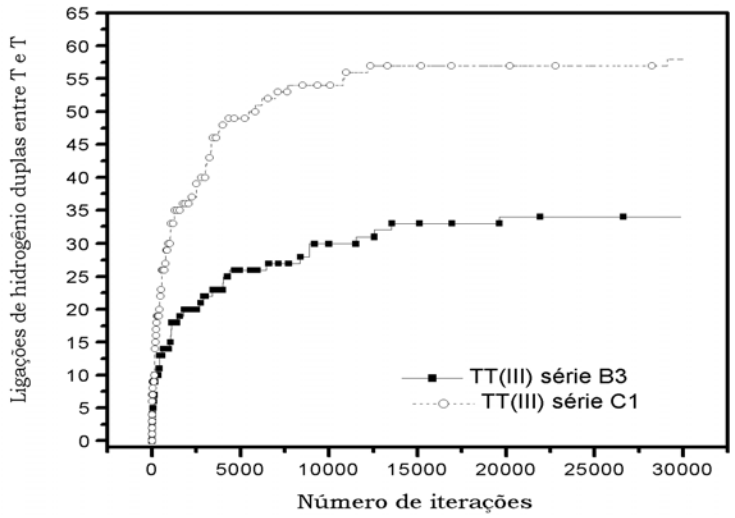

Figura 4.10: Ligações duplas entre T e T: experimentos B3 e C1

quanto ao número de ligações para o modelo TT(III), encontrado entre os experi-

mentos das séries B3 e C1, como se observa na figura 4.10. Este comportamento pode ser justificado pela diferença existente entre os parâmetros das duas séries, ou seja, estes apresentam uma variação de 20% no comprimento da ligação de hidrogênio (varmax).

4.2.4 Ligações simples e duplas entre as bases Guanina e Citosina

Com relação as ligações simples entre as bases G e C, existem seis tipos, que são reconhecidos como 1pa (GO6,CN4), 1pb (GN1,CN3), 1pc (GN2,CO2), 1pd (GN1,CO2), 1pe (GN2,CN3) e 1pf (GN3,CN4) onde, por exemplo, a notação 1pa (GO6,CN4) significa uma ligação de hidrogênio simples entre o oxigênio 6 da Guanina e o nitrogênio 4 da Citosina.

Assim, verifica-se na figura 4.11, para o experimento da série C2, que os modelos 1pa na relação 108/269 e 1pc na relação 91/269 ocorrem em maior quantidade relativamente aos modelos 1pb na relação 38/269, 1pd na relação 12/269, 1pe na relação 19/269 e 1pf na relação 1/269, onde 269 corresponde ao número total de ligações simples ocorridas neste experimento. Como mencionado anteriormente,

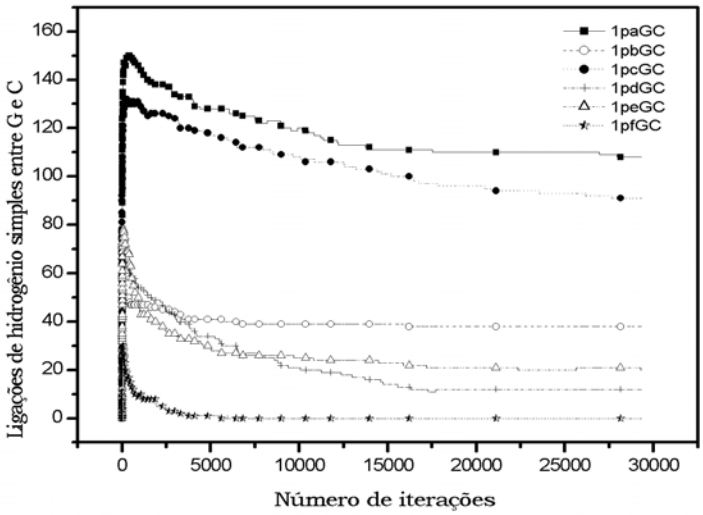

Figura 4.11: Ligações de hidrogênio simples entre G e C: experimento C2

os modelos de ligações simples são de suma importância na formação das ligações múltiplas, uma vez que cada uma destas representa o somatório das ligações de hidrogênio simples. Neste contexto, verifica-se que existem três modelos de ligações múltiplas, entre as bases Guanina e Citosina dados por: GCWC (Guanina e Citosina de Watson e Crick) que envolve em sua formação as ligações 1pa, 1pb e 1pc; GCrWC (Guanina e Citosina reverso de Watson e Crick) para 1pd e 1pe e GC(II) (Guanina e Citosina II) para 1pe e 1pf, assim denominadas segundo dados da literatura [45].

Desta forma, analisando o comportamento para os modelos com ligações múltiplas na figura 4.12, verificou-se que os tipos GCWC e GCrWC foram os que apresentaram o maior número com 129/330 e 120/330 ligações, respectivamente. O modelo GC(II) apresentou 81/330 ligações, onde 330 corresponde ao número total de ligações múltiplas ocorridas no experimento da série C2.

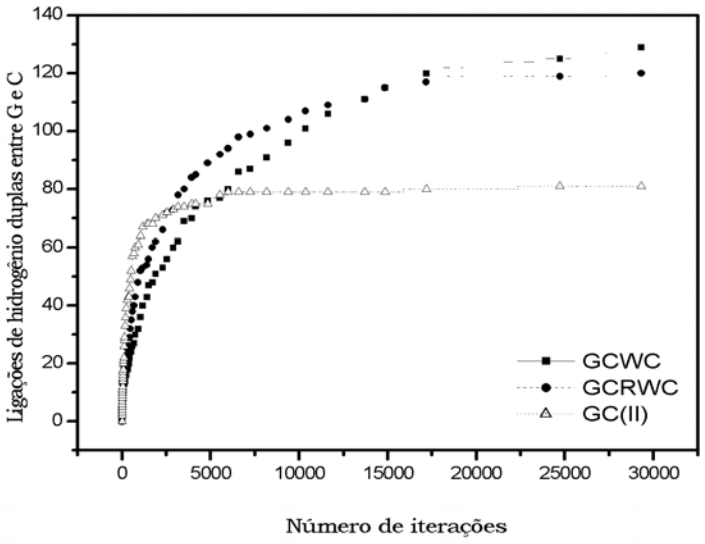

Figura 4.12: Ligações de hidrogênio múltiplas entre G e C: experimento C2

Também foi observada a evolução das ligações de hidrogênio simples e múltiplas totais entre as bases Guanina e Citosina. Assim, na figura 4.13 verifica-se que no início do processo iterativo o número de ligações simples é relativamente superior aquele formado por múltiplas.

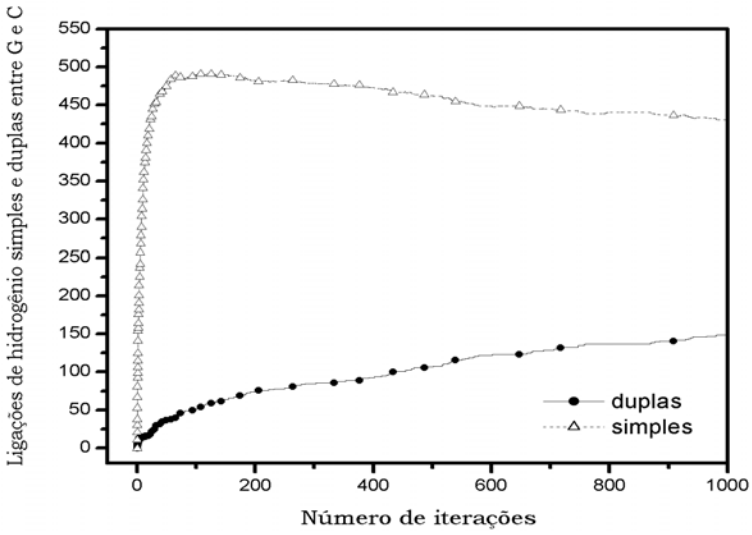

Figura 4.13: Ligações de hidrogênio simples e duplas entre G e C: experimento C2

4.2.5 Ligações simples e duplas entre as bases Guanina e Guanina

De acordo com as características dos modelos de ligações simples entre as bases G com G, existem sete tipos, designadas por 1pa (GO6,GN1), 1pb (GN1,GO6), 1pc (GN2,GN7), 1pd(GN1,GN7), 1pe (GN2,GO6), 1pf (GN3,GN2) e 1pg (GN2,GN3) onde, por exemplo, a notação 1pa (GO6,GN1) significa uma ligação de hidrogênio simples entre o oxigênio 6 da Guanina e o nitrogênio 1 da outra Guanina. Destaca-se que os modelos 1pb e 1pf foram os que apresentaram o menor número de ligações relativamente aos demais.

Quanto aos modelos de ligações múltiplas entre G e G, sabe-se que existem 4 tipos, os quais são estabelecidas por (1pa e 1pb) para GG(I), (1pb e 1pc) para GG(II), (1pd e 1pe) para GG(III) e (1pf e 1pg) para GG(IV). Neste sentido, os que apresentaram o maior número de ligações foram GG(I) e GG(IV), como pode-se observar na figura 4.14.

Assim, considerando apenas o critério geométrico, conclui-se que a preferência na formação de uns modelos em relação a outros está relacionada a acessi-

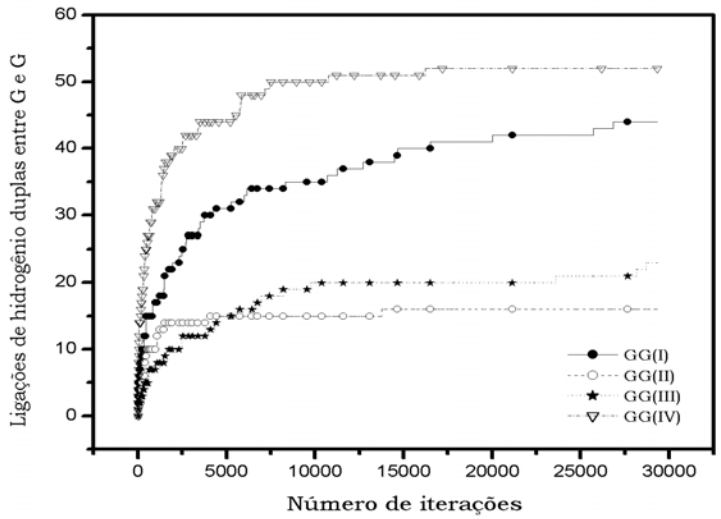

Figura 4.14: Ligações de hidrogênio duplas entre G e G: experimento C2

bilidade geométrica do par, bem como aos tipos de ligações de hidrogênio simples envolvidas no modelo.

4.2.6 Ligações simples e duplas entre as bases Citosina e Citosina

Finalmente, para as ligações de hidrogênio simples entre as bases Citosina e Citosina, destacam-se dois tipos: 1pa definido por (CN4 e CN3) e 1pb (CN3 e CN4). Como só há a possibilidade de um único modelo de ligação dupla entre Citosina e Citosina, não há necessidade de fazer uma análise comparativa entre os modelos deste par.

4.3 Simulações com o critério energético

Para facilitar a avaliação da contribuição energética nos processos de formação dos pares de bases, os resultados serão apresentados de forma comparativa àqueles obtidos anteriormente com o critério geométrico. Neste sentido, é importante salientar que o processo de verificação e análise na formação das ligações dos pares

de bases divide-se em duas etapas: primeiro inicia-se o cálculo da probabilidade de formação dos modelos, considerando apenas o *critério geométrico* .

A segunda etapa consiste na verificação da formação do par de bases com o critério *energético*. Deste modo, utiliza-se os mesmos parâmetros de simulação quando da análise com o critério geométrico. É importante salientar que este processo engloba o critério geométrico, pois para se proceder a análise da probabilidade energética deve-se considerar também as características geométricas do modelo.

Salienta-se que na etapa de verificação com o critério energético, mesmo que geometricamente exista a possibilidade de formação de um determinado modelo (distância e orientação adequadas), se energeticamente a probabilidade de ligação do par não for favorável, a ligação não é aceita. Assim, as bases nitrogenadas são "soltas" no espaço de simulação e podem concorrer para a formação de novos modelos (A-A, A-T ou T-T).

É oportuno dizer que duas bases verificadas para a formação de um determinado modelo (por exemplo o par A-T), não necessariamente formarão um par da mesma classe (heterogênea). Um dos fatores determinante na ocorrência de modelos de uma classe ou outra é a acessibilidade geométrica na formação do par de bases.

4.3.1 Ligações simples e duplas entre as bases Adenina e Timina

Apresenta-se uma análise do comportamento dos modelos heterogêneos e homogêneos entre as bases Adenina e Timina. Para facilitar a verificação dos resultados, indica-se na tabela 4.2 os valores dos fatores de Boltzmann, obtidos no cálculo de estabilização dos pares pelo método semi-empírico AM1.

O número de ligações totais entre as bases Adenina e Timina, com o critério energético, foi na média inferior àquele obtido com o critério geométrico. Considere, como por exemplo, o experimento da série C2, onde obteve-se:

- 575 ligações totais (simples e duplas) entre A e T, para o *critério geométrico*;
- 545 ligações totais (simples e duplas) entre A e T, para o *critério energético.*

Por outro lado, para os modelos homogêneos entre as bases Adenina com Adenina e Timina com Timina observou-se com o critério energético um acréscimo no número de ligações totais. Assim, ainda considerando o exemplo da série C2, tem-se:

- 221 ligações totais (simples e duplas) entre A e A, para o *critério geométrico*;
- 234 ligações totais (simples e duplas) entre A e A, para o *critério energético*;
- 219 ligações totais (simples e duplas) entre T e T, para o *critério geométrico*;
- 235 ligações totais (simples e duplas) entre T e T, para o *critério energético.*

Contudo, observa-se que o número total de ligações duplas e simples para os modelos homogêneos e heterogêneos (AA+ AT+TT) foram semelhantes. Por exemplo, no experimento da série C2 com os critérios geométrico e energético ocorreram 1015 e 1014 ligações respectivamente. Comportamentos semelhantes foram também observados para outras séries da planilha de experimentos.

Entre as características qualitativas mais relevantes na análise comparativa dos resultados com ambos critérios para os modelos entre Adenina e Timina, cabe ressaltar a diferença significativa quanto ao número de ligações simples no início do processo iterativo.

Tabela 4.2: Valores dos fatores de Boltzmann entre A e T

Modelos de pares de bases	*ligações-H simples*	*Fator de Boltzmann*
AT(1pa)	N6-H...O4	0.14
AT(1pb)	N1...H-N3	0.10
AT(1pc)	N6-H...O2	0.11
AT(1pd)	N7...H-N3	0.10
	ligações-H duplas	*Fator de Boltzmann*
AT(WC)	N6-H...O4 e N1...H-N3	0.85
AT(rWC)	N1...H-N3 e N6-H...O2	0.68
AT(H)	N6-H...O4 e N7...H-N3	1
AT(rH)	N6-H...O2 e N7...H-N3	0.86

Na figura 4.15, para o experimento da série C2, é visível que a curva de formação de ligações simples com o critério geométrico apresenta um número de ligações superior relativamente aquela com o critério energético.

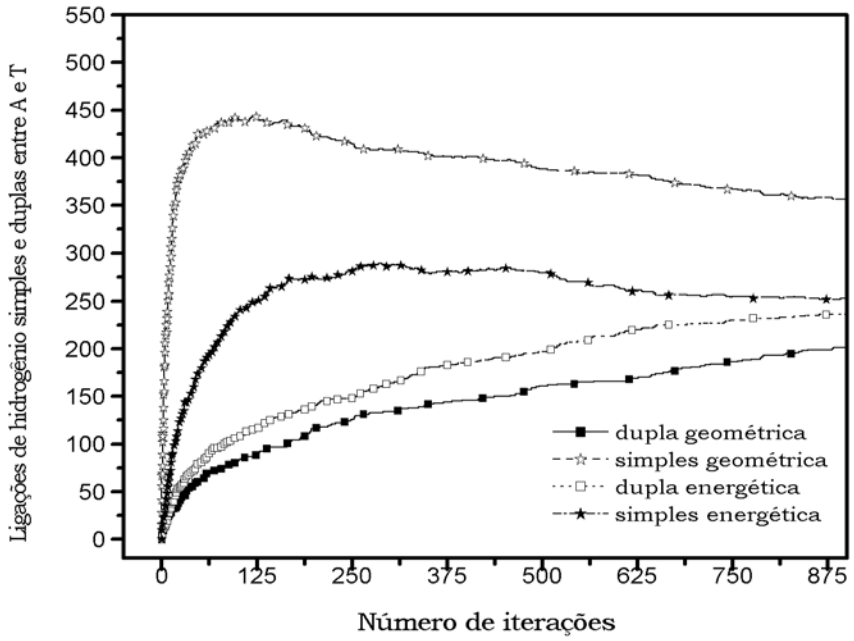

Figura 4.15: Ligações de hidrogênio simples e duplas entre A e T com critérios geométrico e energético: experimento C2.

Esta informação torna-se mais clara na análise comparativa da "taxa" de formação das ligações dos pares em relação a um número particular de iterações (tempo computacional). Neste sentido, tomando-se por exemplo o experimento da série C2 entre A e T, na figura 4.15 tem-se, para 125 iterações:

- 450 *ligações simples* com o *critério geométrico*;
- 250 *ligações simples* com o *critério energético*;
- 98 *ligações duplas* com o *critério geométrico*;
- 110 *ligações duplas* com o *critério energético*.

Esta diferença dita que existe maior flexibilidade na formação dos pares quando se considera o critério geométrico. Desta forma, na verificação das ligações

simples com o critério geométrico, a direcionalidade (orientações ideais) não é estritamente considerada. Salienta-se ainda que na análise com o critério energético, existe um fator de probabilidade (energético) associado a cada tipo de ligação de hidrogênio que influi na formação dos pares. Outros aspectos que ainda podem ser ressaltados é que o critério energético "acelera"a formação de ligações múltiplas e facilita a conversão das ligações simples em duplas.

Verifica-se ainda que o critério energético "filtra"os modelos de ligações simples e "escolhe"aqueles que têm a maior probabilidade de ocorrência, ou seja, aqueles que estão num estado energeticamente favorável a formação do par.

Com respeito ao comportamento das curvas de ligações duplas entre as bases Adenina e Timina observa-se, no gráfico da figura 4.15, que tanto com o critério geométrico quanto com o energético, estas não apresentaram diferenças tão significativas como aquelas das curvas de ligações simples com os mesmos critérios. Uma justificativa para a semelhança no número de ligações duplas em ambos critérios vem do fato que os pares não ocorrem naturalmente por ligações simples.

Adicionalmente, um dos fatores também relacionado à importância do critério energético são as cargas atômicas, as quais tem função importante, não somente na determinação da energia de estabilização da ligação de hidrogênio do par, mas também em determinar a estrutura de mínimo global.

Destaca-se ainda que a ordem de ocorrência encontrada para os referidos modelos de ligações simples (1pa, 1pb, 1pc, 1pd), com o critério energético, concorda com os valores do fator de probabilidade de Boltzmann obtido no cálculo de estabilização de energia (veja tabela 4.2). Ou seja, os modelos de pares *1pa* e *1pc* são os que quantitativamente mais ocorrem, e são energeticamente os mais favoráveis.

De um modo geral, as curvas de formação de ligações de hidrogênio simples entre as bases Adenina e Timina apresentaram um comportamento qualitativo semelhante em ambos critérios. Deste modo, conforme a figura 4.16, no início do

processo iterativo há um grande número de ligações (mais acentuada para a simulação com o critério geométrico) seguida por um decréscimo destas, que contribuem para a formação de ligações duplas.

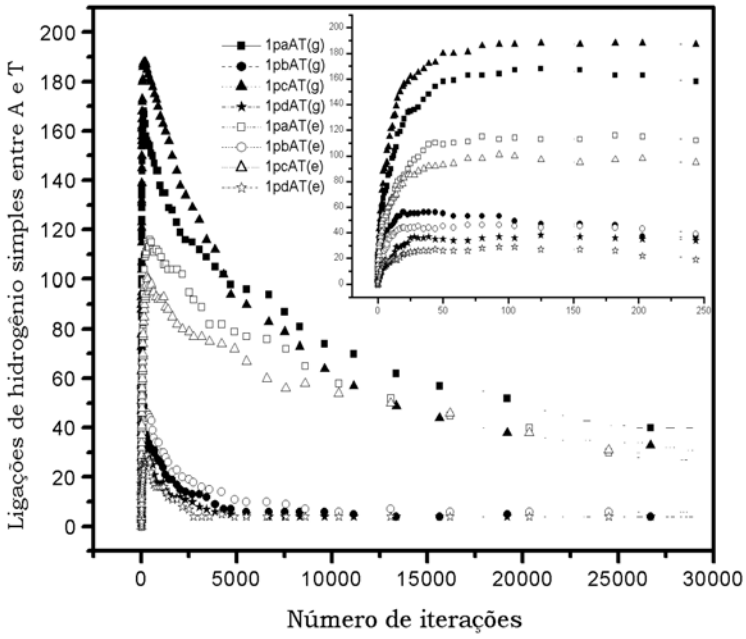

Figura 4.16: Ligações de hidrogênio simples entre A e T com critérios geométrico e energético: experimento C2,(g) geométrico-(e) energético

Tomando-se o intervalo de 200 iterações, na figura 4.16 (região em expansão), tem-se:

- 163 *ligações simples 1pa* com o *critério geométrico*;
- 110 *ligações simples 1pa* com o *critério energético*;
- 37 *ligações simples 1pb* com o *critério geométrico*;
- 45 *ligações simples 1pb* com o *critério energético*;
- 188 *ligações simples 1pc* com o *critério geométrico*;
- 93 *ligações simples 1pc* com o *critério energético*;
- 35 *ligações simples 1pd* com o *critério geométrico*;
- 26 *ligações simples 1pd* com o *critério energético*.

Os números das ligações para os modelos de pares ATWC, ATrWC, ATH, ATrH apresentaram características semelhantes, tanto com o critério geométrico quanto com o energético. Este comportamento é mais facilmente entendido analisando a figura 4.17 e a quantidade de ligações formadas em cada modelo para o experimento da série C2, dadas por:

- 141 ligações de ATWC com critério geométrico;
- 172 ligações de ATWC com critério energético;
- 169 ligações de ATrWC com critério geométrico;
- 149 ligações de ATrWC com critério energético;
- 88 ligações de ATH com critério geométrico;
- 88 ligações de ATH com critério energético;
- 98 ligações de ATrH com critério geométrico;
- 75 ligações de ATrH com critério energético.

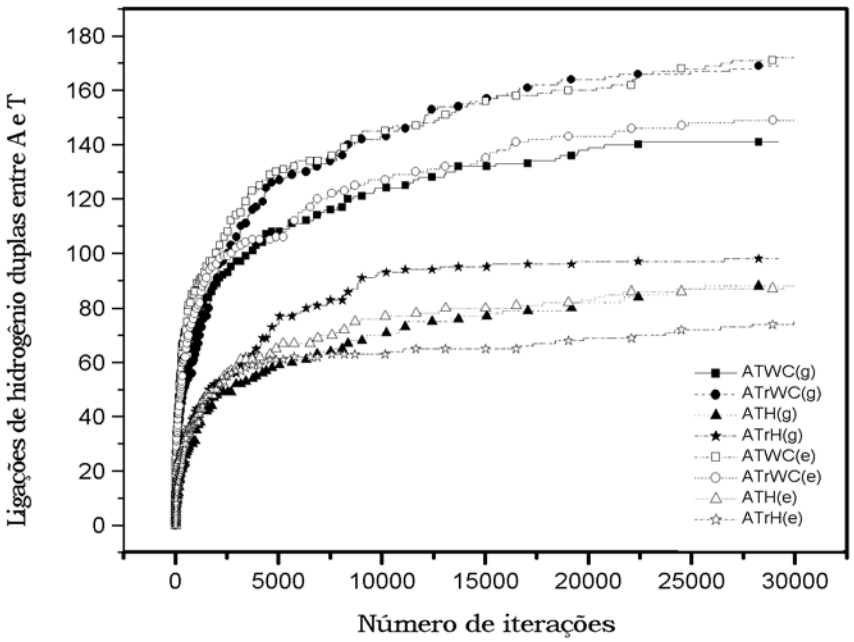

Figura 4.17: Ligações de hidrogênio duplas entre A e T com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

Apesar da diferença observada na ordem de ocorrência das ligações entre os critérios geométrico e energético para o experimento da série C2, tem-se que para a maioria dos experimentos a ordem destes modelos é dada por: $ATWC > ATrWC > ATrH > ATH$. Ressalta-se que, no cálculo da energia de estabilização dos pares pelo método semi-empírico AM1, entre os modelos com ligações duplas entre as bases A e T, o modelo ATH foi o que apresentou o fator de Boltzmann com maior probabilidade energética, conforme tabela 4.2. Mas, este não foi o modelo que apresentou a maior possibilidade em nosso experimento, como pode-se observar na figura 4.17.

Considerando-se as ligações simples como precedentes para a formação de "um"ou "dois"dos modelos com ligações duplas, e a relativa acessibilidade geométrica na formação do par de bases, pode-se concluir que estes são os fatores que provavelmente interferiram na preferência de um determinado par de bases a outro. Assim, destaca-se que os modelos 1pa, 1pb e 1pc precedentes para a formação dos modelos de ligações duplas ATWC e ATrWC apresentam, no experimento da série C2, tanto com o critério geométrico quanto com o energético, quantidades de ligações viáveis, que contribuem para a formação de ligações destes modelos, como pode-se observar na região em expansão da já citada figura 4.16.

Ou seja, a probabilidade de formação de uma dada ligação múltipla não depende apenas da energia desta ligação, mas da acessibilidade geométrica desta e também da acessibilidade geométrica das ligações simples que lhe são precursoras.

4.3.2 Ligações simples e duplas entre as bases Adenina e Adenina

Também se apresenta, na tabela 4.3, os valores dos fatores de Boltzmann para os pares A e A.

O comportamento é semelhante ao observado para os modelos heterogêneos entre as bases Adenina e Timina. Também é importante ressaltar a diferença

Tabela 4.3: Valores dos fatores de Boltzmann entre A e A

Modelos de pares de bases	*ligações-H simples*	*Fator de Boltzmann*
AA(1pa)	N6-H...N1	0.02
AA(1pb)	N1...H-N6	0.10
AA(1pc)	N6-H...N7	0.06
AA(1pd)	N7...H-N6	0.10
	ligações-H duplas	*Fator de Boltzmann*
AA(I)	N6-H...N1 e N1...H-N6	1
AA(II)	N1...H-N6 e N6-H...N7	0.63
AA(III)	N6-H...N7 e N7...H-N6	0.51
AA(IV)	N6-H...N1 e N7...H-N6	0.62

na formação das ligações de hidrogênio simples entre as bases Adenina e Adenina apresentada entre os critérios geométrico e energético.

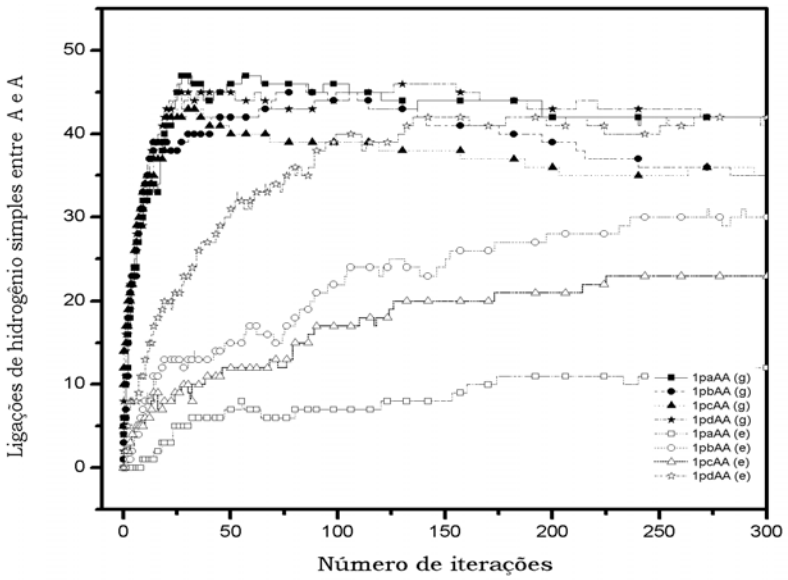

Figura 4.18: Ligações de hidrogênio simples entre A e A com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

Verifique o comportamento no início do processo iterativo, como se observa na figura 4.18. Considerando-se um número de 100 iterações para ambos critérios, nota-se que no geométrico o número de ligações simples, para cada modelo, é superior ao obtido com o critério energético. Verifica-se, ainda na figura 4.18,

que as curvas de ligações simples com o critério geométrico não apresentam uma discrepância significativa no número de ligações entre os modelos, ao contrário do caso energético.

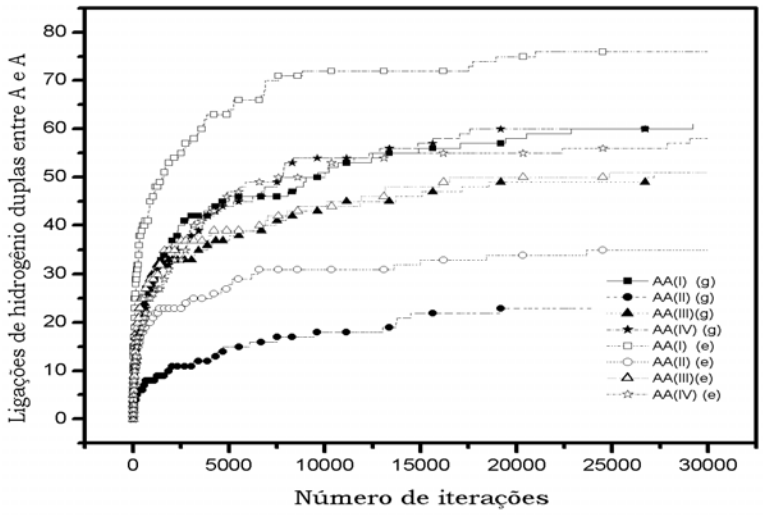

Figura 4.19: Ligações de hidrogênio duplas entre A e A com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

Para as ligações múltiplas entre os pares Adenina e Adenina, como observado na figura 4.19 para ambos critérios, tem-se que dos 4 tipos existentes o que apresenta a maior possibilidade é o modelo AA(I). Ainda, salienta-se que este concorda com o resultado encontrado nos cálculos energéticos, como o mais energeticamente favorável, conforme indicado na tabela 4.3.

Observa-se que o número de ligações múltiplas totais entre as bases Adenina e Adenina é maior com o critério energético. Um fator que contribui para este comportamento é o número elevado de ligações múltiplas ocorrida no modelo AA(I) com o critério energético. Destaca-se ainda que a ordem de ocorrência de ligações duplas entre Adenina e Adenina é basicamente a mesma para ambos critérios.

4.3.3 Ligações simples e duplas entre as bases Timina e Timina

Inicialmente, apresenta-se na tabela 4.4 os valores dos fatores de Boltzmann para os modelos entre T e T.

Tabela 4.4: Valores dos fatores de Boltzmann entre T e T

Modelos de pares de bases	*ligações-H simples*	Fator de Boltzmann
TT(1pa)	O4...H-N3	0.39
TT(1pb)	N3-H...O2	0.28
TT(1pc)	O2...H-N3	0.15
TT(1pd)	N3-H...O4	0.15
	ligações duplas	*Fator de Boltzmann*
TT(I)	O4...H-N3 e N3-H...O2	0.94
TT(II)	O4...H-N3 e N3-H...O4	1
TT(III)	N3-H...O2 e O2...H-N3	0.88
TT(IV)	O2...H-N3 e N3-H...O4	0.94

Os resultados das curvas de ligações de hidrogênio com os critérios geométrico e energético entre as bases Timina e Timina apresentaram comportamentos qualitativos semelhantes aos observados nos pareamentos envolvendo as bases Adenina e Adenina. Para a maioria dos experimentos envolvendo ligações simples (1pa, 1pb, 1pc, 1pd), com o critério geométrico, os tipos de ligações 1pa e 1pc foram os que apresentaram o maior número de ligações frente aos modelos 1pb e 1pd.

Quando da simulação envolvendo o critério energético, observa-se semelhança na preferência das ligações simples, ou seja, os modelos que apresentaram maior número de ocorrência foram 1pa e 1pc. Dos 4 tipos de ligações duplas existente, a maior possibilidade de ocorrência observada é para TT(I), formada pelas ligações simples 1pa e 1pb.

Visando uma melhor comparação das ligações com os critérios geométrico e energético, apresenta-se na figura 4.20 as curvas de formação de ligações duplas e simples entre Timina e Timina em função do número de etapas computacionais (iterações). Sendo assim, observa-se que a taxa de formação envolvendo as ligações

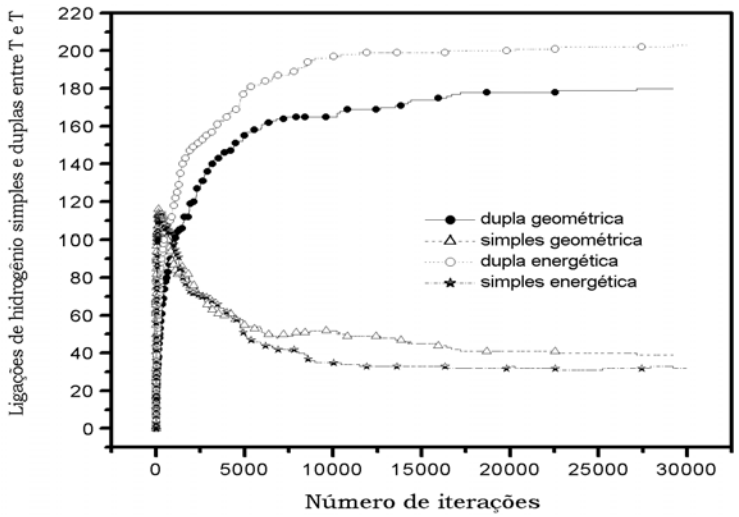

Figura 4.20: Ligações de hidrogênio simples e duplas entre T e T com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

duplas, com o critério energético, apresentou um valor relativamente superior aquelas envolvendo o geométrico. Porém, vale salientar que a diferença quanto ao número de ligações simples, com ambos critérios, não foram tão significativas.

4.3.4 Ligações simples e duplas entre as bases Guanina e Citosina

O número total de ligações para os modelos heterogêneos, com o critério energético, foram na média inferiores ao número de ligações obtido com o critério geométrico. A tabela 4.5 apresenta os valores dos fatores de Boltzmann, para os modelos de pares de bases entre G e C. Os baixos valores dos fatores de Boltzmann contribuem, neste caso, para "desfavorecer"a formação com o critério energético.

Tomando-se, como exemplo a comparação com ambos critérios para o experimento da série C2 entre G e C, tem-se:

- 599 ligações totais (simples e duplas) entre G e C, para o *critério geométrico*;
- 237 ligações totais (simples e duplas) entre G e C, para o *critério energético*.

Tabela 4.5: Valores dos fatores de Boltzmann entre G e C

Modelos de pares de bases	*ligações-H simples*	*Fator de Boltzmann*
GC(1pa)	O6...H-N4	1.54×10^{-6}
GC(1pb)	N1-H...N3	2.5×10^{-5}
GC(1pc)	N2-H...O2	0.011
GC(1pd)	N1-H...O2	1.75×10^{-5}
GC(1pe)	N2-H...N3	2.73×10^{-3}
GC(1pf)	N3...H-N4	5.61×10^{-6}
	ligações-H duplas	*Fator de Boltzmann*
GC(WC)	O6... H-N4, N1...H-N3 e N2-H...O2	1
GC(rWC)	N1-H...O2 e N2-H...N3	0.30
GC(II)	N3... H-N4 e N2-H...N3	0.00341

Por outro lado, para os pares Guanina com Guanina e Citosina com Citosina, observa-se que houve um acréscimo no número de ligações totais com o critério energético. Assim, ainda considerando o exemplo da série C2, tem-se:

- 211 ligações totais (simples e duplas) entre G e G, para o *critério geométrico*;
- 392 ligações totais (simples e duplas) entre G e G, para o *critério energético*;
- 201 ligações totais (simples e duplas) entre C e C, para o *critério geométrico*;
- 357 ligações totais (simples e duplas) entre C e C, para o *critério energético.*

Verifica-se que o número total de ligações simples formadas com o critério geométrico apresenta valor superior relativamente aquele obtido com o critério energético. Tomando-se como exemplo 80 iterações no gráfico da figura 4.21, tem-se:

- 484 *ligações simples totais* com o *critério geométrico*;
- 17 *ligações simples totais* com o *critério energético*;
- 49 *ligações duplas totais* com o *critério geométrico*;
- 19 *ligações duplas totais* com o *critério energético.*

Além disso, na figura 4.21, observa-se que com o critério geométrico a curva de formação de ligações simples apresenta o seguinte comportamento: no início do processo iterativo há a formação de um número elevado de ligações, seguida

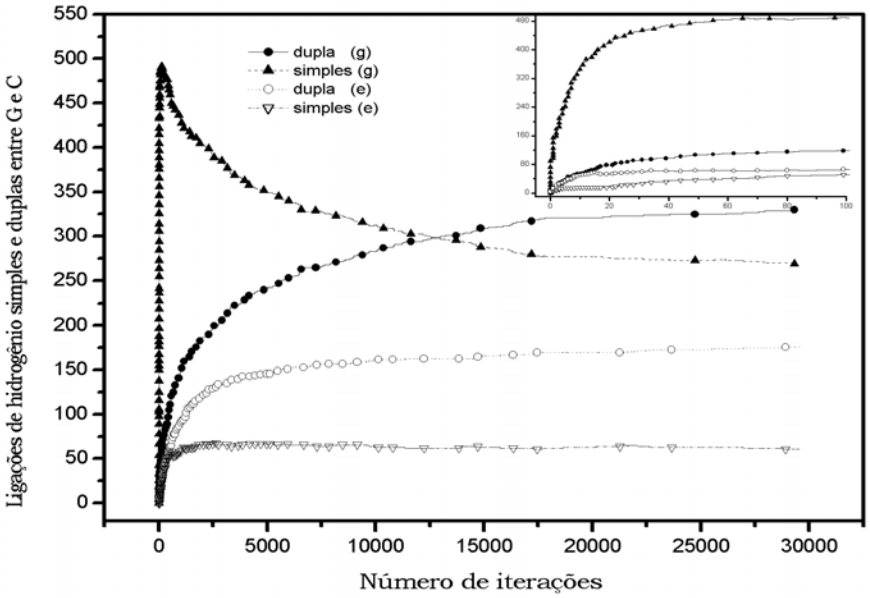

Figura 4.21: Ligações de hidrogênio simples e duplas entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

por um decréscimo desta que contribui para a formação da curva de ligações duplas.

Observa-se que o número de ligações simples quando aplicado com o critério energético, no início do processo iterativo, é relativamente inferior aquele apresentado com o critério geométrico. Este comportamento é justificado pelos baixos fatores de probabilidade que os modelos de ligações simples, entre as bases Guanina e Citosina, possuem.

A figura 4.22 apresenta as curvas de formação de pares de bases em função do número de etapas computacionais dos 6 tipos de ligações simples existentes entre as bases Guanina e Citosina (1pa, 1pb, 1pc, 1pd, 1pe e 1pf) com os critérios geométrico e energético.

Salienta-se que os modelos dominantes, ou seja, os que apresentam a maior probabilidade de ocorrência, com o critério geométrico são *1pa* e *1pc*; o modelo dominante com o critério energético é o tipo *1pc*. Este comportamento, verificado

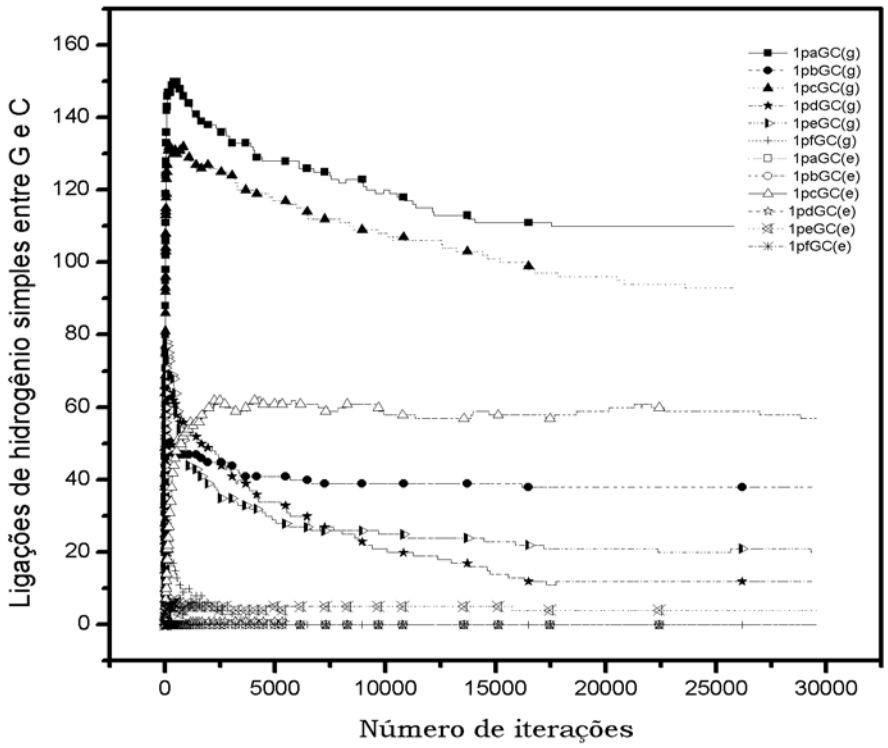

Figura 4.22: Ligações de hidrogênio simples entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

para ligações simples com o critério energético, é esperado pois o modelo 1pc é energeticamente o mais estável (veja tabela 4.5).

Portanto, este apresentará uma freqüência superior aos demais tipos de ligações simples, como observado na figura 4.22. Como os valores dos fatores de Boltzmann, para alguns tipos de ligações simples, entre as bases Guanina e Citosina são muito baixos, a possibilidade de ligações múltiplas envolvendo estes tipos de ligações simples será baixa.

Na figura 4.23, observa-se a formação dos 3 tipos de ligações múltiplas entre as bases Guanina e Citosina com os critérios geométrico e energético, respectivamente. Pode-se verificar que, para ambos critérios, os modelos GCWC e o GCrWC apresentaram a maior freqüência de ocorrência, uma vez que dos três tipos

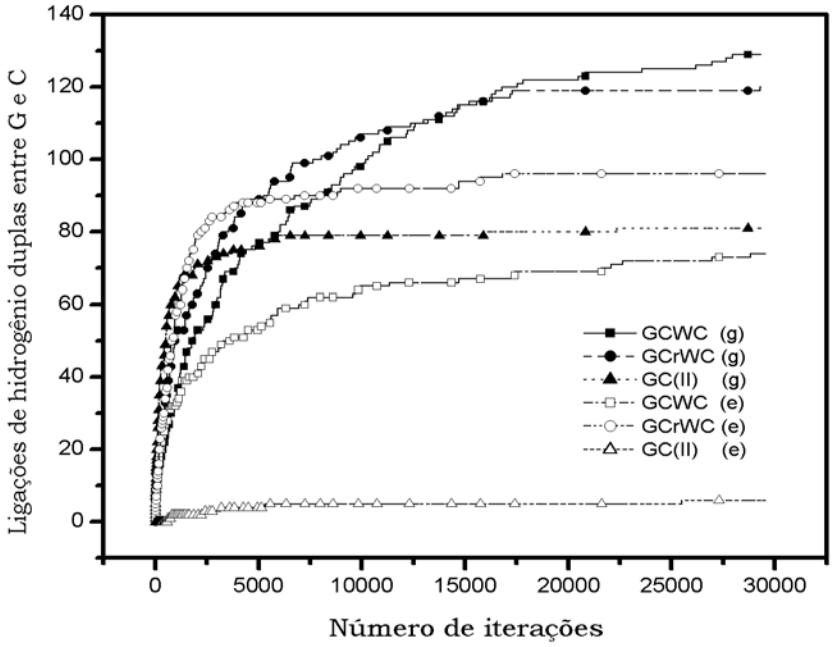

Figura 4.23: Ligações de hidrogênio duplas entre G e C com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

de modelos entre as bases Guanina e Citosina, o GC(II) é energeticamente o menos estável.

4.3.5 Ligações simples e duplas entre as bases Guanina e Guanina

Na tabela 4.6 apresenta-se os valores dos fatores de Boltzmann, obtidos para cada modelo de par de bases entre G e G.

Existem 7 tipos de ligações simples entre as bases Guanina e Guanina. Assim, pode-se verificar uma diferença significativa no número de ligações simples e duplas. Na figura 4.24 verifica-se que a curva de ligações simples, formada com o critério geométrico, apresenta um número de ligações inferior aquela obtida com o critério energético.

Tabela 4.6: Valores dos fatores de Boltzmann entre G e G

Modelos de pares de bases	*ligações-H simples*	*Fator de Boltzmann*
GG(1pa)	O6...H-N1	0.10
GG(1pb)	N1-H...O6	0.40
GG(1pc)	N2-H...N7	9.05×10^{-5}
GG(1pd)	N1-H...N7	2.95×10^{-9}
GG(1pe)	N2-H...O6	1.023×10^{-5}
GG(1pf)	N3...H-N2	1.59×10^{-11}
GG(1pg)	N2-H...N3	1.40×10^{-9}
	ligações-H duplas	*Fator de Boltzmann*
GG(I)	O6... H-N1 e N1-H...O6	1
GG(II)	N1-H...O6 e N2-H...N7	4×10^{-4}
GG(III)	N1-H... N7 e N2-H...O6	8.88×10^{-5}
GG(IV)	N2-H... N3 e N3...H-N2	2.95×10^{-9}

Este resultado é surpreendente, uma vez que nos modelos homogêneos, citados anteriormente, (AA, TT) a curva de formação de ligações simples com o critério geométrico geralmente apresentou um número de ligações superior aquela obtida com o critério energético.

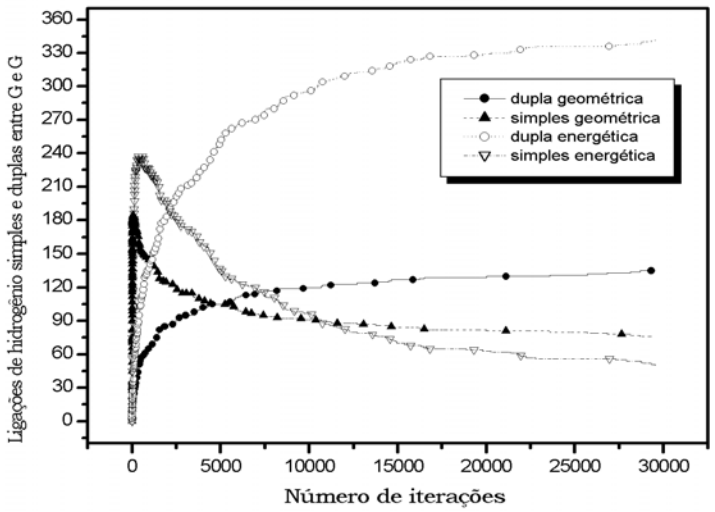

Figura 4.24: Ligações de hidrogênio simples e duplas entre G e G com os critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

O mesmo comportamento pode ser observado para as curvas de ligações duplas entre as bases Guanina e Guanina. Assim, verifica-se, na figura 4.24, que a curva de ligações duplas com o critério energético apresenta um número de ligações superior em relação aquela com o critério geométrico.

Pode-se concluir que os principais fatores que contribuem para o elevado número de ligações simples e duplas para os modelos G e G, com o critério energético, são: o processo intermediário envolvido na formação das ligações de hidrogênio simples e duplas e os valores do fator de Boltzmann associados a cada modelo de par de bases.

Neste sentido, é importante salientar que no processo de formação de ligações envolvendo as bases G e C, existe a possibilidade de formação de pares com as classes G-G, G-C e C-C. Assim, de acordo com os cálculos de energia de estabilização, verificou-se que os valores dos fatores de Boltzmann foram muito baixos para ligações simples em alguns dos modelos das classes G-C e G-G (conforme tabelas 4.5 e 4.6).

Deste modo, os tipos de ligações simples que apresentam fatores de probabilidade baixos provavelmente não ocorrem com o critério energético. A formação de modelos envolvendo estes tipos de ligações simples (com baixo fator de Boltzmann) ocorrerá, somente quando a geometria do par de bases já iniciar diretamente enquanto ligações múltiplas.

Além disso, para os pareamentos entre G-G destacam-se dois tipos de ligações simples com fator de probabilidade viável para a ocorrência de ligações (1pa e 1pb) e, uma vez que estes tipos são precedentes para o modelo de ligação múltipla GG(I), a probabilidade energética de formação de pares envolvendo G e G, com ligações simples e duplas, é elevada. Assim, por exemplo, um modelo de ligação simples G-C formado com o critério geométrico, mas que não constate ligação com o critério energético, pode competir para a ocorrência de outros modelos (G-C, G-G

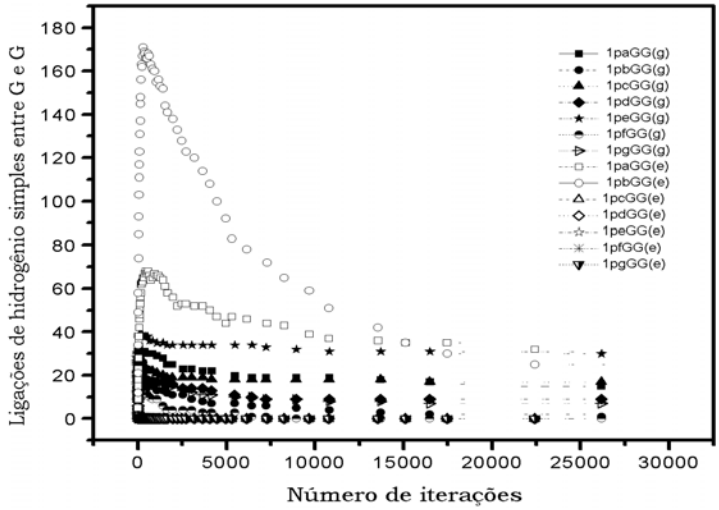

Figura 4.25: Ligações de hidrogênio simples entre G e G com critérios geométrico e energético: experimento C2, (g) geométrico-(e) energético

ou C-C) com ligações simples ou duplas, mas com maior tendência energética na formação para os modelos G-G e C-C.

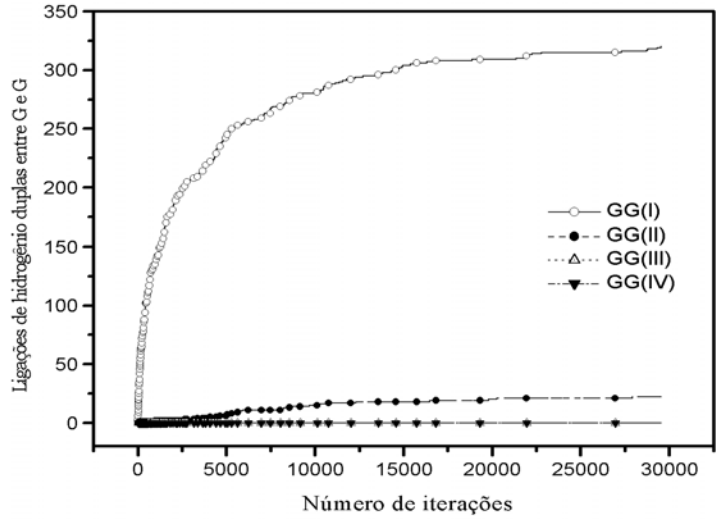

Figura 4.26: Ligações de hidrogênio duplas entre G e G com critério energético: experimento C2.

Assim, a figura 4.25 apresenta o comportamento das curvas de formação para cada um dos modelos de ligações simples entre as bases Guanina e Guanina com os critérios geométrico e energético, respectivamente. Observa-se que os tipos de ligações simples que apresentam a maior probabilidade de ocorrência com o critério energético são os modelos 1pa e 1pb.

Deste modo, como estes tipos são precedentes para o modelo de ligações múltiplas GG(I), conclui-se que com o critério energético este modelo é o que apresentará a maior probabilidade de ocorrência, como se observa na figura 4.26.

4.3.6 Ligações simples e duplas entre as bases Citosina e Citosina

A tabela 4.7 apresenta os valores do fator de Boltzmann para os modelos entre C e C.

Tabela 4.7: Valores dos fatores de Boltzmann entre C e C

Modelos de pares de bases	*ligações-H simples*	*Fator de Boltzmann*
CC (1pa)	N4-H...N3	0.010
CC(1pb)	N3...H-N4	0.010
	ligações-H duplas	*Fator de Boltzmann*
CC	N4-H...H-N3 e N3...H-N4	1

Na figura 4.27 encontra-se a evolução das ligações simples e duplas entre as bases Citosina e Citosina envolvendo os critérios geométrico e energético. Isto permite-nos inferir que o número de ligações duplas com critério energético apresenta um valor superior aquele com o critério geométrico. Por outro lado, observa-se que os modelos de ligações simples com o critério geométrico apresentaram maior número de ligações em relação aquele com o critério energético.

Quanto ao comportamento das ligações duplas entre os critérios tem-se a seguinte justificativa. Sabe-se que entre os pareamentos das bases Citosina e Citosina existe a possibilidade de formação de apenas um modelo de ligação múltipla, sendo este o mais provável como se verifica na figura 4.28. Além disso, os tipos de

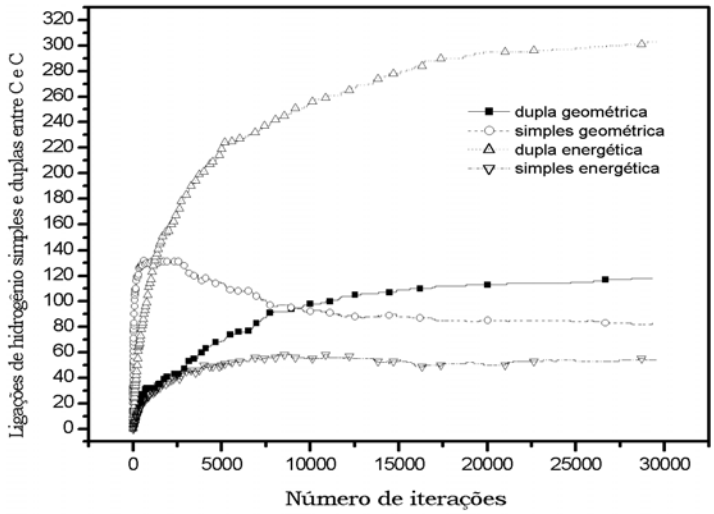

Figura 4.27: Ligações de hidrogênio simples e duplas entre C e C com critérios geométrico e energético: experimento C2.

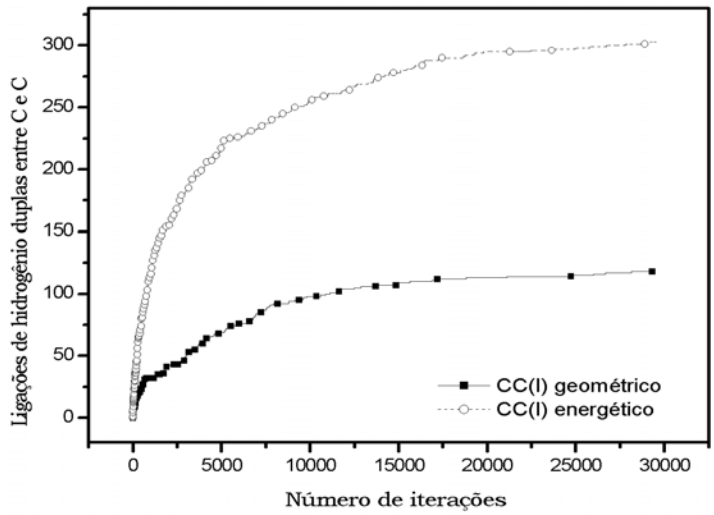

Figura 4.28: Ligações de hidrogênio duplas entre C e C com critérios geométrico e energético: experimento C2.

ligações simples existentes entre as bases Citosina e Citosina são apenas dois, *1pa* e *1pb*.

De um modo geral, destaca-se que existe forte influência do processo intermediário na formação das ligações de hidrogênio simples, bem como da acessibilidade do par quando da verificação das ligações com o critério geométrico. Por outro lado, observa-se que com o critério energético, os modelos que possuem um baixo fator de probabilidade, mesmo que constatem possibilidade de ligações com o critério geométrico, são naturalmente eliminados.

A seguir apresenta-se as principais conclusões bem como as perspectivas deste trabalho.

5 CONCLUSÕES E PERSPECTIVAS

Este capítulo apresenta as principais conclusões resultantes do trabalho realizado, destacando as contribuições, os resultados, as experiências adquiridas e sugestões para desenvolvimento em trabalhos futuros.

5.1 Conclusões dos resultados obtidos

Baseado nos resultados e discussões apresentados nos capítulos anteriores, conclui-se que é possível interpretar, em termos de conceitos simples, a probabilidade da formação de ligações de hidrogênio envolvendo as bases nitrogenadas A, T, G e C. De acordo com os resultados computacionais, verifica-se que há diferenças na atuação dos critérios empregados na simulação dos processos de formação dos pares de bases. Assim descreve-se a seguir as características de cada critério.

Com o critério geométrico destaca-se que a acessibilidade geométrica é fator determinante na formação dos pares (preferência geométrica na formação de alguns modelos) e há existência de certa liberdade na formação das ligações de hidrogênio simples (única restrição geométrica: distância) [16]. Por outro lado, com o critério energético tem-se que as transformações no sistema ocorrem por influências dos fatores energéticos, pela seleção dos modelos mais estáveis.

A seguir conclui-se sobre os resultados obtidos em função de cada critério. Desta maneira, o critério geométrico indica que:

- dos 4 modelos de ligações de hidrogênio múltiplas heterogêneas entre as bases A e T, os que apresentaram maior número de ligações foram ATWC e ATrWC. Os modelos de ATH e ATrH ocorrem em menor quantidade;

- quanto aos modelos homogêneos, tem-se que entre A e A, os modelos AA(I), AA(III) e AA(IV) são os que apresentam maior possibilidade de ligações. Para os pares entre T e T podemos destacar que os modelos TT(I), TT(II) e TT(III) apresentam quantidades de ligações comparavéis.

- dos 3 modelos de ligações múltiplas entre as bases G e C, os que apresentaram maior probabilidade de formação foram GCrWC e GCWC; o modelo de GC(II) ocorre em menor quantidade.

- quanto aos modelos homogêneos, tem-se que dos pares entre G e G, os GG(I) e GG(IV) apresentaram maior quantidade de ligações.

Ainda neste sentido, é importante salientar o comportamento dos modelos com ligações de hidrogênio simples. Pelos resultados observados nas ligações simples e duplas entre as bases nitrogenadas; admite-se assim que uma das principais etapas no processo de formação dos pares de bases esteja associada a formação das ligações simples, uma vez que estas são precedentes dos modelos de ligações múltiplas (duplas).

Por outro lado, na análise dos resultados com o critério energético tem-se que:

- dos quatro tipos existentes entre G e G, o GG(I) foi o que apresentou a maior probabilidade de ocorrência. Este resultado está correto uma vez, que segundo dados da literatura [44], este é o modelo que apresenta a estrutura com maior estabilidade.

- quanto ao modelo entre C e C, verifica-se uma grande probabilidade de ocorrência pois, segundo dados da literatura [44], o valor do fator de probabilidade para este tipo de estrutura é favorável para a sua ocorrência.

- com as estruturas G e C, os resultados mostraram que a maior probabilidade de ligações é dada para os modelos GCWC e GCrWC.

- dos 4 tipos de pares entre A e A, o mais estável é o modelo AA(I). Por outro lado, para os modelos envolvendo as bases T e T, o modelo TT(II) é o mais estável segundo dados da literatura [44]; em nosso experimento a estrutura que apresentou a maior probabilidade de ocorrência foi TT(I). Mas, uma vez que os valores do fator de probabilidade energética para cada um dos 4 modelos com ligações duplas são semelhantes e a diferença no número de ligações entre os modelos é pequena, este comportamento é coerente.

- para a classe entre A e T, dos 4 tipos existentes, o que apresentou a maior probabilidade de ocorrência foi o modelo ATWC. Também, destaca-se que este não é o modelo considerado mais estável. Mas, como estes apresentam valores de energia de estabilização semelhantes, segundo dados da literatura [44], este comportamento é adequado.

Assim, como descrito anteriormente, nosso algoritmo gera uma amostra das estruturas dos modelos mais prováveis dentro de uma determinada classe, de acordo com critérios geométrico e energético (as probabilidades de Boltzmann). Um fato surpreendente que encontramos nos resultados é que a probabilidade de ocorrência para alguns modelos de pares não segue a estrutura mais provável dentro da classe segundo o fator de probabilidade de Boltzmann. Isto sugere que existe uma forte influência da etapa intermediária na formação dos pares (consideração de modelos formados inicialmente com ligações simples).

Com estes resultados pode-se destacar que alguns dos modelos considerados os mais estáveis (que apresentam características energéticas favoráveis a formação do par), não são necessariamente os que apresentam a maior probabilidade de ocorrência.

5.2 Contribuições da tese

Foi implementado um algoritmo para o estudo do processo de formação dos pares de bases considerando critérios geométrico e energético. Assim, salienta-se que:

- o algoritmo implementado viabiliza a manipulação através de índices, tornando o código fonte mais legível com a redução do número de variáveis para manipular a mesma estrutura. Por suas características e estratégias empregadas o algoritmo mostrou-se prático e eficiente, pois não exigiu grande número de iterações. Inúmeros parâmetros de simulação foram analisados para comprovarem o funcionamento e a eficiência do algoritmo proposto. Também destaca-se que o uso de composição de matrizes traz um tratamento mais geral na modelagem das moléculas.

- os resultados obtidos nas simulações com os critérios estabelecidos fornecem a probabilidade de formação para cada modelo dentro de uma determinada classe de par de bases;

- a técnica de Monte Carlo conforme neste trabalho utilizada permite que um sistema grande possa ser modelado num número de configurações aleatórias, e os dados podem ser usados para descrever o sistema como um todo.

- a verificação dos pares de bases considerando como etapa intermediária a ligação de hidrogênio simples (1 ponte de hidrogênio), permite destacar a importância da acessibilidade geométrica na formação dos modelos;

- a análise na formação das ligações com cada critério nos permite distinguir a contribuição e diferença no comportamento de cada um na formação dos pares;

- a relação entre os modelos ditos energeticamente como os mais estáveis e os que apresentaram a maior probabilidade de ocorrência com os critérios aqui estabelecidos;

- o desempenho obtido com o algoritmo indica a importância do emprego em análise de regiões específicas com repetições de seqüências de nucleotídeos.

O presente trabalho mostra que é possível, mediante o algoritmo de colocação e movimentação aleatórias, bem como da codificação da seqüência de etapas definidora, aplicar o Método de Monte Carlo para simular os processos de ligações de hidrogênio para a formação dos pares, considerando os critérios geométrico e energético de forma simples e eficiente.

Além disto, a combinação de ambos os critérios reproduz, na maior parte dos casos, a preponderância daquelas combinações que ocorrem naturalmente, mostrando que estas dependem da acessibilidade geométrica e da energia de interação, como explicitado antes. Assim, devido a variedade estrutural existente entre os modelos de ligações de hidrogênio nos pares de bases, um estudo dos que apresentam a maior probabilidade de ocorrência, numa determinada região de seqüência repetitiva, pode auxiliar no entendimento do comportamento biológico.

5.3 Sugestões para trabalhos futuros

Falar em trabalhos futuros no campo da modelagem molecular é uma tarefa difícil, pois há muito para ser feito. Algumas idéias que surgem são:

- consideração das moléculas do solvente e análise dos resultados com os aqui obtidos.

- análise dos processos de ligações de hidrogênio, também com modelos A-C, A-G, T-C e T-G com critérios geométrico e energético.

- Aprimoramento do algoritmo para a análise de estruturas maiores e mais complexas. Como por exemplo, a análise do processo de formação de pares de bases, considerando a estrutura de um nucleotídeo completo (base nitrogenada, açúcar e fosfato).

- análise da probabilidade de formação destes modelos inseridos na dupla-hélice do DNA, pois os pares isolados são altamente dependentes da variação da seqüência na posição em que se encontram;

Trabalhos completos publicados em Anais e apresentados em Congressos

- ***IV Pan-American Workshop Applied Computational Mathematics***, *Faculdade de Matemática, Astronomia e Física. Universidade de Córdoba, Argentina, 1-5 julho de 2002.*
- ***VII ERMAC - Encontro Regional de Matemática Aplicada e Computacional***, *Anais pp. 279-283, PUC-RS, 20-21 junho de 2002.*
- ***XXV CNMAC - Congresso Nacional de Matemática Aplicada e Computacional***, *Anais p.430, Nova Friburgo -RJ, 16-19 setembro de 2002.*
- ***Formation process of free biological base pairs studied by Monte Carlo, based on geometrical and energetic probabilistic principles, Novel Approaches to the Structure and Dynamics of liquids: Experiments Theories and Simulations***, *Anais p.55, Rhodes-Grece, 7-15 setembro de 2002.*
- ***XV CBECIMAT - Congresso Brasileiro de Engenharia e Ciência dos Materiais***, *Natal-RN, 09-13 novembro de 2002.*
- ***XXVI ENFMAC - Encontro Nacional de Física da Matéria Condensada, Caxambu-MG***, *06-10 maio de 2003.*

Trabalho submetido

Carvalho,S., Samios, D., Bortoli, A. L. de, Justo D. A. R. e Netz, P., Monte Carlo probabilistic method for free biological base pairs formation based on geometrical principles- *Preprint submitted to Applied Numerical Mathematics, Elsevier, 2003* .

REFERÊNCIAS

[1] ALLEN, M.P.; TILDESLEY, D. J. *Computer simulation of liquids.* Oxford: Clarendon Press, 1987.

[2] ALLINGER, N.L.; ZHOU, X.; BERGSMA, J. Molecular mechanics parameters. *J. Mol. Struct.,* Amsterdam, v.312, n. 69, 1994.

[3] ALTMAN, R. Probabilistic structure calculations: A three-dimensional tRNA structure from sequence correlation data. First International Conference on Intelligent Systems for Molecular Biology. National Library of Medicine, Bethesda, 1993.

[4] ARNOTT, S. The geometry of nucleic acids. *Prog. Biophys. Mol. Biol,* Oxford, v. 21, p. 265–319, 1970.

[5] BABCOCK, M. S.; PEDNAULT, E. P. D.; OLSON, W. K. Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Mol. Biol.,* v. 237, n. 1, p. 125–156, 1994.

[6] BABCOCK, M. S.; OLSON, W. K. The effect of mathematics and coordinate system on comparability and dependencies of nucleic acid structure parameters. *J. Mol. Biol.,* v. 237, n. 1, p. 98–124, 1994.

[7] BAUMGAERTNER, A. *Simulations of polymer models in applications of the Monte Carlo.* Berlin: Springer Verlang, 1987.

[8] BINDER, K.; STAUFFER, D. A simple introduction to Monte Carlo simulation and some specialized topics in applications of the Monte Carlo Methods in Statical Physics. *AIAA- American Institute of Aeronautics and Astronautics*, p. 1–9, 2001.

[9] BINDER, K. *Applications of the Monte Carlo method in statical physics.* Berlin: Springer-Verlag, 1986.

[10] Bisch, P. M.; Pascutti, P. G. Modelagem e dinâmica de biomoléculas, 2000. IBCCF(Instituto de Biofísica Carlos Chagas Filho) /UFRJ.

[11] Blattner, F. R. et al. The complete genome sequence of Escherichia coli k-12. *Science* v. 277, p. 1453–1474, 1997.

[12] Brooks III, C. L. Methodological advances in molecular dynamics simulation of biological systems. *Cur. Op. Struc. Biol.* v. 5, p. 211–215, 1995.

[13] CAChe. *CAChe 5.0 for Windows.* Oxford Molecular Limited, sep. 2001.

[14] Calladine, C. R. *Understanding DNA: the molecule and how it works.* Academic Press, Inc, 1997.

[15] Carvalho, S.; Samios, D.; Bortoli, A. L. de; Justo, D. A. R.; Netz, P. Monte Carlo probabilistic method for free biological base pairs formation based on geometrical principles. *Preprint submitted to Applied Numerical Mathematics.*

[16] Carvalho, S.; Samios, D.; Bortoli, A. L. de; Justo, D. A. R.; Netz, P. Formation process of free biological base pairs studied by Monte Carlo, based on geometrical and energetic probabilistic principles. *Novel Approaches to the Structure and Dynamics of liquids: Experiments Theories and Simulations. Anais in Rhodes-Grece 7-15 september*, p. 55, 2002.

[17] Claverie, P.; Caron, F.; Boevue, J. C. *Int. J. Quantum Chem.*, v. 19, p. 229, 1981.

[18] Clowney, L. S. et al. Geometric parameters in nucleic acids: nitrogenous bases. *J. Am. Chem. Soc.*, n. 118, p. 509–518, 1996.

[19] Cornell, W. D. et al. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, n. 117, p. 5179–5197, 1995.

[20] DEL BENE, J. E. Ab initio computation of the enthalpies of some gas-phase hydration reactions. *J. Phys. Chem.*, n. 87, p. 3279, 1983.

[21] DEMIDOVITCH, B. P. *Computational mathematics.* Moscou: Mir, 1987.

[22] DICKERSON, R. E. et al. Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, n. 208, p. 787–791, 1989.

[23] DOKHOLYAN, N. V. et al. Distribution of base pair repeats in coding and noncoding DNA sequences. *Phys. Rev. Lett.*, Woodlrury, v. 79, n. 25, p. 5182–5185, Dec.1997.

[24] DONOHUE, J. Hydrogen-bonded helical configurations of polynucleotides. *Chemistry*, n. 42, p. 61–65, 1955.

[25] DONOHUE, J. Hydrogen-bonded helical configurations of polynucleotides. *Pro. Natn. Acad. Sci.*, n. 42, p. 60–365, 1956.

[26] DONOHUE, J. ; TRUEBLOOD, K. N. Base-pairing in DNA. *J. Mol. Biol.*, n. 2, p. 363–371, 1960.

[27] EBBING, D. D. *Química geral.* São Paulo: MacGraw-Hill, 1996.

[28] EHRLICH, R. *Physics and Computers.* Boston: Mifflin Company, 1973.

[29] ELSTENER, M.; HOBZA, P. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density- functional-theory based treatment. *J. Chem. Phys.*, Melville, v. 114, n. 12, p. 5140–5156, Mar. 2001.

[30] ESCHENMOSER, A. Chemical etiology of nucleic acid structure. *Science*, n. 284, p. 2118, 1999.

[31] FABIOLA, F.; BERTRAM, R.; KOROSTELEV, A.; CHAPMAN, M. An improved hydrogen bond potential: Impact on medium resolution protein structures. *Protein Science*, v. 11, p. 1415–1423, 2002.

[32] FAPESP, N. *Genoma Humano, especial Biologia, a ciência do século XXI.* FAPESP, jun. 1999.

[33] FERRÃO, M. F. *Aplicação do método de Monte Carlo no estudo dos mecanismos de cura co-reativa de resinas epoxi com anidridos dicarboxilicos em presença de amina terciária. 1992.* 226 f.. Dissertação (Mestrado em Química) Instituto de Química, Universidade Federal do Rio Grande do Sul, 1992.

[34] FRENKEL, D.; SMIT, B. Understanding molecular simulation: from algorithms to applications. San Diego: Academic Press, Inc, 1996.

[35] GAVEZZOTTI, A.; FILIPPINI, G. Geometry of the intermolecular X-H...Y(X,Y=N,O) hydrogen bond and the calibration of empirical hydrogen-bond potentials. *J. Phys. Chem.*, n. 98, p. 4831–4837, 1994.

[36] GELBIN, A. et al. Geometric parameters in nucleic acids: Sugar and phosphate constituintes. *J. Am. Chem. Soc.*, n. 118, p.519–529, 1996.

[37] GENDRON, P.; MAJOR, F. Quantitative Analysis of Nucleic Acid Three-dimensional Structures. *J. Mol. Biol.*, v. 308, n. 5, p. 919–936, May. 2001.

[38] GONZALEZ, E. et al. Simulación de la hidratación de los pares de bases adenina-timina por el método de Monte Carlo. *Rev. Mex. Fís.*, México, n. 5, v. 44, p. 473–478, Oct. 1998.

[39] GRIFFITHS, A. J. F.; MILLER, J. H.; SUZUKI, D. T.; LEWONTIN, R. C.; GELBART, W. M. *Introdução à genética.* Rio de Janeiro: Guanabara Koogan, 1997.

[40] GROSBERG, A. YU. ; KHOKHLOV, A. R. *Giant Molecules: here, there, and everywhere...* Academic Press, 1997.

[41] HAMMERSLEY, J. M.; HANDSCOMB, D. C. *Monte Carlo methods.* London: Methuen, 1964.

[42] HERRMANN, H. J.; TSALLIS, C. T. Biogenesis and the growth of DNA-like polymer chains: a computer simulation. *Physica A, Amsterdam,* v.153, n. 2, 202–216, Nov. 1988.

[43] HERSKOWITZ, I. H. *Princípios básicos de genética molecular.* São Paulo: Nacional, 1971.

[44] HOBZA, P. et al. Ability of empirical potentials (AMBER, CHARMM, CVFF, OPLS, Poltev) and semi-empirical quantum chemical methods (AM1, MNDO/M, PM3) to describe H-bonding in DNA base pairs; comparison with *ab initio* results. *Chem. Phys. Lett.*, Amsterdam, v. 257, n. 112, p. 31–35, 1996.

[45] HOBZA, P.; SANDORFY, C. Nonempirical calculations on all the 29 possible DNA base pair. *J. Am. Soc.*, v. 109, n. 2, p. 1302–1307, June 1987.

[46] HOBZA, P.; SPONER, J. Thermodynamic characteristics for the formation of H-bonded DNA base pairs. *Chem. Phys. Lett., Amsterdam,* v. 261, n. 3, p. 379–384, 1996.

[47] HOBZA, P.; SPONER, J.; POSALEK.; M. H-bonded and stacked DNA base pairs: Cytosine Dimer an ab Initio Second-Order Moller Study. *J. Am. Soc.*, v. 117, p. 792–798, 1995.

[48] HOOGSTEEN, K. Structure of a crystal containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallophica* v.12, p. 822–823, 1959.

[49] INDA, M. A. *"Simulação Monte Carlo para o modelo CCA de samios e netz"*1995. 54 f. Dissertação (Mestrado em Matemática), Instituto de Matemática, Universidade Ferderal do Rio Grande do Sul, Porto Alegre, 1995.

[50] IPPOLITO, J.A.; ALEXANDER, R. S.; CRISTIANSON, D. W. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.*, v. 215, p. 457–471, 1990.

[51] KALOS, M. H. *Monte Carlo methods.* New York: John Wiley, 1986.

[52] KLINGLER, T. M.; BRUTLAG, D. L. Detection of Correlations in tRNA Sequences with Structural Implications. *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology,* v. 1, p. 225–233, 1993.

[53] KNEGTEL, R. M.; RULLMANN, J.; BOELENS, R.; KAPTEIN, R. Monty: a Monte Carlo approach to protein DNA recognition. *J. Mol. Biol.,* v. 235, p. 318–324, 1994.

[54] KVICK, A.; KOETZLE, T. F.; THOMAS, R. Hydrogen bond studies. 89. A neutron diffraction study of hydrogen bonding in i-methylthymine. *Chem. Phys.,* v. 61, n. 7, p. 2711–2719, Oct. 1974.

[55] LECOMTE, S.N.; LIN, C. H.; PATEL, D. J. Additional hydrogen bonds and base-pair kinetics in the symetrical AMP-DNA aptamer complex. *Biophysical Journal,* v. 81, p. 3422–3431, 2001.

[56] LEHNINGER, A. L. *Biochemistry.* Worth Publishers, Inc, New York, 1975.

[57] LEMIEUX, S.; MAJOR, F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Research,* v. 30, p. 4250–4263, 2002.

[58] LESZCZYNSKI, J. Computational Study on Hydrogen Bonding and Stacking Interactions Between Nucleic Acid Bases. *SIAM J. Sci. Stat. Comput.,* v. 10, n. 3, p. 581–605, May 1989.

[59] LINDAUER, K. ; BENDIC, C.; SÜHNEL, J. HBexplore - A new tool for identifying and analyzing hydrogen bonding patterns in biological macromolecules. *Computer Applications in the Biosciences,* v. 12, p. 281–289, 1996.

[60] LUSCOMBE, N. M. ; LASKOWSKI, R. A. Amino acid-base interactions: a three-dimensional analysis of protein-dna interactions at an atomic level. *Nucleic Acids Research,* v. 29, n. 13, p. 2860–2874, 2001.

[61] MACKAY, D. H.; HAGLER, A. T. *The role of energy minimization in simulation strategies of biomolecular system, in Prediction of Protein Structure and the Principles of Protein Conformation.* Plenun press, New York, 1989.

[62] MARRA, G.; SCHAR, P. Recognition of DNA alterations by the mismatch repair system. *Biochemical Journal,* London, v. 338, p. 1–13, 1999.

[63] MARTINS, J. B. L. Termodinâmica aplicada a potencial químico, 2003. Disponível na internet: http://www.unb.br/iq/lqc/Joao/extras/.

[64] MAXWELL, E. A. *Methods of plane Projetive Geometry Based on the Use of General Homogeneous Coordinates.* University Press, Cambridge, 1946.

[65] METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.,* New York, v. 21, n. 6, p. 1087–1092, June 1953.

[66] MITRA, R.; PETTIT, B. M.; BLAKE, R. D. Conformational states governing the rates of spontaneous transition mutations. *Biopolymers,* v. 36, p. 169–179, 1995.

[67] MORGON, N. H. Computação em química teórica: informações técnicas. *Quím. Nova, São Paulo,* v. 24, n. 5, p. 676–682, 2001.

[68] MUNDIM, K. C. Modelagem molecular aplicada a sólidos e biomoléculas aplicada a biomoléculas, 2002. IV Escola de Inverno do CBPF, Universidade de Brasília.

[69] NAGASWAMY, U.; VOSS, N.; ZHANG, Z.; FOX, G. E. Database of non-canonical base-pairs found in known RNA structures. *Nucleic Acids Research, Oxford,* v. 8, n. 1, p. 375–376, 2000.

[70] NAGASWAMY, U. et al. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Research, Oxford,* v. 30, n. 1 , p. 395–397, 2002.

[71] NETZ, P. A.; GONZÁLEZ ORTEGA, G. *Fundamentos de físico-química.* Porto Alegre: Artmed, 2002.

[72] NETZ, P. A. *Simulação computacional de processos de reticulação: aplicação do método de Monte Carlo no estudo da reação de resinas epóxi com anidrido e amina terciária como iniciador* 1992. 170 f. Dissertação (mestrado em Química), Instituto de Química, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1992.

[73] NIR, E. Pairing of isolated nucleic-acid bases in the absence of the DNA backbone. *Nature, London,* v. 408, n. 338, p. 21–28, 2000.

[74] OLBY, R. C. *The path to the double helix: the discovery of DNA.* Dover Publications Inc, New York, 1974.

[75] PALADINI, A. *Macromoléculas.* Washington: Union Panamericana, 1968.

[76] POLTEV, V. I.; TEPLUKHIN, A. V.; E CHUPRINA, V. P. Monte Carlo simulation of DNA duplex hydration. B and B' conformations of poly (dA)-poly(dT) have different hydration shells. *J. Biomol. Struct. Dyn.,* v. 6, n. 3, p. 575–586, 1985.

[77] POLTEV, V. I. ; SHULYUPINA, N. V. Hydrogen bonding, stacking and cation binding of DNA bases. *J. Biomol. Struct. Dyn.*, v. 3, p. 739, 1986.

[78] ROGERS, D. F. *Mathematical elements for computer graphics.* New York: McGraw-Hill, 1973.

[79] SAENGER, W. *Principles of Nucleic Acid Structure.* Springer-Verlang, New York, 1984.

[80] SALZBERG, S. L.; SEARLS, D. B.; KASIF, S. *Computational methods in molecular biology. Amsterdam:* Elsevier, 1999.

[81] SANTAMARIA, R. et al. Vibrational spectra of nucleic acid bases and their Watson Crick pair complexes. *J. Comput. Chem.,* v. 20, p. 511–530, 1999.

[82] SANT'ANA, C. M. R. Glossário de termos usados no planejamento de fármacos (recomendações da IUPAC). *Química Nova,* v. 25, n. 3, p. 505–512, 2002.

[83] SARAI, A.; SAITO, M. Theoretical studies on the interaction of proteins with base pairs. Effect of external of H-bond interactions on the stability of Guanine-Cytosine and Non-Watson-Crick Pairs. *International Journal of Quantum Chemistry,* v. 28, p. 399–409, 1985.

[84] SAYLE, R. *RASMOL Molecular Randers (For Linux).* Inc., may 1995.

[85] SCHUTZ, E. A modeling experiment demonstrating H-bonding of Purine and Pyrimidine bases, 1998. Disponível na internet: <http://www. molecules.org/experiments /Schutz/Schutz.html>. Acesso em 16 ago 2002.

[86] SHOWALTER, A. K. ; TSAI, M. D. A DNA Polymerase with specificity for five base pairs. *J. Am. Chem. Soc.,* v. 123, p. 1776–1777, 2001.

[87] SHUSTERMAN, G.P.; SHUSTERMAN, A. J. Teaching chemistry with electron density models. *Journal of Chemical Education, Easton,* v. 74, p. 1–16, 1997.

[88] SINDEN, R. R. Mutagenesis - DNA repair lecture notes, 2001. Laboratory of DNA Structure and Mutagenesis Center for Genome Research.

[89] SOUMPASIS, D. M. ; HUMMER, D. M. A rigorous base-pair oriented description of DNA structures. *J. Biomol. Struct and Dyn.* v. 6, p. 397–420, 1988.

[90] SPONER, J.; FLORIAN, J.; HOBZA, P. Nonplanar DNA Base Pairs. *J. Biomol. Struct Dynam.*, v. 13, 1996.

[91] SPONER, J.; JERZY, L.; HOBZA, P. Structures and energies of hydrogen-bonded DNA base pairs. A Nonempirical Study with Inclusion of Eletron Correlation. *J.Phys. Chem., Washington,* v. 100, n. 2, p. 1965–1974, 1996.

[92] SPONER, J.; HOBZA, P. MP2 and CCSD(T) study on hydrogen bonding; aromatic stacking and nonaromatic stacking. *Chem. Phys. Lett., Amsterdam,* v. 267, n. 314, p. 263–270, 1997.

[93] SPONER, J.; HOBZA, P. Nonempirical ab initio calculations on DNA base pairs. *Chemical Physics, Amsterdam,* v. 204, n. 213, p. 365–372, Apr. 1996.

[94] STRYER, L. *Biochemistry.* W. H. New York: Freeman, 1995.

[95] SUHNEL, J.; LINDAUER, K. Beyond intra-base-pair hydrogen bonds in DNA structures: A comprehensive analysis. *Chemistry,* v. 42, p. 61–65, 1955.

[96] SUZUKI, M. et al. *A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry.* AIST-NIBHT Structural Biology Centre, Jan. 1999.

[97] TINOCO, I. *Structure of base pairs involving at least two hydrogen bonds.* Cold Spring Harbor Press, 1993.

[98] TINOCO, I. *Physical chemistry: principles and applications in biological sciences.* Prentice Hall Inc, 1995.

[99] TSALLIS, C. T.; FERREIRA, R. On the origin of self-replicating information - containing polymers from oligomeric mixtures. *Physics Letters A*, Amsterdam, v. 9, p. 461–463, Dec. 1983.

[100] VAN HOLDE, K.; ZLATANOVA, J. Unusual DNA structures, chromatin and transcription. *Bioessays* v. 1, p. 59–68, 1994.

[101] VAN HOLDE, K. *Principles of physical biochemistry.* Prentice Hall, Upper Saddle River, New Jersey, 1998.

[102] VASCONCELOS, A. T.; RUSSO, C. Análise computacional de bancos de dados genéticos, 2000. IBCCF(Instituto de Biofísica Carlos Chagas Filho) /UFRJ.

[103] WATSON, J. D.; CRICK, F. Molecular structure of Nucleic Acids: A structure for deoxyribose nucleic. *Nature* v.25, n. 4356, p. 737–73 8, Apr. 1960.

[104] WATSON, J.D.; CRICK, F. Genetical implications of the structure of deoxyribonucleic acid. *Nature* v. 30, n. 4361, p. 964–967, May. 1953.

[105] ZAHA, A. *Biologia molecular básica.* Porto Alegre: Mercado Aberto, 1996.

Apêndice A ENERGIA DE ESTABILIZAÇÃO DE LIGAÇÕES DE HIDROGÊNIO DOS PARES DE BASES

Neste apêndice apresenta-se a energia de estabilização dos pares de bases obtidos usando o método semi-empírico AM1

Tabela A.1: Energia de estabilização (em Kcal/mol) dos pares de bases (ligações múltiplas) obtidos pelo método semi-empírico AM1.

Par de bases	*Energia de estabilização*	*Energia relativa*	*Fator de Boltzmann*
TT1	-5.93	0.04	0.94
TT2	-5.97	0	1
TT3	-5.90	0.07	0.88
TT4	-5.93	0.04	0.94
AA1	-3.28	0	1
AA2	-3.00	0.28	0.63
AA3	-2.89	0.39	0.51
AA4	-3.00	0.28	0.62
ATWC	-4.96	0.09	0.85
ATrWC	-4.83	0.22	0.68
ATH	-5.06	0	1
ATrH	-4.97	0.089	0.86
GCWC	-14.48	0	1
GCrWC	-13.78	0.7	0.30
GC(I)	-11.09	3.39	0.03
GG(I)	-15.92	0	1
GG(II)	-11.26	4.66	$4\text{x}10^{-4}$
GG(III)	-10.36	5.56	$8.88\text{x}10^{-5}$
GG(IV)	-4.21	11.71	$2.95\text{x}10^{-9}$
CC	-9.55	0	1

Tabela A.2: Energia de estabilização (em Kcal/mol) dos pares de bases (ligações simples) obtidos pelo método semi-empírico AM1.

Par de bases	*Energia de estabilização*	*Energia relativa*	*Fator de Boltzmann*
TT(1pa)	-5.42	0.55	0.39
TT(1pb)	-5.23	0.74	0.28
TT(1pc)	-4.84	1.13	0.15
TT(1pd)	-4.84	1.13	0.15
AA(1pa)	-1.03	2.24	0.02
AA(1pb)	-1.91	1.37	0.10
AA(1pc)	-1.60	1.68	0.06
AA(1pd)	-1.91	1.37	0.10
AT(1pa)	-3.91	1.14	0.14
AT(1pb)	-3.68	1.37	0.10
AT(1pc)	-3.75	1.30	0.11
AT(1pd)	-3.71	1.34	0.10
GC(1pa)	-6.49	7.98	$1.54\text{x}10^{-6}$
GC(1pb)	-8.17	6.30	$2.59\text{x}10^{-5}$
GC(1pc)	-11.97	2.68	0.011
GC(1pd)	-7.94	6.53	$1.75\text{x}10^{-5}$
GC(1pe)	-10.96	3.52	$2.73\text{x}10^{-3}$
GC(1pf)	-7.27	7.21	$5.61\text{x}10^{-6}$
GG(1pa)	-14.61	1.31	0.10
GG(1pb)	-15.394	0.53	0.40
GG(1pc)	-10.37	5.55	$9.05\text{x}10^{-5}$
GG(1pd)	-4.21	11.71	$2.95\text{x}10^{-9}$
GG(1pe)	-9.07	6.85	$1.02\text{x}10^{-5}$
GG(1pf)	-1.10	14.82	$1.59\text{x}10^{-11}$
GG(1pg)	-3.77	12.15	$1.40\text{x}10^{-9}$
CC(1pa)	-6.82	-2.73	0.010
CC(1pb)	-6.82	-2.73	0.010