Ministério da
Ciência e Tecnologia

INPE-15742-TDI/1487

CARACTERIZAÇÃO COMPUTACIONAL DE PADRÕES ESTRUTURAIS EM SEQUÊNCIAS DE DNA RELACIONADAS A PROCESSOS EM REDES METABÓLICAS

Laurita dos Santos

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Reinaldo Roberto Rosa e Günter J. L. Gerhardt, aprovada em 26 de fevereiro de 2009.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m18@80/2009/02.16.18.57>

INPE
São José dos Campos
2009

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3945-6911/6923
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

Ministério da
Ciência e Tecnologia

INPE-15742-TDI/1487

CARACTERIZAÇÃO COMPUTACIONAL DE PADRÕES ESTRUTURAIS EM SEQUÊNCIAS DE DNA RELACIONADAS A PROCESSOS EM REDES METABÓLICAS

Laurita dos Santos

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Reinaldo Roberto Rosa e Günter J. L. Gerhardt, aprovada em 26 de fevereiro de 2009.

Registro do documento original:
<http://urlib.net/sid.inpe.br/mtc-m18@80/2009/02.16.18.57>

INPE
São José dos Campos
2009

Dados Internacionais de Catalogação na Publicação (CIP)

Santos, Laurita.

S59c Caracterização Computacional de padrões estruturais em sequências de DNA relacionadas a processos em redes metabólicas / Laurita dos Santos. – São José dos Campos : INPE, 2009.

126p. ; (INPE-15742-TDI/1487)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2009.

Orientadores : Dr. Reinaldo Roberto Rosa e Günter J. L. Gerhardt.

1. Computação Aplicada. 2. Sequências Genéticas. 3. Exobiologia. 4. Dendred Fluctuation Analysis (DFA). 5. Análise spectral Gradiente. I.Título.

CDU 004 : 575.112

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada

Dr. Nandamudi Lankalapalli Vijaykumar

Presidente / INPE / SJCampos - SP

Dr. Reinaldo Roberto Rosa

Orientador(a) / INPE / SJCampos - SP

Dr. Gunther J. L. Gerhardt

Orientador(a) / UERGS/UCS / Caxias do Sul - RS

Dr. Ezzat Selim Chalhoub

Membro da Banca / INPE / SJCampos - SP

Dra. Rita Maria Cunha de Almeida

Convidado(a) / UFRGS / Porto Alegre - RS

Aluno (a): Laurita dos Santos

São José dos Campos, 26 de fevereiro de 2009

Até o Fim
Humberto Gessinger

Não vim até aqui pra desistir agora
Entendo você se você quiser ir embora
Não vai ser a primeira vez
Nas últimas 24 horas
Mas eu não vim até aqui pra desistir agora

Minhas raízes estão no ar
Minha casa é qualquer lugar
Se depender de mim eu vou até o fim
Voando sem instrumentos
Ao sabor do vento
Se depender de mim eu vou até o fim

Não vim até aqui pra desistir agora
Entendo você se você quiser ir embora
Não vai ser a primeira vez
Em menos de 24 horas
Mas eu não vim até aqui pra desistir agora

A ilha não se curva noite a dentro vida afora
Toda a vida, o dia inteiro
Não seria exagero
Se depender de mim eu vou até o fim

Cada célula, todo fio de cabelo
Falando assim parece exagero
Mas se depender de mim
Eu vou até o fim

Não vim até aqui pra desistir agora
Não vim até aqui pra desistir

Humberto Gessinger
Cantor e compositor.

A meus pais Tereza e Valny.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, pois sem a crença do Ser Superior e de que Nele tudo posso não teria chegado até este momento.

Em especial minha gratidão aos meus pais, que incondicionalmente me apoiaram nesta árdua caminhada.

À FAPESP (processo $2006/02281-3$) pelo suporte financeiro e por apoiar este projeto. Ao INPE pela oportunidade de fazer parte da Computação Aplicada.

Ao meu orientador Dr. Reinaldo R. Rosa pela amizade, apoio, idéias e auxílio em momentos necessários. Sua participação foi essencial na realização deste trabalho.

Ao meu orientador Dr. Günther J. L. Gerhardt, que me acompanha desde a iniciação científica e que foi fundamental para que estivesse no INPE. Primeiramente pelo incentivo "tu vais sim" e depois por todas as contribuições realizadas neste trabalho e por partilhar todas as suas longas histórias.

Ao professor Dr. Fernando M. Ramos por todo o apoio e crédito na realização deste projeto, aos demais professores da CAP e Dr. Marcus Hauser por idéias que contribuiram para o trabalho.

Agradeço à "família" de amigos que encontrei nesta cidade e neste Instituto. Sem eles tudo teria sido mais difícil. Em especial à Mariana Baroni pela amizade, por nossas conversas arbitrárias, paciência e principalmente contribuições extremamente valiosas na finalização deste trabalho. Ao Eduardo Fávero por toda ajuda (principalmente suporte computacional). Aos meus colegas de sala: Helaine Furtado, Rosângela Bageston, Rudinei Martins e mais recentemente o Francisco por todo apoio, idéias, conversas e momentos de descontração.

Aos demais amigos da CAP: Plínio Ribeiro, Roberta Panzera, Rodolfo Maduro, Sóstenes Gomes, Flávia Mendonça, Juliana Guerra, Thalita Veronese, Ramom Freitas e Murilo Dantas (principalmente por partilhar sua pesquisa). Deixo estes representando todos os colegas que de uma forma ou de outra estiveram presentes, mesmo nas horas não dedicadas aos estudos.

Às secretárias da CAP e do LAC em especial Vanessa Oliveira, Neusa Buto, Claúdia Carraro e Maria Cristina Peloggia por toda ajuda.

E finalmente, à Banca Examinadora pelas sugestões e comentários relacionados a este trabalho.

RESUMO

Nas últimas décadas, uma enorme quantidade de informação sobre o funcionamento de sistemas biológicos foram disponibilizadas em bancos de dados de acesso público. A Computação Aplicada à Biologia ou Bioinformática tem contribuído para análise computacional de dados biológicos cada vez mais ricos em informação. Neste contexto, este trabalho tem por objetivo analisar e caracterizar a estrutura do Ácido Nucléico DNA através de técnicas matemáticas e computacionais. As técnicas de caracterização empregadas são: a análise de flutuação "destendenciada", o coeficiente de dispersão e a análise espectral gradiente. São utilizadas as seqüências gênicas e não gênicas dos seguintes organismos: a *Escherichia coli*, uma bactéria do Reino Eubacteria; a *Thermoplasma acidophilum*, uma arquea do Reino Archaea e a *Saccharomyces cerevisiae*, uma levedura do Reino Fungi. Estes organismos são importantes em estudos de Exobiologia ou Astrobiologia, uma vez que, representam origens evolutivas distintas. Os principais resultados evidenciam diferenças estruturais robustas entre os três organismos e validam as técnicas utilizadas para análise de seqüência genéticas.

CHARACTERIZATION COMPUTATIONAL OF THE STRUCTURAL PATTERNS IN DNA SEQUENCE RELATED WITH METABOLIC PATHWAYS

ABSTRACT

In the last years, an amount of information about the biological systems were available in public databases. The Computer Science Applied to Biology or Bioinformatics has contributed to computational analysis of biological data giving a lot of information on biological processes and patterns in natural systems. In this context, this study aims to examine and characterize the structure of nucleic acid (DNA) through mathematical and computational techniques. The characterization techniques used in this work are: Detrended Fluctuation Analysis, Dispersion Coefficient and Gradient Spectra Analysis. Coding and non-coding sequences of the following organisms are used: *Escherichia coli* which is a bacterium of the Eubacteria Kingdom, the *Thermoplasma acidophilum* which is an archaeal of the Archaea Kingdom and *Saccharomyces cerevisiae* which is a yeast of the Fungi Kingdom. Such organisms are important in the exobiological scenario due to their distinct evolutionary origins. The main results have shown robust structural differences among the three organisms and were important in order to validate the techniques for genetic sequences analysis.

SUMÁRIO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

INPE	–	Instituto Nacional de Pesquisas Espaciais
LAC	–	Laboratório Associado de Computação e Matemática Aplicada
DNA	–	ADN - Ácido Desoxirribonucléico
RNA	–	ARN - Ácido Ribonucléico
DFA	–	*Detrended Fluctuation Analysis*
GSA	–	*Gradient Spectra Analysis*
PCHIP	–	Piecewise Cubic Hermite Interpolating Polynomial
WMA	–	Wavelet Multiresolution Analysis (Análise de Multirresolução por Wavelets)
GPA	–	Gradient Pattern Analysis (Análise de Padrões-Gradiente)

1 INTRODUÇÃO

Nas últimas décadas, uma enorme quantidade de informações sobre o funcionamento de sistemas biológicos foram disponibilizadas em bancos de dados de acesso público. A Computação Aplicada à Biologia (sob os nomes de Bioinformática e Biologia Computacional) tem contribuído tanto na modelagem e simulação, como na análise computacional de dados cada vez mais ricos em informação.

Em particular, a descoberta da estrutura do DNA[1] em 1953 marcou o advento da Genética e o surgimento de novas tecnologias capazes de auxiliar o homem na busca pelo conhecimento da natureza dos organismos vivos (ver Apêndice A).

O conhecimento gerado pela descoberta da estrutura do DNA por James D. Watson e Francis Crick em 1953 foi fundamental para a compreensão da genética dos organismos. Sabe-se que os Ácidos Nucléicos (DNA e RNA [2]) constituem os organismos vivos (incluindo os vírus). Esse material genético por sua vez age conjuntamente com o sistema celular[3] disponível no organismo. Essa ação se torna direta ou indiretamente responsável por todas as características de um ser vivo.

Um aspecto relevante em relação à codificação do DNA em proteínas é o fato de o código genético ser degenerativo, ou seja, a estrutura base do DNA denominada nucleotídeo (adenina (A), guanina (G), timina (T) e citosina (C)) combina-se em trios. Cada trio de nucleotídeo é denominado códon ou triplete sendo que existem 64 possíveis códons ou tripletes dos quatros nucleotídeos. No final do processo de transmissão de informação, o RNA (obtido pela transcrição do DNA) traduz apenas 20 aminoácidos que constituirão todas as proteínas conhecidas dos organismos vivos.

Ainda em relação à codificação de aminoácidos, Trifonov (1999) sugeriu uma possível ordem de aparecimento durante a evolução dos organismos. Para ele, os primeiros aminoácidos que apareceram foram a glicina (Gly), alanina (Ala) e o ácido aspártico (Asp). Miller (1953) realizou um experimento onde simulava a constituição da atmosfera planetária primitiva na tentativa de produzir compostos orgânicos. No

[1] Neste trabalho é adotado a sigla em inglês DNA (*Deoxyribonucleic Acid*) por ser mais usual. A tradução é Ácido Desoxirribonucléico - ADN.

[2] Neste trabalho é adotado a sigla em inglês RNA (*Ribonucleic Acid*) por ser mais usual. A tradução é Ácido Ribonucléico - ARN.

[3] Sistema celular, neste contexto, compreende o conjunto de moléculas e organelas que juntamente com o material genético desempenham as funções necessárias para que as células mantenham-se em atividade.

final do experimento, Miller detectou que haviam se formado os aminoácidos Gly, Ala e Asp. Sabe-se que os códons responsáveis pela tradução desses aminoácidos devem ter o G na primeira posição do códon, o G ou o C na segunda posição e qualquer um dos quatro nucleotídeos na terceira posição do códon. Dessa forma a quantidade de bases nitrogenadas G e C presentes no DNA, denominada conteúdo GC é um padrão importante para as estruturas das seqüências genéticas. Além de estar relacionado à codificação do que se acredita serem os primeiros aminoácidos, o conteúdo GC e consequentemente de Gly são diferentes entre os seres vivos durante a separação dos reinos dos organismos (Eubacteria, Archaea, Protista, Fungi, Plantae e Animalia). Portanto, pode-se utilizar a quantidade de Gly presente em um organismo para inferir o reino ao qual ele pertence (TRIFONOV, 1999).

Esses aspectos fazem parte da estrutura dos Ácidos Nucléicos e compreender essas relações estruturais pode contribuir em diferentes estudos, como por exemplo, aqueles relacionados à evolução dos seres vivos e até mesmo a compreensão de como estabeleceu-se vida no planeta Terra. O interesse relacionado à evolução dos organismos e aparecimento de vida terrestre é um dos objetivos da Exobiologia[4], ciência que se dedica ao estudo da origem, evolução e distribuição de vida na Terra e no Universo. Portanto, considerando o interesse da Exobiologia, este trabalho tem como principal objetivo caracterizar padrões estruturais de diferentes organismos representados pelas suas respectivas seqüências genéticas. Para isso são selecionadas técnicas de caracterização compatíveis com a representação numérica das seqüências genéticas.

1.1 Motivação

Dentro do contexto exobiológico, a caracterização de seqüências de DNA visando a estruturação das regiões gênicas (éxons)[5] e não gênicas (íntrons)[6], fornece informações relevantes que podem contribuir com estudos de filogenia[7] (podendo ser relacionados à Exobiologia) e estudos relacionados à homologia[8] de uma determinada seqüência.

[4]Exobiologia ou Astrobiologia é a ciência que se dedica ao estudo da origem e evolução da vida no planeta Terra. Para detalhamento da Exobiologia ver Apêndice B.

[5]Região gênica ou éxon é a região do DNA que é transcrita para RNA originando uma proteína pelo processo de tradução.

[6]Região não gênica ou íntron é a região do DNA que não é transcrita para RNA.

[7]Termo utilizado para relações evolutivas de um grupo de organismos com a finalidade de determinar as relações ancestrais entre os seres vivos (NELSON; COX, 2004).

[8]Termo utilizado para designar semelhanças entre estruturas de diferentes organismos que podem possuir a mesma origem (LESK, 2007).

Usualmente, a caracterização estrutural entre éxons e íntrons é realizada através das seguintes técnicas. A seguir tem-se alguns exemplos:

- o alinhamento de seqüências, que é usado para inferir a função de uma determinada seqüência baseado em homologia. O alinhamento é uma técnica que consiste basicamente em encontrar seqüências similares aquela que há interesse em classificar, inferindo sua função através da similaridade com outras seqüência (SILVA, 2006);
- a análise de flutuação "destendenciada" (DFA - *dentrended fluctuation analysis*), que fornece leis de potência similares àquelas obtidas através do Espectro de Potência[9] $1/f^{\beta}$ (PENG et al., 1994);
- a análise de similaridade entre genes de diferentes espécies baseada na distribuição de probabilidades de transição de cadeias de Markov (BAI et al., 2007; PODOBNIK et al., 2007).

Em geral, essas técnicas usuais demandam uma grande quantidade de nucleotídeos (na ordem de mais de 10^4 nucleotídeos) para a robustez e eficiência das técnicas. Portanto, há restrição quando o objetivo é analisar seqüências ou partes de seqüências que contenham menos que essa quantidade de nucleotídeos.

Neste contexto, o objetivo principal do trabalho é caracterizar estruturalmente seqüências de DNA dos organismos filogeneticamente diferentes (a bactéria *Escherichia coli*, a arquea *Thermoplasma acidophilum* e a levedura *Saccharomyces cerevisiae*) através de técnicas matemáticas e computacionais verificando a distinção entre regiões gênicas e não gênicas, com restrições de tamanho, de um mesmo organismo e entre os organismos.

Para atingir este objetivo, neste trabalho são usadas técnicas computacionais que permitem verificar a existência de diferenciação estrutural nas seqüências de DNA com poucos nucleotídeos. Duas características fundamentais são verificadas: (i) quando uma seqüência ou parte dela é gênica (responsável pela formação de genes) e (ii) quando a seqüência não é gênica (não há, ou não se conhece, produto transcrito). As técnicas selecionadas para caracterizar as estruturas genéticas, no contexto deste

[9]Espectro de Potência é obtido a partir da Transformada de Fourier, através da qual é possível transformar uma série no domínio do tempo em sua representação equivalente no domínio da freqüência.

trabalho, são as seguintes: (i) o cálculo do coeficiente de dispersão, calculado a partir das seqüências genéticas baseadas em teoria de redes complexas (GERHARDT et al., 2006), (ii) análise de flutuação "destendenciada" (PENG et al., 1994) e (iii) a análise espectral gradiente que classifica uma estrutura pelo seu coeficiente de assimetria gradiente (ROSA et al., 2008).

A caracterização da estruturação do DNA através das técnicas propostas é importante pois pode contribuir para análise de características ligadas as propriedades de redes complexas e assimetrias presentes nas seqüências. Essas características uma vez presentes podem evidenciar as diferenças entre as regiões gênicas e não gênicas em uma seqüência de DNA.

1.2 Organização do texto

Considerando o tema interdisciplinar no qual se insere este trabalho (Computação Aplicada à Biologia), esta dissertação está organizada como segue:

- No Capítulo 2, o detalhamento dos dados e as técnicas de caracterização são descritos;
- No Capítulo 3 são descritas as análises e interpretações oriundas dos resultados obtidos através do emprego das técnicas matemáticas e computacionais de caracterização;
- No Capítulo 4 são apresentadas as conclusões finais e perspectivas futuras objetivando a continuidade desta linha de pesquisa no Laboratório Associado de Computação e Matemática Aplicada - LAC/INPE.

2 DESCRIÇÃO DOS DADOS E TÉCNICAS DE CARACTERIZAÇÃO

2.1 Genomas selecionados

Para verificar o comportamento da estrutura dos Ácidos Nucléicos e se esse comportamento é distingüível entre éxons e íntrons foram selecionados previamente os seguintes genomas[1]: um organismo do Reino Eubacteria, a bactéria *Escherichia coli* (*E. coli*) (Figura 2.1); um organismo do Reino Archaea, a arquea *Thermoplasma acidophilum* (*T. acidophilum*) (Figura 2.2) e um organismo do Reino Fungi, a levedura *Saccharomyces cerevisiae*[2] (*S. cerevisiae*) (Figura 2.3).

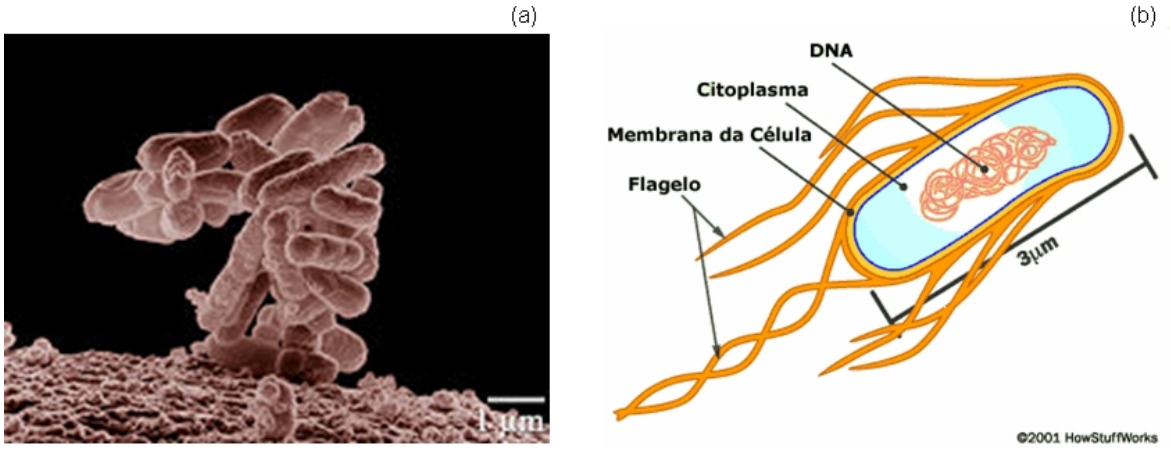


Figura 2.1 - (a) Microscopia eletrônica da *E. coli*, onde cada célula é uma bactéria. (b) Esquema de uma bactéria evidenciando suas partes.

Fonte: (a) http://pt.wikipedia.org/wiki/Escherichia_coli, (b) http://ciencia.hsw.uol.com.br/celulas1.htm.

Esses organismos foram selecionados pois muitas de suas características já são conhecidas e estudadas, permitindo assim melhor compreensão da sua natureza. Neste trabalho, os organismos são representantes de três reinos distintos do ponto de vista evolutivo, desta forma sendo considerados uma distinta amostra filogenética onde pode-se inferir relações no contexto exobiológico. A bactéria e a levedura são consid-

[1]Genoma é o conjunto de todos as seqüências de DNA ou cromossomos presentes nas células dos organismos. Neste contexto, cada organismo é unicelular, sendo que a bactéria e a arquea são constituídas por um genoma de apenas um cromossomo e a levedura possui um genoma com 16 cromossomos no núcleo da célula.

[2]Para uma descrição detalhada das seqüências genéticas deste trabalho ver Apêndice C.

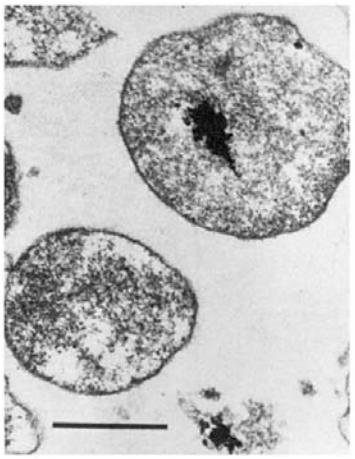


Figura 2.2 - Microscopia eletrônica da *T. acidophilum*.

Fonte: http://cam.bio.ncku.edu.tw/micro/chapter20/.

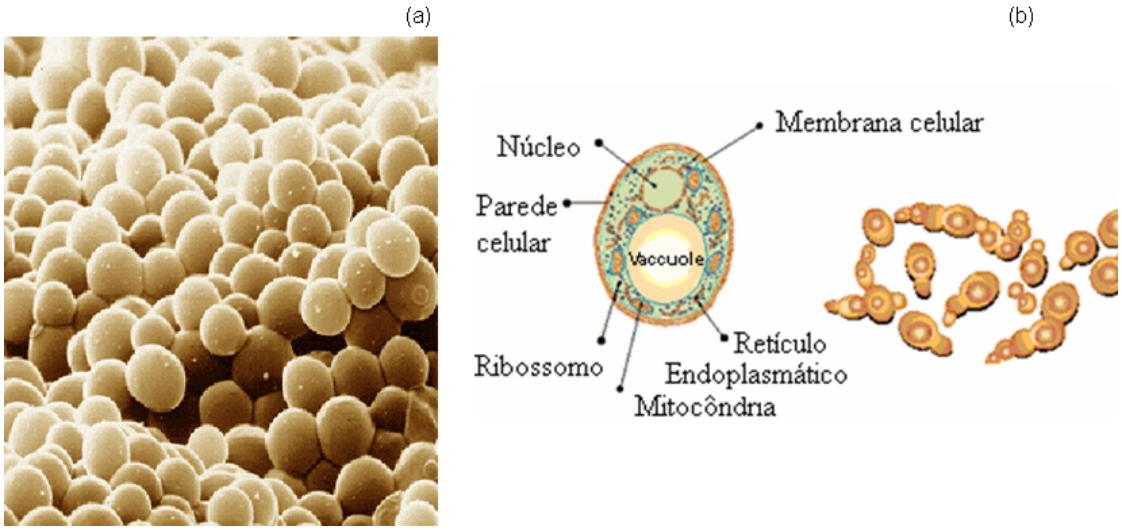


Figura 2.3 - (a) Microscopia eletrônica da *S. cerevisiae*, cada célula é uma levedura. (b) Esquema de uma levedura evidenciando suas partes.

Fonte: (a) http://www.ufrgs.br/alimentus/pao/ingredientes/ing_fermento01.htm, (b) http://www.delaval.com.br/Dairy_Knowledge/EfficientCooling/Why_Cool_Milk.htm

eradas protótipos: uma vez verificado determinado comportamento pode-se inferir características similares em outros procariotos[3] e eucariotos[4].

[3]Procarioto é um organismo cuja célula é desprovida de membrana nuclear.
[4]Eucarioto é um organismo cuja célula possui membrana nuclear.

Os genomas dos organismos escolhidos foram obtidos a partir dos bancos de dados de domínio público que disponibilizam a seqüência de nucleotídeos de uma das fitas do DNA dos organismos (BENSON et al., 2002). Os bancos de dados mais utilizados que disponibilizam os genomas e informações das rotas metabólicas são os seguintes:

- *GenBank*[5];
- *EcoCyc - Encyclopedia of Escherichia coli* $K-12$ *Genes and Metabolism*[6];
- *EcoGene - Database of Escherichia coli Sequence and Function*[7];
- *SGD - Saccharomyces Genome Database*[8];
- *KEGG: Kyoto Encyclopedia of Genes and Genomes*[9].

A *E. coli* e *T. acidophilum* são organismos procariotos, com uma estrutura celular mais simplificada quando comparada à estrutura de um organismo eucarioto (como a levedura). Mas essa aparente simplicidade não pode ser afirmada em relação à estruturação do material genético presente. Portanto, analisar o genoma destes organismos pode contribuir para verificar a estruturação do material genético. No caso da bactéria e arquea tem-se apenas um cromossomo e da levedura tem-se um genoma nuclear constituído por 16 cromossomos.

Analisando o genoma da *E. coli* observa-se um DNA circular presente no citoplasma bacteriano que é responsável por denotar todo o genótipo (conjunto de genes). Portanto, mesmo sendo uma seqüência menor quando comparado a um eucarioto, a bactéria deve produzir todo seu aparato genético. Estima-se que a *E. coli* possua aproximadamente 4000 proteínas (podendo existir uma pequena variação nesse número dependendo da subespécie considerada) (RUDD, 2000). A arquea *T. acidophilum* é um procarioto termoacidófilo[10] com cromossomo circular como as bactérias, podendo ser encontrado em minas de carvão. Esta arquea é considerada um dos menores organismos sequenciados com 1482 proteínas (RUEPP et al., 2000).

[5] $http://www.ncbi.nlm.nih.gov/Genbank/$
[6] $http://ecocyc.org/$
[7] $http://ecogene.org/$
[8] $http://www.yeastgenome.org/$
[9] $http://www.genome.jp/kegg/$
[10] Organismo que possui a capacidade de crescer em altas temperaturas e com baixo pH. Possui crescimento ótimo em torno de 55^oC e 65^oC e pH 2.

A levedura *S. cerevisiae* é um organismo eucarioto cujo genoma é constituído de 16 cromossomos nucleares e 1 cromossomo mitocondrial, possuindo em torno de 6000 genes. Ambos organismos, a bactéria e a levedura, são utilizados para diversos estudos, como por exemplo, verificar a evolução molecular de determinadas moléculas inferindo-se árvores filogenéticas (ISHIGAMI et al., 1996).

A análise das seqüências genéticas desses organismos ocorre através da obtenção da seqüência de nucleotídeos de seus respectivos genomas. No banco de dados GenBank estão disponibilizadas informações a respeito dos genomas, incluindo a seqüência de nucleotídeos propriamente dita. Para avaliar essas seqüências são necessárias transformações dos dados coletados. Neste trabalho, as seqüências de DNA são tratadas de uma duas formas: (i) a partir das seqüências de DNA são obtidos *DNAs walk* e (ii) as seqüências são analisadas como redes complexas. A Figura 2.4 mostra a seqüência da enzima Glucokinase ou Hexokinase da *Saccharomyces cerevisiae* conforme é disponibilizado nos bancos de dados, a partir dela, é realizada as transformações necessárias para análise.

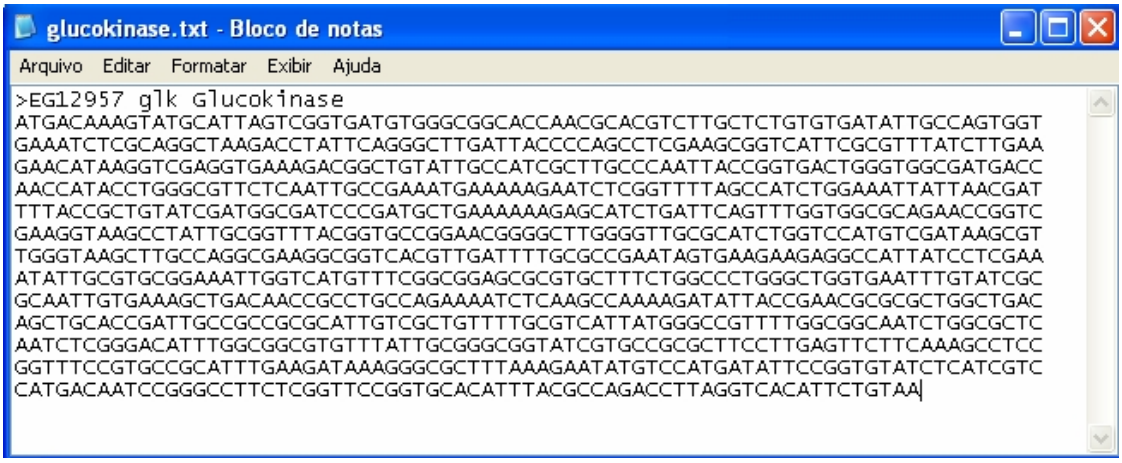


Figura 2.4 - Seqüência de nucleotídeos do gene da enzima Glucokinase ou Hexokinase da *Saccharomyces cerevisiae*.

Inicialmente, um processo importante a ser investigado através da análise da seqüência genética dos organismos é a glicólise, uma vez que, é um processo metabólico comum a todos organismos.

2.1.1 Glicólise

Inicialmente, para caracterizar éxons e íntrons são selecionados genes e isogenes (genes que traduzem[11] enzimas distintas que podem participar da mesma etapa metabólica originando o mesmo produto final) similares entre os organismos escolhidos que participam da primeira etapa metabólica de produção de energia a partir da glicose, a Glicólise. A Glicólise é a primeira etapa do processo metabólico responsável pela produção de energia a partir das moléculas de glicose (carboidrato com 6 átomos de carbono, 12 de hidrogênio e 6 de oxigênio - $C_6H_{12}O_6$). Depois que a glicose entra na célula pelo processo de difusão (sem dispêndio de energia) sofre 10 reações com a participação das seguintes enzimas (GARRETT; GRISHAM, 1999):

- 1^a reação: Enzima Hexokinase;
- 2^a reação: Enzima Fosfoglicoisomerase;
- 3^a reação: Enzima Fosfofrutokinase;
- 4^a reação: Enzima Aldolase;
- 5^a reação: Enzima Triose fosfato isomerase;
- 6^a reação: Enzima Gliceraldeido-3-fosfato dehidrogenase;
- 7^a reação: Enzima Fosfoglicerato kinase;
- 8^a reação: Enzima Fosfoglicerato mutase;
- 9^a reação: Enzima Enolase;
- 10^a reação: Enzima Piruvato kinase.

O nome adotado das enzimas neste trabalho são os mais utilizados, entretanto na literatura podem-se encontrar pequenas alterações no nome dependendo da fonte utilizada. As reações químicas são oriundas da presença do substrato adequado.

Essas enzimas estão em todas as células, mas podem ocorrer em concentrações diferentes dependendo do aparato presente e na especialização celular. A Figura 2.5 ilustra as etapas que a glicose sofre para ser quebrada em duas moléculas de piruvato.

[11] O termo tradução no contexto genético dá-se quando o RNA obtido a partir do gene participa do processo de produção de proteínas.

As enzimas relacionadas a Glicólise são obtidas de seqüências de DNA específicas e conhecidas (GARRETT; GRISHAM, 1999).

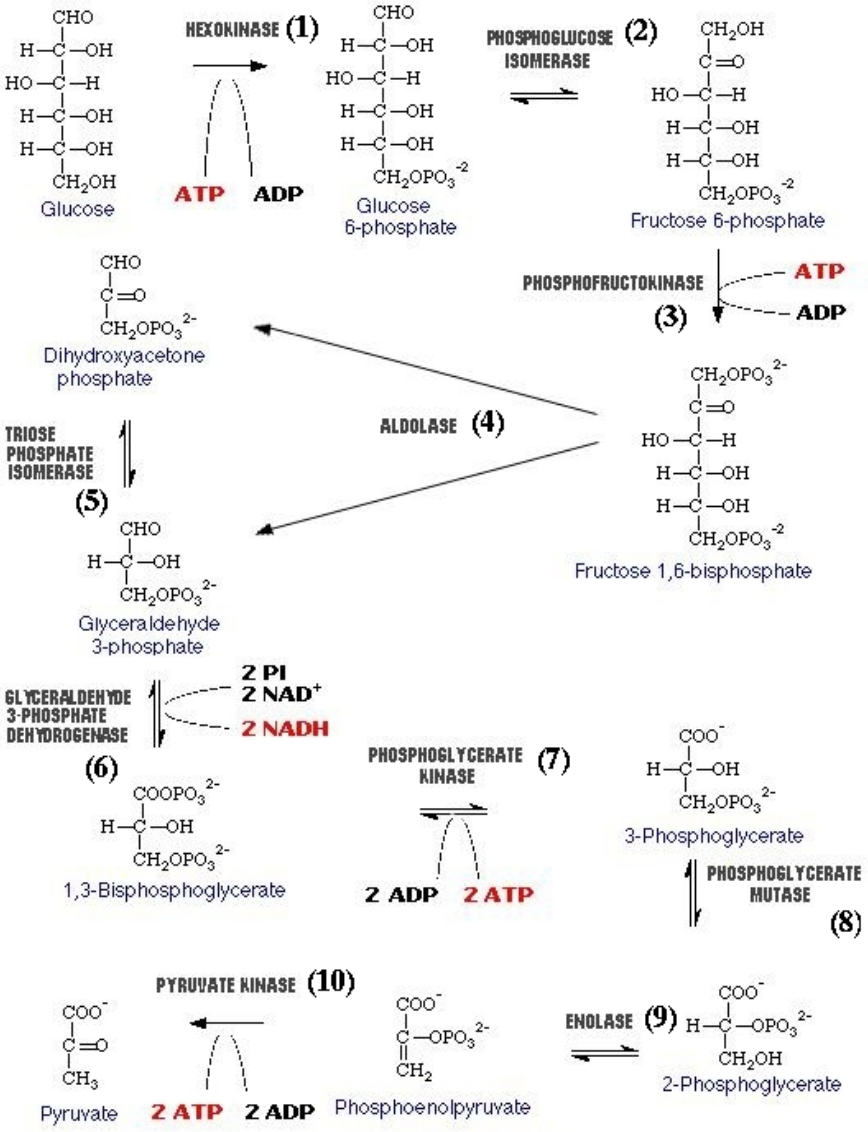


Figura 2.5 - Representação das etapas de degradação da glicose. Os números de 1 até 10 identificam as reações químicas ocorridas em cada etapa e as enzimas responsáveis.

Considerando os genomas selecionados, bem como o processo da Glicólise, são descritas a seguir as três técnicas escolhidas para caracterizar os padrões genéticos estruturais relacionados a cada organismo.

2.2 Técnicas de caracterização

Para caracterização estrutural dos dados descritos na Seção 2.1 são propostas técnicas de caracterização distintas porém complementares em relação à identificação de características de padrões genéticos de cada estrutura.

É importante destacar que as técnicas selecionadas para a análise dos dados de DNA foram utilizadas neste projeto de forma direta como ferramentas previamente implementadas e testadas em diferentes contextos. Para o cálculo da flutuação espectral-gradiente média foram utilizados os códigos escritos em Matlab®, desenvolvidos por Dantas (2008). O código usado para o cálculo do coeficiente de dispersao foi desenvolvido em linguagem C (GERHARDT et al., 2006). Para a análise de flutuação "destendenciada" foi desenvolvido um programa em Matlab® no segundo ano deste projeto (SANTOS, 2008).

A seguir, inicialmente é apresentada a técnica do coeficiente de dispersão D sendo necessária a contextualização de redes complexas (para compreensão das redes de DNA e do coeficiente de aglomeração). A técnica da DFA, proposta por Peng et al. (1994), é usada para analisar seqüências de DNA e possui aplicabilidade em outras séries biológicas como eletrocardiogramas, por exemplo (GUERRA, 2008). Por fim, o conceito de GSA é descrito no contexto das assimetrias estruturais do DNA.

2.2.1 Coeficiente de dispersão

Uma seqüência de DNA pode ser representada por um grafo, determinado por dois conjuntos $G = (v, k)$, onde v é o conjunto de vértices e k é o conjunto de arestas. Essa rede de seqüência de DNA é constituída dos tripletes (vértices) e da justaposição entre dois vértices (arestas) em algum ponto da seqüência linear original, sendo N o número de vértices e L o número de arestas. Como visto na Subseção A.1.1, o DNA possui até 64 tripletes, portanto, sua rede fica restrita a 64 possíveis vértices.

A Figura 2.6 mostra como ocorre a formação dessa rede de DNA. A representação é meramente ilustrativa e verifica-se apenas a justaposição dos tripletes, não observando demais propriedades de um grafo.

O coeficiente de dispersão (D) é uma medida estatística utilizada para determinar a localização de uma seqüência de DNA em relação a outra, gerada para ser o grupo controle, chamada aqui de "pseudo-aleatória", devido a diferentes percentagens de

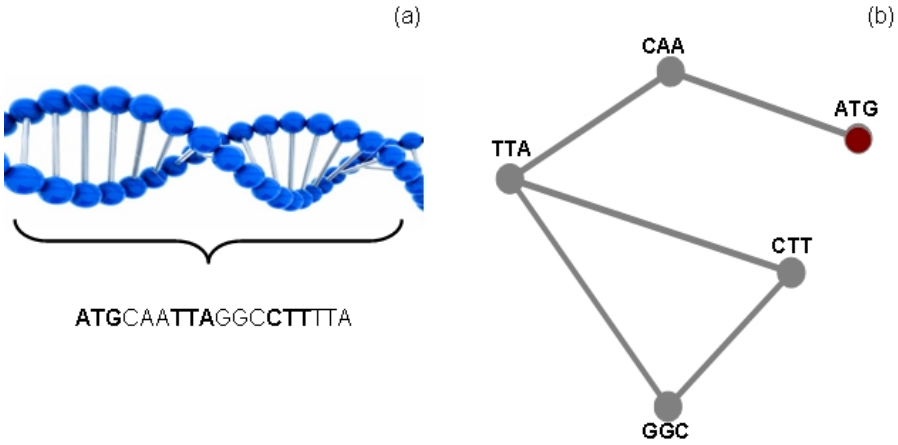


Figura 2.6 - (a) Seqüência de 6 tripletes de uma fita de DNA. (b) Transformação da seqüência (a) em grafo. Cada vértice representa um triplete, sendo o vértice vermelho o primeiro da seqüência.

GC ao longo da seqüência[12]. Essa medida pode ser importante para estabelecer a evolução do DNA de um organismo.

Para obter o coeficiente de dispersão é preciso elaborar uma rede de tripletes de DNA que sirva de grupo controle. Esse grupo de controle, varia seu conteúdo GC ao longo da seqüência até conter somente nucleotídeos G e C. A escolha inicial do tamanho de L ao longo da seqüência é arbitrária, sendo escolhida neste trabalho como $L = 250$, onde uma seqüência de DNA corresponde a $1500pb$ (GERHARDT et al., 2006). Em muitos aspectos o conteúdo GC e a própria seqüência introduzem uma natural organização para a rede e para a aglomeração.

A Figura 2.7 apresenta duas diferentes redes de tripletes, mostrando a natural organização de cada seqüência. Nesse exemplo tem-se os primeiros 450 nucleotídeos da seqüência de DNA de cada um dos organismos utilizados. No primeiro grafo tem-se os nucleotídeos do *Plasmodium falciparum*, que é um protozoário (organismo eucarioto que é encontrado no ambiente) cujo DNA é formado em sua maioria de nucleotídeos A e T, e no segundo grafo tem-se os nucleotídeos do *Thermus thermophilus*, que é um procarioto encontrado em ambientes cuja temperatura é alta e seu DNA é constituído na maioria de nucleotídeos G e C. Essas características podem ser visualizadas nas redes. Neste exemplo não são considerados segmentos similares entre os dois organismos. Arbitrariamente são utilizados os primeiros nu-

[12]No Apêndice E há uma abordagem sobre redes complexas com a finalidade de contextualização para os sistemas biológicos.

cleotídeos dos genomas correspondentes, ilustrando a organização diferenciada entre eles.

A formulação da rede de tripletes é meramente ilustrativa. As ligações não possuem direção e os vértices que estão isolados são aqueles que não estão presentes na seqüência. Nesse caso há o interesse na hierarquia de triângulos ou tricodons formados e na relação com o coeficiente de aglomeração $\bar{C}$, dado por (ALBERT; BARABáSI, 2002):

$$\bar{C} = \frac{1}{L}\sum_{i=1}^{L} c_i \tag{2.1}$$

sendo c_i o coeficiente de aglomeração local dado na Equação E.1.

O coeficiente de dispersão D usado nesse trabalho é definido como (GERHARDT et al., 2006):

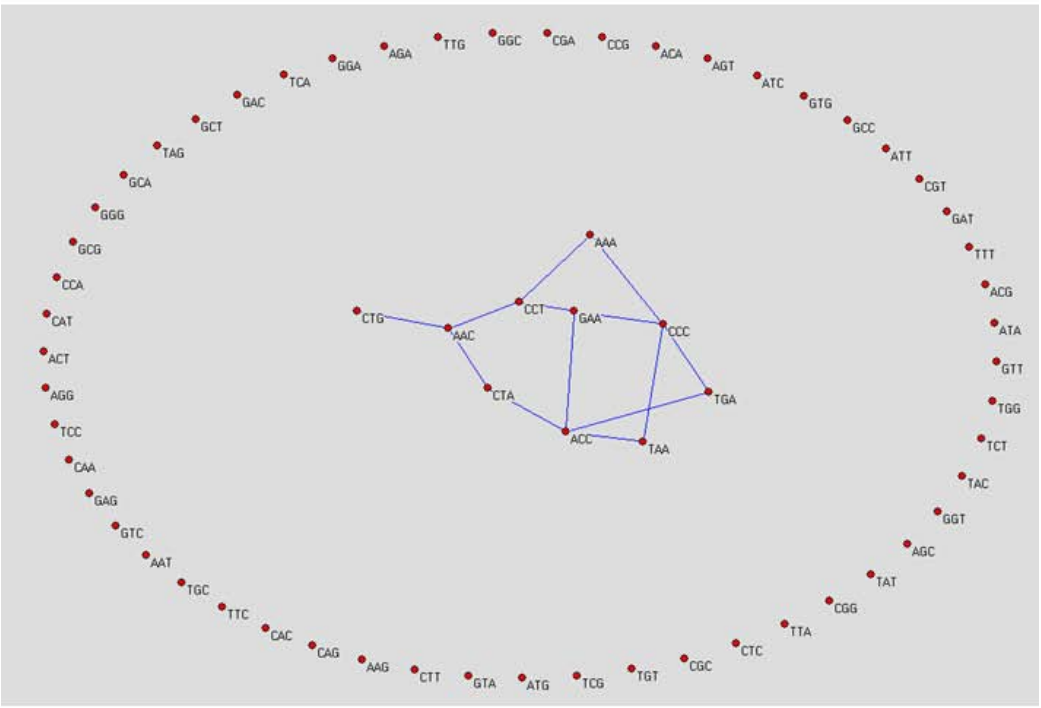
$$D(GC)_L \equiv \frac{1}{n_b}\sum_{i=1}^{n_b} \frac{\left[\bar{C}_i(GC) - \bar{C}_{rand}(GC)\right]}{\sigma_{rand}(GC)} \tag{2.2}$$

onde n_b é o número de janelas de tamanho L da seqüência analisada e $\bar{C}_{rand}$ é o coeficiente de aglomeração médio do grupo de controle. Nesse contexto, σ representa o desvio padrão de $\bar{C}_{rand}(GC)$, que pode ser encontrado numericamente. A Figura 2.8 apresenta a representação da curva obtida com seqüências "pseudo-aleatórias" usando $\bar{C}_{rand}(GC)$ com $L = 250$.

Esse resultado mostra que existem duas regiões de máximo local (GC em torno de 15% e 85%). Estas regiões são decorrência natural do próprio coeficiente de aglomeração e da forma como se gerou as curvas "pseudo-aleatória". Esse comportamento é uma competição entre dois fatores: a densidade de conexão de triângulos, que é natural da medida podendo privilegiar dois pares de bases, e a quebra desse triângulos quando não é mais possivel ter um laço de conexão em função da alta densidade deste mesmo conjunto de pares de bases.

Calculando o σ_{rand} ao longo da seqüência simulada, tem-se o comportamento apresentado na Figura 2.9. Observa-se que nos picos em $\bar{C}_{rand}$ (15% e 85% de GC) o desvio padrão é maior. Para cálculo do σ_{rand} considera-se, nesse caso, intervalos de 10 pontos dos valores de $\bar{C}_{rand}$.

a)



b)

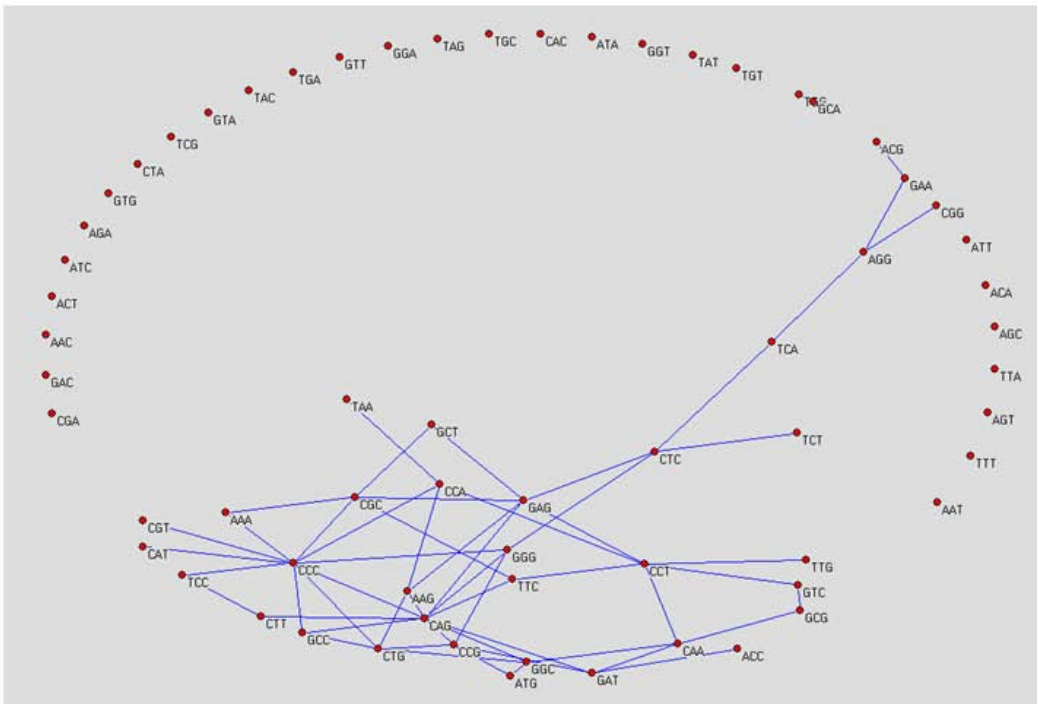


Figura 2.7 - Duas redes de tripletes de diferentes organismos. (a) *Plasmodium falciparum* e (b) *Thermus thermophilus*. Para cada rede foram considerados apenas os primeiros 450 nucletídeos das respectivas seqüências permitindo uma visualização da organização da rede. Nesse caso, os vértices que não possuem arestas são aqueles que não estão presentes nesse trecho do DNA de ambos organismos analisados.

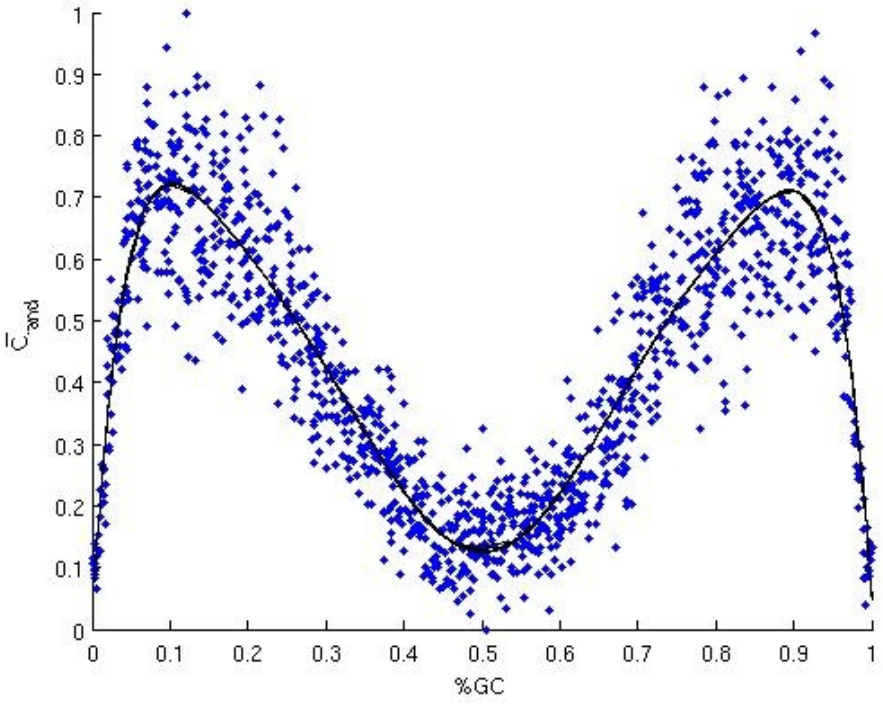


Figura 2.8 - Coeficiente de aglomeração de uma seqüência "pseudo-randômica" com $L = 250$ considerando diferentes percentagens de GC.

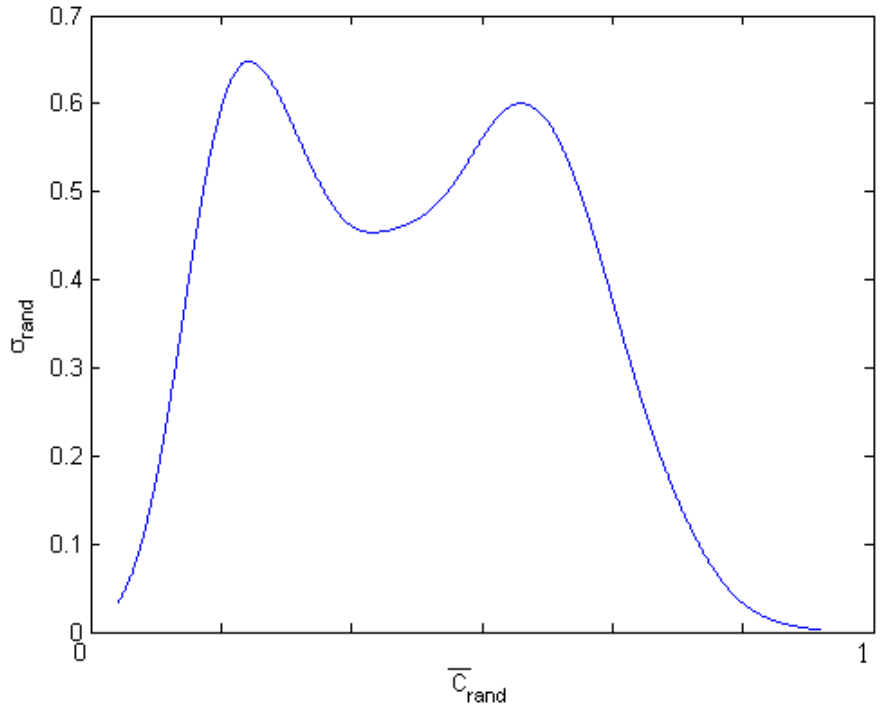


Figura 2.9 - Valores de σ_{rand} ao longo da série considerando intervalos de 10 pontos de $\bar{C}_{rand}$.

2.2.2 Análise da flutuação "destendenciada"

Peng et al. (1994) propuseram analisar seqüências de DNA com uma técnica que verifica a flutuação "destendenciada" do chamado *DNA walk*. Esta técnica é denominada DFA (*Detrended Fluctuation Analysis*) e tem sido aplicada na caracterização de padrões de variabilidade de diversas medidas, na sua maioria séries temporais.

Nesta técnica, inicialmente a seqüência genética é transformada em um caminho de DNA, conhecido como *DNA walk*. O *DNA walk* é obtido substituindo as purinas (A ou G) por -1 e as pirimidinas (T ou C) por 1 (PENG et al., 1992; PENG et al., 1994) (ver Apêndice D), de forma que um perfil de amplitude pseudo-aleatória é gerado pela seguinte fórmula:

$$y(n) = \sum_{i=1}^{n} u(i) \tag{2.3}$$

onde n é o número de nucleotídeos, $u(i) \equiv -1$ para A ou G e $u(i) \equiv 1$ para T ou C. Um exemplo de *DNA walk* é apresentado na Figura 2.10.

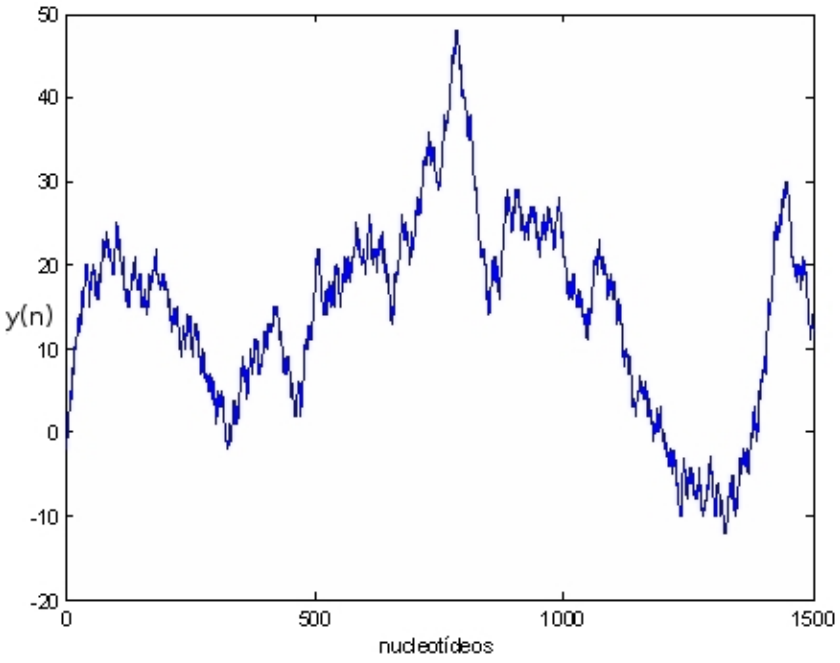


Figura 2.10 - *DNA walk* gerado a partir da seqüência dos primeiros 1500 nucleotídeos gênicos do cromossomo 1 da *S. cerevisiae*.

Segundo Peng et al. (1994) a técnica DFA compreende os seguintes passos:

a) A série $y(n)$ é dividida em escalas l. As escalas l adotadas neste trabalho variam de 4 a 1024 pontos, também utilizada nos trabalhos de Peng et al. (1994). A tendência local de cada segmento de tamanho l é definida através de um ajuste linear por mínimos quadrados (ver Figura 2.11). Essa tendência local é a reta que melhor descreve o conjunto de dados analisados.

b) Calcula-se a diferença entre a seqüência original e a tendência local (ajuste) em todos os pontos (nucleotídeos) para cada escala l. Desta forma tem-se y_l.

c) Calcula-se a variância de y_l sobre cada escala e calcula-se a média dessas variâncias da seguinte maneira:

$$F_d^2(l) \equiv \frac{1}{N} \sum_{n=1}^{N} y_l^2(n) \tag{2.4}$$

onde N é a quantidade de escalas obtidas.

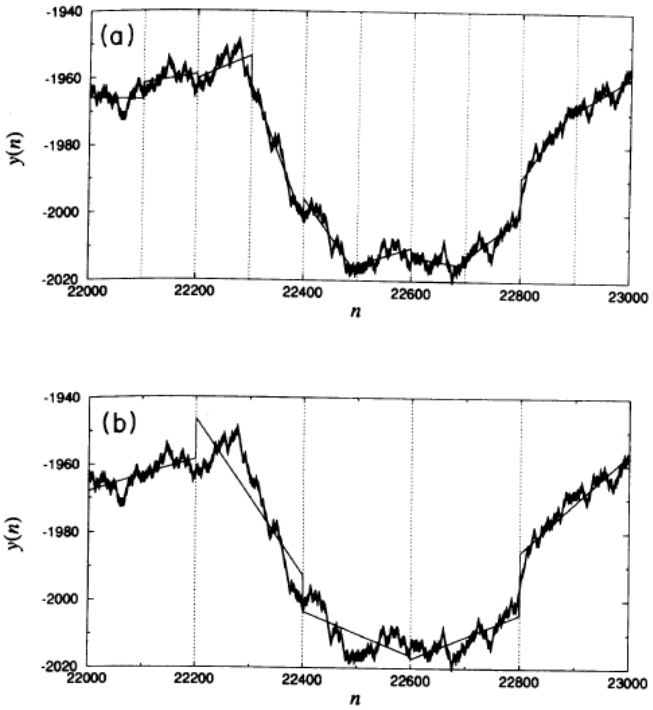


Figura 2.11 - *DNA walk* gerado pela subseqüência do genoma do Bacteriófago λ. O DFA é aplicado em (a) com $l = 100$ e (b) com $l = 200$ (PENG et al., 1994).

Através da DFA, obtém-se um α similar ao obtido com o Espectro de Potência[13] $1/f^{\beta}$ com $\alpha = 2\beta + 1$. A Figura 2.12 apresenta o gráfico $log_{10}l$ versus $log_{10}F_d$ e o valor de α referente a inclinação do ajuste linear correspondente.

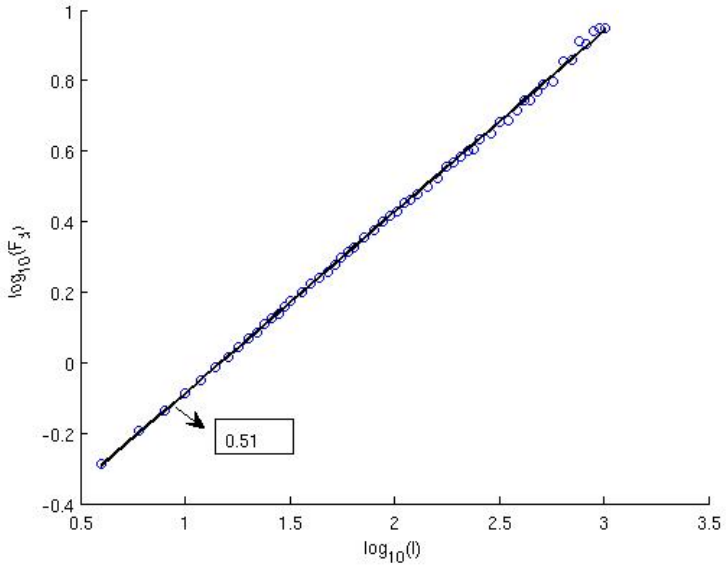


Figura 2.12 - Valor de α obtido a partir, do gráfico $log_{10}l$ versus $log_{10}F_d$, da seqüência do Bacteriófago λ que possui $48502pb$ (PENG et al., 1994).

De acordo com Peng et al. (1994) quando as seqüências apresentam correlações de curto-alcance (ou não apresentam) na série de nucleotídeos, então o caminho de DNA pode ter propriedades de um caminho aleatório (*random walk*) sendo $F_d(l) \sim l^{\frac{1}{2}}$. Entretanto, se há correlações de longo-alcance tem-se $F_d(l) \sim l^{\alpha}$ com $\alpha \neq \frac{1}{2}$.

A principal desvantagem dessa técnica está no fato que para determinação robusta do α, assim como o α do Espectro de Potência, é necessário uma seqüência longa, com $N >> 10^3$ pontos. Essa característica restringe a aplicação dessa técnica em regiões muito pequenas como genes específicos ou regiões não codificantes presentes, por exemplo, em procariotos.

[13]Espectro de Potência é obtido a partir da Transformada de Fourier, através da qual é possível transformar uma série no domínio do tempo em sua representação equivalente no domínio da freqüência.

2.2.3 Análise Espectral Gradiente

Esta técnica, desenvolvida pelo grupo de computação científica do LAC/INPE (ROSA et al., 2008; DANTAS, 2008), analisa o padrão de variabilidade de uma série temporal curta[14]. A Análise Espectral Gradiente[15] (*Gradient Spectral Analysis* - (GSA)), neste contexto, é aplicada em séries espaciais (como são consideradas neste trabalho as seqüências genéticas). Esta técnica baseia-se na Análise de Padrões Gradientes (*Gradient Pattern Analysis*) (GPA) (ROSA et al., 1999; ASSIREU et al., 2002; ROSA et al., 2003) conjuntamente com a Análise Multirresolução por Wavelets (*Wavelets Multiresolution Analysis*) (WMA) (HAGELBERG; GAMAGE, 1994; MALLAT, 1989), permitindo a caracterização do padrão de variabilidade de uma série. Enquanto a GPA tem seus princípios fundamentados na álgebra matricial e na geometria convexa, a WMA fundamenta-se na análise funcional (GUERRA, 2008)

Basicamente, a GSA consiste de uma seqüência ordenada de quatro operações em uma série temporal ou espacial $\{A_i\}_N$ que representa um conjunto de N medidas discretas da amplitude de uma variável genérica $A(i)$:

- determinação da escala de máxima coerência da série analisada;
- representação da multirresolução da série;
- cálculo da potência do coeficiente de assimetria;
- obtenção do espectro-gradiente e da medida de flutuação do espectro-gradiente.

A escala de máxima coerência (λ_{mc}) é obtida através do ponto de inflexão (derivada nula) em um gráfico da variância da ondeleta versus a escala de dilatação a[16]: $var(W_\Psi) \times a$ (DANTAS, 2008). Trata-se da variância de uma transformada ondeleta,

[14]Segundo Dantas (2008), uma série temporal é considerada série temporal curta quando o número de medidas (pontos da série) for $N \sim 10^3$ pontos.

[15]Esta técnica é recente e concatena várias etapas. Portanto, o Apêndice F possui informações complementares.

[16]A dilatação a diz respeito a janela de tempo que varia para wavelets. Isso ocorre porque wavelets possui um conjunto infinito de possíveis funções-base, permitindo a flexibilidade na representação do domínio tempo. Esta flexibilidade é possível dada a variáveis que permitem controlar a onda geradora (wavelet-mãe) para que altere sua escala de freqüência através de uma dilatação denominada dilatação a.

$W_\Psi[A(t)]$ como função das possíveis escalas de dilatação, a, associadas à ondeleta-mãe $\Psi(t - b/a)$. Em geral, essas escalas expressam a existência de estruturas coerentes associadas aos processos dinâmicos inerentes ao sistema analisado. No caso da seqüência de DNA essa propriedade representa repetições de longa distância nas distribuições dos nucleotídeos ao longo da seqüência. A Figura 2.13 apresenta um exemplo de λ_{mc} para uma seqüência gênica escolhida arbitrariamente. Neste caso, a escala de máxima coerência é correspondente a escala saturada no periodograma. O valor da escala para este exemplo é $\lambda_{mc} = 400$ pontos caracterizando a maior coerência da série para duas regiões de mínima amplitude da série.

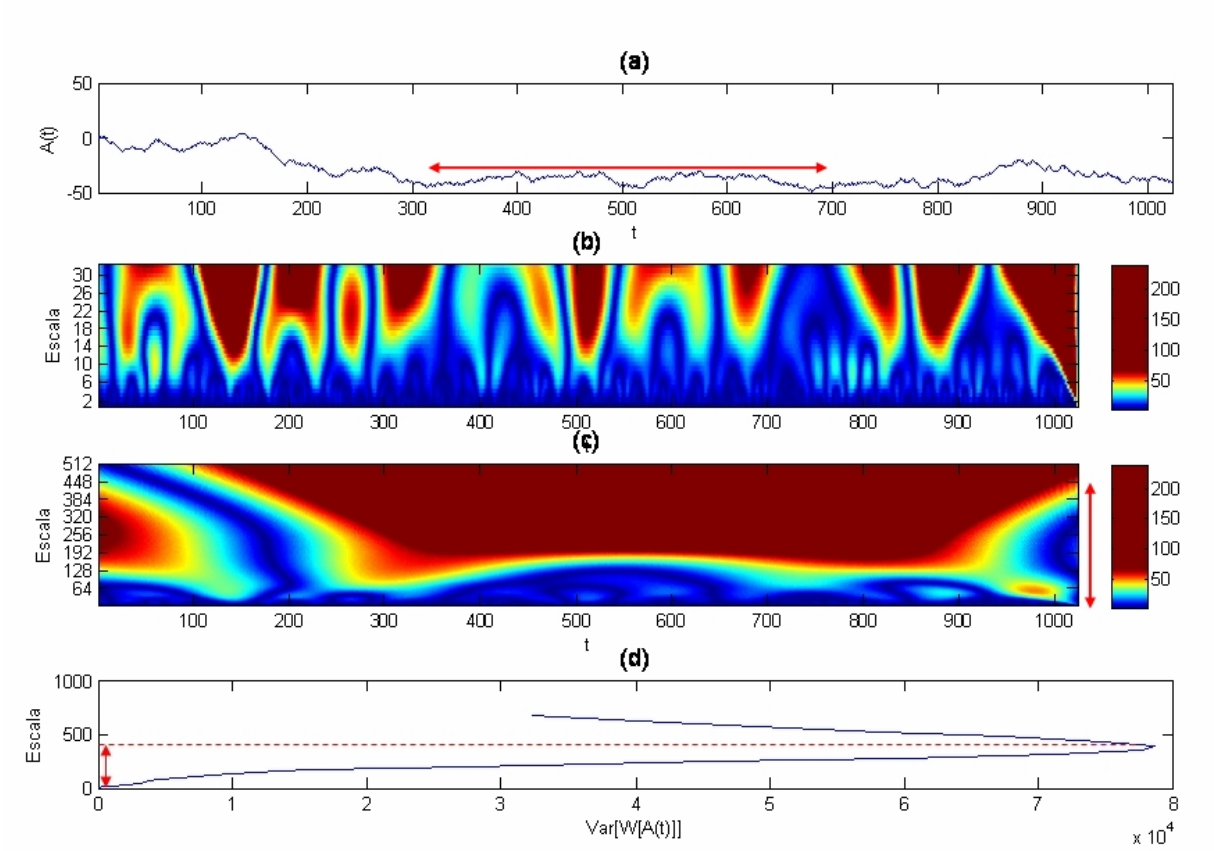


Figura 2.13 - Cálculo da λ_{mc} para um exemplo de série gênica. A escala obtida é de 400 pontos. a) seqüência analisada com destaque para a escala obtida, b) periodograma da seqüência rica em escalas, c) periodograma mostrando a escala saturada e d) escala onde a variância é máxima. Na Figura (a), $A(t)$ representa $y(n)$.

A representação da multirresolução consiste da decomposição e reconstrução da série

temporal ou espacial por meio de uma ondeleta que mantenha as características estruturais do sinal em todas as suas componentes ω_j. O número possível de decomposições é diretamente associado com a ondeleta-mãe, que deve ser escolhida para revelar da melhor forma a estrutura do sinal em consideração. Para escalas de variabilidade muito curtas sob modulação não-linear, a ondeleta-mãe mais estável é a biortogonal. Portanto, utilizando um algoritmo para a transformada discreta biortogonal (DANTAS, 2008), neste caso, a ondeleta "bior 6.8". Primeiramente são obtidos os componentes de aproximação. Nesta abordagem os coeficientes da ondeleta biortogonal têm valores discretos, em que as classes de decomposição e reconstrução seguem uma escala diádica. A Figura 2.14 mostra a saída do algoritmo aplicado na seqüência gênica composta por 1024 pontos. Dessa forma, todas as componentes de escala para a seqüência original são obtidas, a partir da aproximação completa biortogonal para o padrão típico de variabilidade desta estrutura. Para esse exemplo, são suficientes 5 componentes. Para cada componente da seqüência é calculado o coeficiente de assimetria.

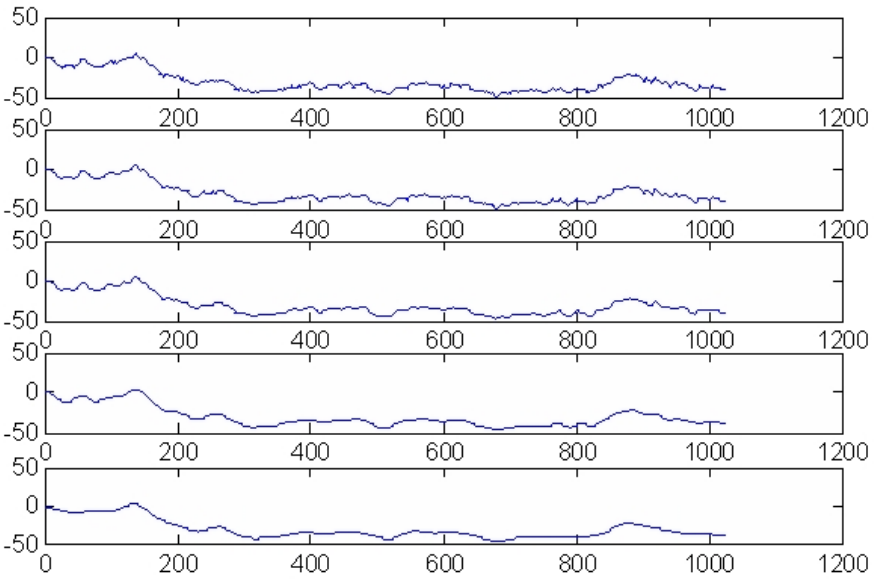


Figura 2.14 - Componentes de aproximação biortogonal para um exemplo de seqüência gênica.

A terceira operação da GSA, baseada na GPA, é o cálculo o coeficiente de assimetria G_A, na escala λ_{mc} aproximada, para cada componente reconstruído. O coeficiente

de assimetria G_A é definido como:

$$G_A = \frac{N_c - N_v}{N_v} \tag{2.5}$$

onde N_v é o número total de vetores assimétricos[17] remanescentes após a remoção dos pares simétricos e N_c é o número de conexões entre esses vetores (ROSA et al., 1999; BARONI et al., 2009). A Figura 2.15 ilustra como é realizado o cálculo do coeficiente de assimetria. Nesta Figura, o primeiro passo representa a aproximação do valor da λ_{mc} e a seqüência janelada (para obter uma matriz quadrada). O segundo passo representa cada janela da seqüência dividida em segmentos iguais para elaboração da matriz quadrada. O terceiro passo representa o cálculo do gradiente da matriz quadrada e a triangulação de Delaunay. Ambas operações são necessárias para o cálculo do coeficiente de assimetria.

Na última operação é calculado a potência do coeficiente de assimetria $G_{POT} \times \omega_j$, definido como:

$$G_{POT} =< G_{A,\omega_j} >^{\lambda_{mc}} \tag{2.6}$$

O espectro-gradiente é obtido interpolando, por ajuste não-linear, os valores obtidos nos gráfico $G_A(\omega) \times \omega$ (ROSA et al., 2008; DANTAS, 2008). Para essa reamostragem o espectro-gradiente é interpolado com mil pontos[18] entre os valores de G_{POT}. Através de análises empíricas a interpolação escolhida é a interpolação cúbica *P-chip*[19] gerando resultados mais robustos. A Figura 2.16 apresenta um exemplo de espectro-gradiente interpolado para uma seqüência de $1024pb$. A partir do espectro-gradiente interpolado obtém-se a flutuação espectral gradiente média, definida por:

$$f_{eg} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_{POT} - < G_{POT} >)^2} \tag{2.7}$$

sendo N a quantidade de componentes de aproximação ω_j para cada sinal.

A Figura 2.17 apresenta a flutuação do espectro-gradiente para os genes da Glicólise (que tem uma flutuação menor) e as regiões não gênicas que possuem uma flutuação

[17] Neste contexto, vetores assimétricos são aqueles que possuem mesma fase e mesmo módulo.

[18] De acordo com Dantas (2008) a medida f_{eg} converge para reamostragens contendo mais que 120 pontos.

[19] *P-chip* é a função de interpolação cúbica de Hermite (função presente no software Matlab®).

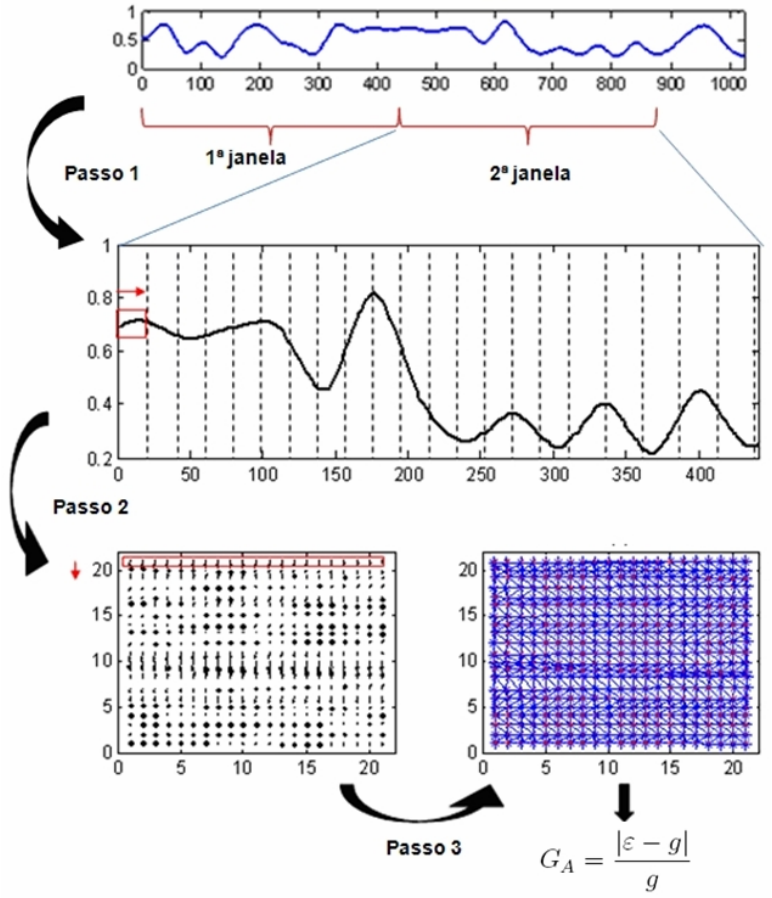


Figura 2.15 - Cálculo do G_A para apenas um fragmento de uma componente de aproximação. Passo 1: janelamento da série no qual é transformado em matrizes quadradas. Passo 2: para cada matriz quadrada (valor obtido pela λ_{mc}) é verificado o gradiente e passo 3: campo de triangulação de Delaunay para os vetores assimétricos (DANTAS, 2008).

maior.

Aplicações em séries temporais canônicas mostram que a GSA serve como ferramenta complementar na análise de padrões de variabilidade temporal irregulares, intermitentes e não-estacionários, comumente gerados por processos não-lineares (DANTAS, 2008). De forma mais contundente é enfatizado a importância da GSA como uma das poucas metodologias robustas para a análise de séries temporais ou espaciais curtas, isto é, aquelas compostas por uma quantidade parcial de medidas capaz de comprometer o desempenho das análises estatísticas convencionais. As quais, em geral, definem medidas que convergem apenas para grandes amostras, dado que o valor esperado sempre envolve medidas de desvios em relação à média estatística da amostra.

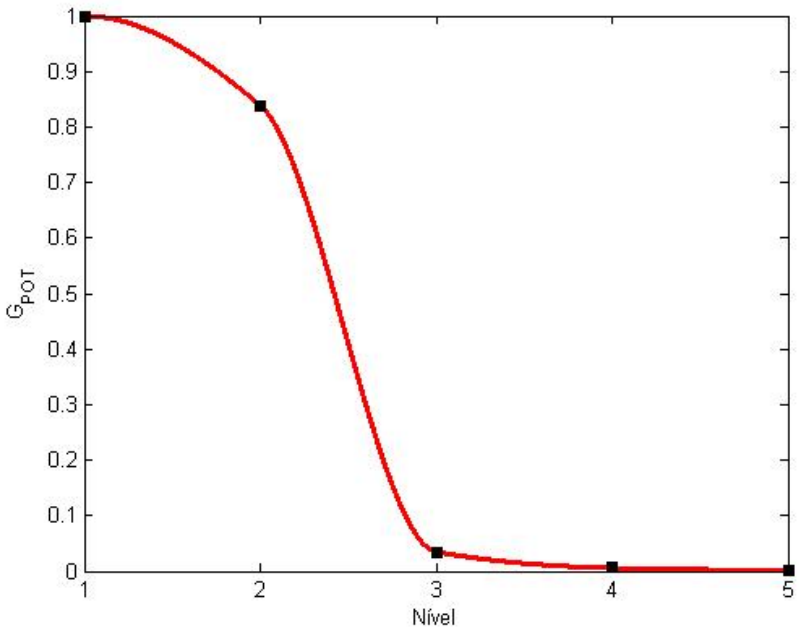


Figura 2.16 - Espectro-gradiente médio normalizado (G'_{POT}) obtido a partir de uma seqüência gênica.

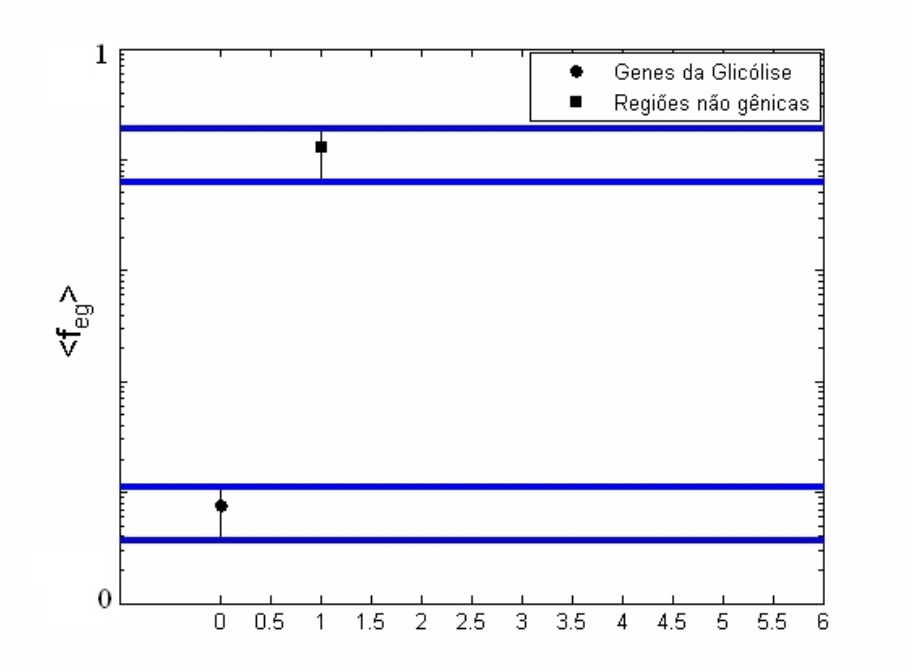


Figura 2.17 - Flutuação do espectro-gradiente para os genes da Glicólise e regiões não gênicas da *E. coli*.

A Figura 2.18 apresenta as etapas para a aplicação da GSA nas seqüências genéticas. As etapas são descritas a seguir:

1 utilizando informações dos bancos de dados são selecionadas seqüências genéticas ou parte de seqüências. São usadas as fitas do DNA que possuem os nucleotídeos transcritos em RNA;

2 através da seqüência de DNA selecionada é gerado o *DNA walk*, conforme já descrito neste Capitulo;

3 o cálculo da escala de máxima coerência λ_{mc} é obtido do *DNA walk*. Este valor será usado em dois momentos diferentes na GSA: (i) para aproximar a janela usada para o cálculo G_A e (ii) para calcular G_{POT};

4 a seqüência de DNA é decomposta e reconstruída através da técnica da multirresolução por ondeletas (HAGELBERG; GAMAGE, 1994). O número possível de decomposições é diretamente associado com a ondeleta-mãe, que deve ser escolhida para revelar da melhor forma a estrutura da seqüência analisada;

5 cada componente da seqüência original é transformada em uma matriz quadrada que é utilizada para o cálculo do coeficiente de assimetria ($G_{A,n}$);

6 é calculado G_{POT} para cada componente. O espectro-gradiente é obtido pela interpolação não linear do G_{POT} (*p-chip*);

7 a flutuação espectral-gradiente média (f_{eg}) é calculada e o gráfico obtido apresenta $< f_{eg} >$ com sua respectiva faixa dinâmica de variação.

O próximo Capítulo apresenta os resultados obtidos da aplicação das técnicas de caracterização descritas neste Capítulo nos genomas selecionados.

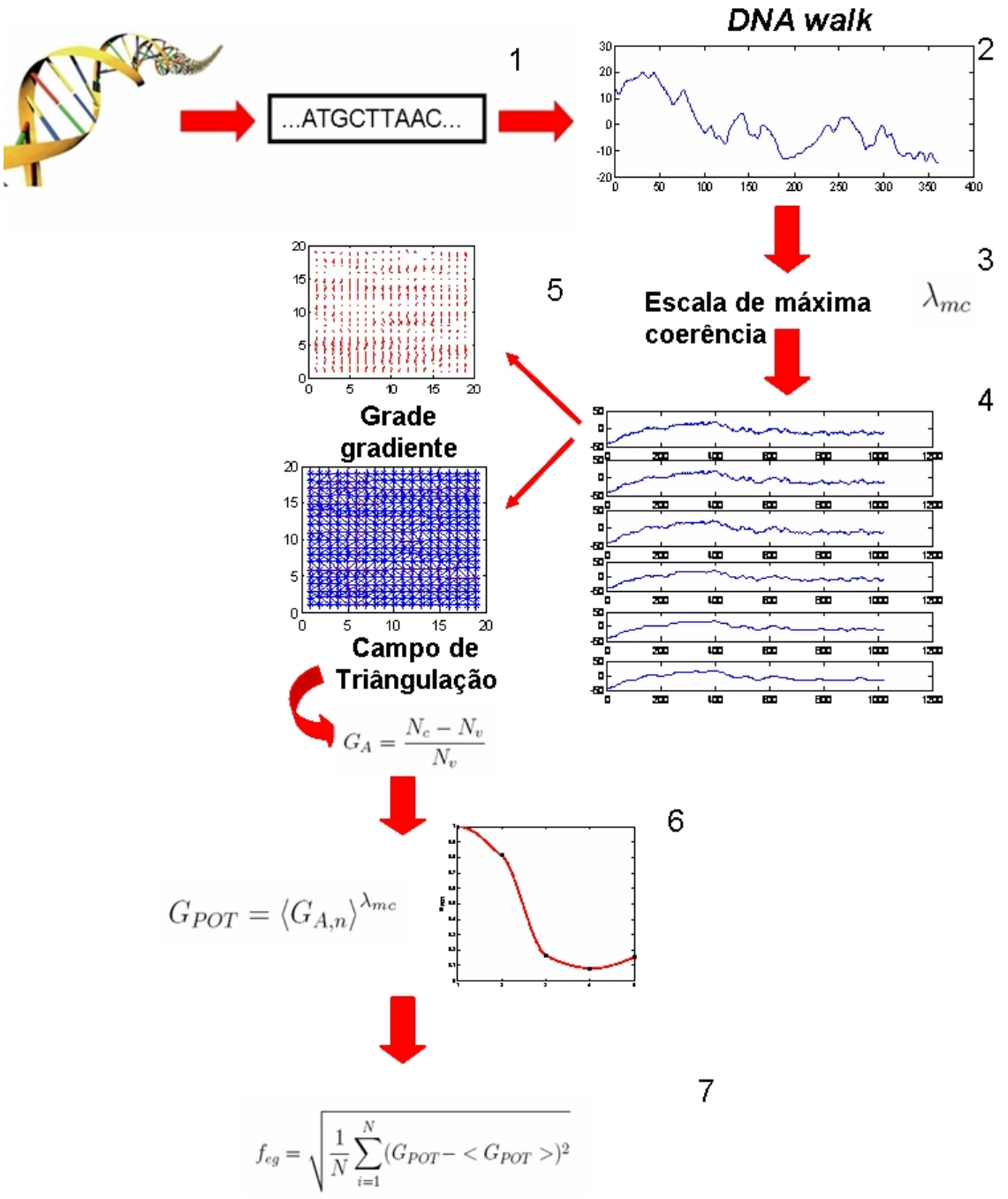


Figura 2.18 - Etapas da técnica GSA para as seqüências genéticas. Os números de 1 a 7 são os passos da técnica. 1 obtenção da seqüência desejada, 2 geração do *DNA walk* para a seqüência, 3 cálculo da escala de máxima coerência, 4 componentes da seqüência genética obtidas pela decomposição e reconstrução da série, 5 cálculo do coeficiente de assimetria, 6 cálculo do G_{POT} e 7 cálculo da flutuação espectral-gradiente média.

3 Análise e Interpretação dos Resultados

Neste Capítulo são apresentados os resultados e interpretações obtidos a partir da análise dos dados, utilizando as técnicas de caracterização descritas no Capítulo anterior. Os resultados são apresentados na seguinte organização:

- a primeira parte apresenta a análise e caracterização para a classificação comparativa entre seqüências não codificantes da *E. coli*, *T. acidophilum* e *S. cerevisiae* (Seção 3.1);
- a segunda parte apresenta a análise e caracterização dos genes da Glicólise e regiões não gênicas para a *E. coli* e *S. cerevisiae* (Seção 3.2);
- a terceira parte apresenta a análise do genoma da *S. cerevisiae*, sendo este analisado regiões gênicas e não gênicas (Seção 3.3).

3.1 Análise comparativa entre os organismos selecionados

Os três organismos selecionados no contexto exobiológico são representantes dos reinos Eubacteria (*E. coli*), Archaea (*T. acidophilum*), ambos organismos procariotos. O terceiro organismo é do Reino Fungi (*S. cerevisiae*) que é representante dos organismos eucariotos (conforme descrito no Apêndice A).

O objetivo desta Seção é caracterizar as regiões não gênicas dos três organismos utilizando-se as técnicas coeficiente de dispersão, DFA e GSA. Como critério de análise, as regiões não codificantes são escolhidas pois não participam diretamente do processo de tradução de RNAs. Arbitrariamente 5 regiões intergênicas (entre os genes) são selecionadas de cada organismo (que não são apontadas pelo GenBank como gênicas) com $1024pb$[1] cada, constituindo portanto uma série com 1024 pontos.

3.1.1 Cálculo do coeficiente de dispersão

Para o cálculo do coeficiente de dispersão, é necessário concatenar as regiões não codificantes de cada um dos organismos (visto que nesse caso é preciso no mínimo 1500 pontos). O número de arestas da rede de tripletes de DNA utilizado para esse trabalho é $L = 250$. Neste caso, um valor de $\bar{C}$ para cada conjunto de $1500pb$ é obtido (GERHARDT et al., 2006).

[1] A designação *pb* - par de base, diz respeito ao nucleotídeo da seqüência. Cada nucleotídeo possui um correspondente na outra fita de DNA.

A Tabela 3.1 apresenta os valores obtidos para GC, média de $\bar{C}_{250}$ ($\langle \bar{C}_{250} \rangle$) e D_{250} para as regiões não codificantes da *E. coli*, *T. acidophilum* e *S. cerevisiae*. A *T. acidophilum* apresenta o menor valor de GC nestas regiões do que os outros dois organismos. Os valores do coeficiente de dispersão mostram que a *S. cerevisiae* tem menor valor ($D \approx 0.3$). Dessa maneira, pode-se afirmar que este organismo está mais próximo do grupo de controle (considerando a diferença igual a 0), do que os dois organismos procariotos (com valores de $D \approx 1$). De uma forma geral, pode-se inferir que as regiões não codificantes da *S. cerevisiae* analisadas possuem uma caracterização mais similar ao grupo controle que as regiões não codificantes dos outros dois organismos.

Tabela 3.1 - Valores de GC, $\langle \bar{C}_{250} \rangle$ e D_{250} para as seqüências não codificantes dos organismos selecionados.

Organismo	GC	$\langle \bar{C}_{250} \rangle$	D_{250}
E. coli	0.4291	0.3990 ± 0.0740	0.8570
T. acidophilum	0.3880	0.4600 ± 0.0555	0.9573
S. cerevisiae	0.4071	0.3936 ± 0.0585	0.2912

3.1.2 Aplicação da DFA

Para a aplicação da técnica DFA é necessário utilizar a seqüência não gênica como um *DNA walk* (descrito Subseção 2.2.2).

A Figura 3.1 mostra os valores de α para cada seqüência não codificante de cada organismo. Note que, quando os valores de *alpha*'s são usados com a finalidade de caracterizar as regiões não codificantes dos três organismos, não é possível estabelecer diferenças entre os α's das seqüências dos organismos analisados. Para identificar a variedade entre os α's, foram calculados o α médio ($\langle \alpha \rangle$), o desvio padrão (σ_α) e o coeficiente de variação C_v[2]. Os resultados são apresentados na Tabela 3.2.

Como a diferença entre os α's médios e seus desvios para os três casos é muito pequena, pode-se dizer que esta técnica não caracterizou as estruturas das regiões não gênicas de cada organismo, o que permitiria distinguí-los como organismos diferentes.

[2]O coeficiente de variação (C_v) é uma medida que permite verificar a variabilidade das seqüências analisadas em relação a média apresentada pelos dados (FERREIRA, 1996). Neste trabalho C_v é definido como $C_v = \sigma_\alpha / \langle \alpha \rangle$. O valor de C_v considerado, neste trabalho como ótimo, é $C_v < 0.05$. Este valor indica que a percentagem de variabilidade das seqüências em relação a média é menor que 5%.

Possivelmente, este resultado está relacionado ao tamanho de cada seqüência não codificante.

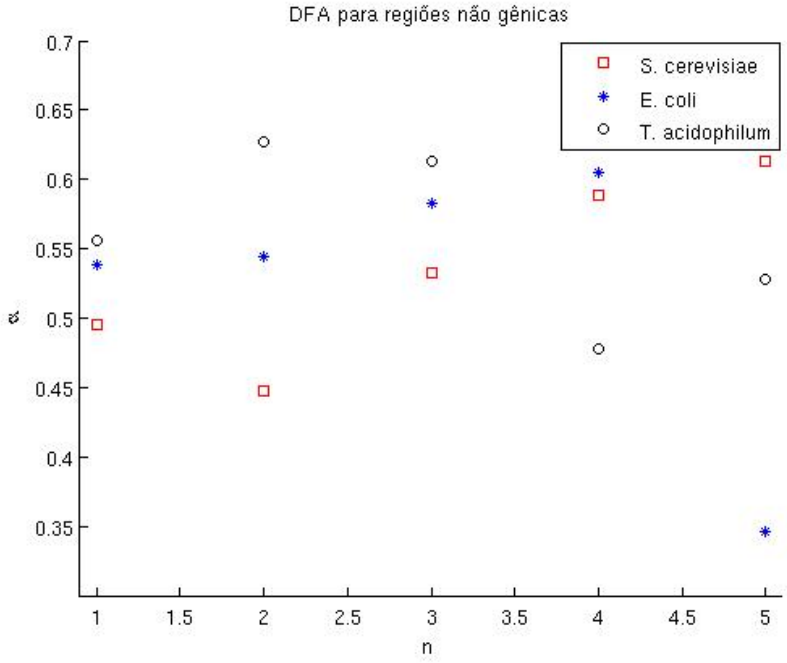


Figura 3.1 - Valores de α obtidos das seqüências não gênicas da *S. cerevisiae*, *E. coli* e *T. acidophilum*. Sendo n a quantidade de seqüências de cada organismo.

Tabela 3.2 - Valores de α, α médio e coeficiente de variação (C_v) discriminados para os três organismos.

	S. cerevisiae		*E. coli*		*T. acidophilum*
α	0.4953	α	0.5390	α	0.5563
	0.4480		0.5448		0.6272
	0.5327		0.5826		0.6136
	0.5885		0.6055		0.4784
	0.6130		0.3457		0.5278
μ	0.5355 ± 0.0673		0.5235 ± 0.1032		0.5508 ± 0.0712
C_v	0.1257		0.1969		0.1293

3.1.3 Aplicação da GSA

Para a aplicação da GSA, cada organismo é caracterizado por seu respectivo espectro gradiente que apresenta um valor típico para a flutuação do coeficiente de assimetria

gradiente (DANTAS, 2008). A Figura 3.2 apresenta o espectro gradiente médio gerado para cada uma das 5 seqüências não codificantes dos três organismos. Na parte superior estão as seqüências da *T. acidophilum* seguidas pela seqüências da *E. coli*, e na parte inferior são apresentadas as seqüências da *S. cerevisiae*. Esta separação pode estar relacionada a propriedades estruturais das seqüências não gênicas de cada organismo: dois procariotos e um eucarioto.

Esta característica também pode ser verificada analisando a Figura 3.3, onde é calculado a flutuação média do espectro gradiente $\langle f_{eg} \rangle$. O grupo dos procariotos mantém-se na parte superior da Figura e o eucarioto na parte inferior, permitindo distinguí-los entre si. Portanto, os resultados obtidos validam a técnica GSA como um novo classificador de seqüências genéticas curtas.

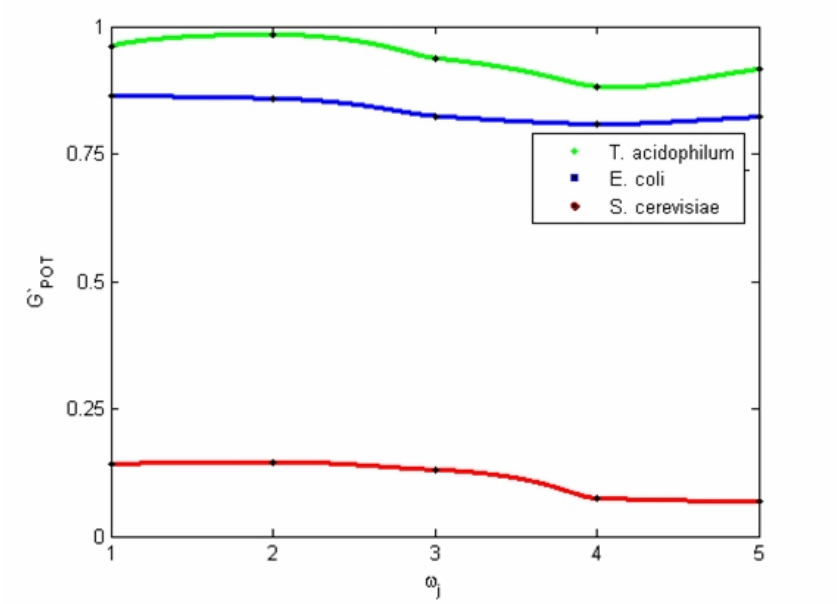


Figura 3.2 - Espectro Gradiente médio normalizado (G'_{POT}) obtido a partir das séries não codificantes dos três organismos.

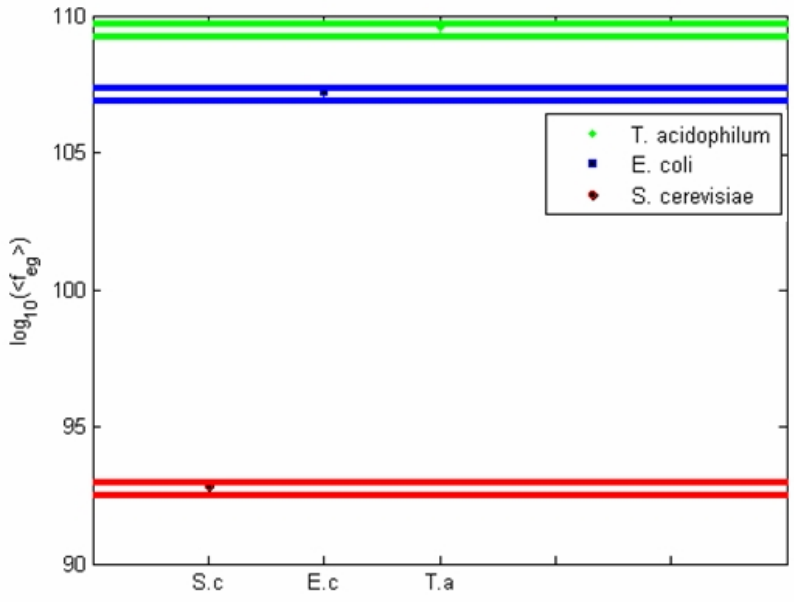


Figura 3.3 - Variação de $\langle f_{eg} \rangle$ para cada conjunto de séries não gênicas dos três organismos.

3.1.4 Discussões dos resultados

Nesta parte do trabalho, a caracterização das seqüências não gênicas[3] dos três organismos que representam distintas origens filogenéticas (característica importante quando considerado o contexto exobiológico), foi realizada pelas três técnicas descritas no Capítulo 2. Em relação aos resultados obtidos pelo emprego das técnicas neste conjunto de dados pode-se dizer que:

- a técnica DFA não apresentou resultado satisfatório, pois não fez distinção entre estes dados. Fato que, pode estar relacionado à restrição de tamanho das seqüências utilizadas (PENG et al., 1994). Os valores obtidos para cada caso apresentam grande discordância detectada pelo coeficiente de variação (C_v);
- o coeficiente de dispersão (D_{250}) caracterizou as seqüências considerando a localização destas em relação a um grupo controle. Com os resultados apresentados verificou-se que os organismos procariotos estão mais distantes do grupo controle do que as séries não gênicas do eucarioto. Os valores de $\bar{C}_{250}$ não distinguiu as seqüências;
- considerando que a *E. coli*, *T. acidophilum* e *S. cerevisiae* são filogeneticamente distantes, a técnica GSA caracterizou as seqüências detectando diferenças finas de assimetrias relacionadas a cada estrutura de DNA.

Analisando a Figura 3.4 verifica-se que a flutuação do coeficiente de assimetria gradiente é menor no caso do eucarioto do que no caso dos dois procariotos. Baseado nos resultados da técnica GSA para os três organismos pode-se conjecturar se é possível relacionar a assimetria encontrada para as estruturas das regiões não gênicas com a evolução dos organismos. Para responder a esta questão é necessário um maior número de análises considerando um grupo contendo mais organismos e mais seqüências de diferentes pontos da árvore filogenética.

[3]Para determinar as regiões não gênicas foram utilizadas informações do banco de dados GenBank, conforme descrito no Apêndice C.

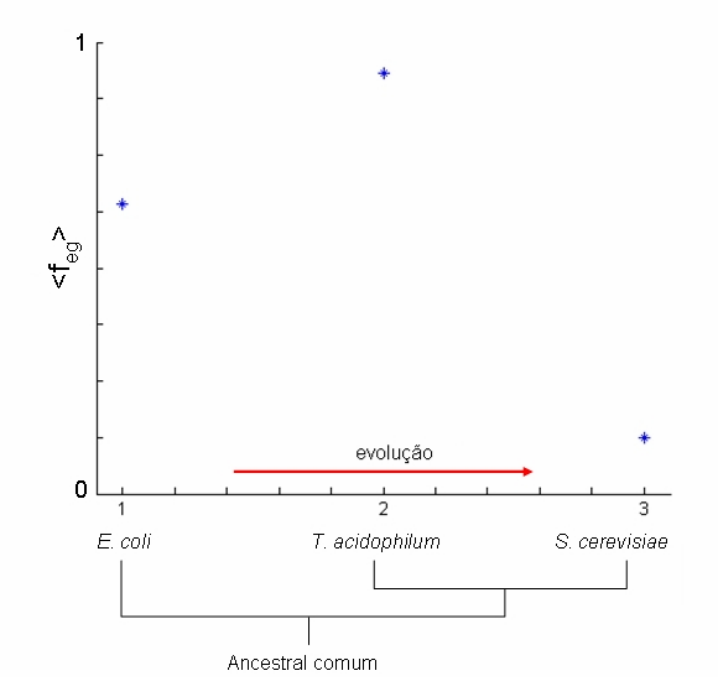


Figura 3.4 - Variação de $\langle f_{eg} \rangle$ para cada conjunto de séries não gênicas relacionado com a evolução dos organismos.

3.2 Genes da Glicólise

Para esta segunda parte do trabalho são considerados apenas dois organismos, a *E. coli* e a *S. cerevisiae*. Estes dois organismos são considerados protótipos, devido ao grande conjunto de informações disponibilizadas sobre eles (ver Apêndice C). Portanto, compreendendo a estruturação dos segmentos gênicos e não gênicos pode-se deduzir as mesmas características aos demais seres procariotos e eucariotos, respectivamente (GARRETT; GRISHAM, 1999).

Nesta análise comparativa são selecionados os genes da Glicólise. A Glicólise é uma etapa metabólica similar em ambos organismos (conforme descrito na Subseção 2.1.1). Aqui são utilizados os genes da Glicólise, alguns isogenes e regiões intergênicas de tamanhos similares aos genes analisados. Estas informações são oriundas dos bancos de dados GenBank, EcoGene, EcoCyc e SGD.

As próximas Subseções apresentam os resultados obtidos com o emprego das três técnicas de caracterização para *E. coli* e *S. cerevisiae*. Primeiramente são apresentados os resultados da *E. coli* seguido pelos resultados da *S. cerevisiae*.

3.2.1 Cálculo do coeficiente de dispersão

Para esta análise considera-se todos os genes da Glicólise juntamente com seus isogenes e todas as regiões não gênicas (intergênicas). As regiões intergênicas têm tamanhos similares aos genes da Glicólise do cromossomo da *E. coli*.

A Figura 3.5 apresenta os valores de $\bar{C}_{250}$ versus $\%GC$ obtidos de cada conjunto de $1500pb$. É possível verificar que os clusteres dos genes permanecem mais aglutinados em relação à quantia de GC do que os clusteres dos segmentos não gênicos. Os clusteres não gênicos possuem uma quantia variável de GC, permanecendo numa faixa constante de $\bar{C}_{250}$. Dessa forma, pode-se dizer que individualmente não há caracterização dos grupos gênico e não gênico.

A Tabela 3.3 mostra os valores de GC, $\langle \bar{C}_{250} \rangle$ e D_{250} para os genes da Glicólise e as regiões não gênicas da *E.coli*. O coeficiente de dispersão D_{250} é capaz de distinguir esses dois grupos (gênico e não gênico) da *E. coli*, sendo que o grupo não gênico está mais próximo do grupo controle que o grupo gênico. Consequentemente, verifica-se uma organização diferenciada entre estes grupos da *E. coli*.

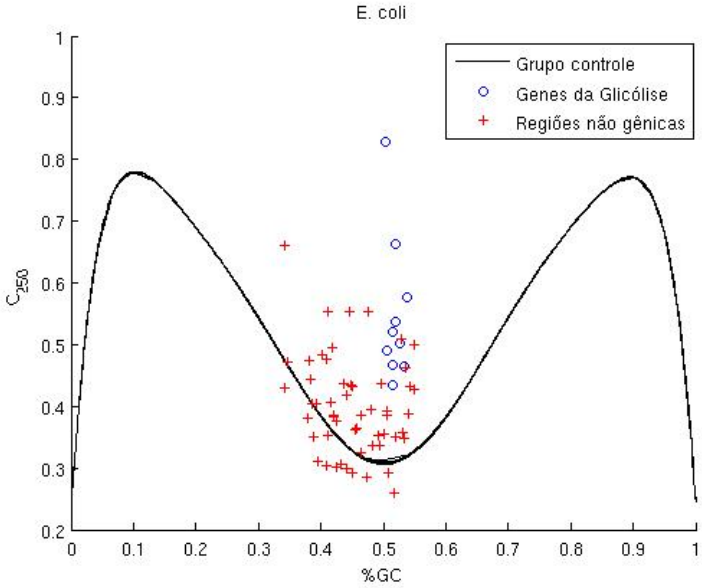


Figura 3.5 - $\bar{C}_{250}$ para todos os genes da Glicólise e todas as regiões não gênicas de tamanhos similares aos dos genes da *E. coli*.

Tabela 3.3 - Valores obtidos usando $L = 250$ para os genes da Glicólise e regiões não gênicas da *E. coli*.

E. coli	GC	$< \bar{C}_{250} >$	D_{250}
Genes da Glicólise	0.5194	0.5483 ± 0.1179	5.1028
Regióes não gênicas	0.4529	0.4002 ± 0.0804	0.9734

A Figura 3.6 apresenta os valores de $\bar{C}_{250}$ para os genes e não genes da *S. cerevisiae*. Nota-se uma separação entre os clusteres das regiões gênicas da Glicólise e das regiões não gênicas. O grupo não gênico está localizado mais próximo do grupo controle do que o grupo gênico. É importante salientar que, na *S. cerevisiae* há mais clusteres não gênicos do que na *E. coli*. Isso é devido ao fato do genoma da *S. cerevisiae* ser maior, constituído assim por mais regiões não codificantes do que a *E. coli*.

A Tabela 3.4 mostra os valores de GC, $\langle \bar{C}_{250} \rangle$ e D_{250} para os genes da Glicólise e as regiões não gênicas da *S. cerevisiae*. Nota-se que D_{250} é capaz de diferenciar os genes dos não genes, sendo o grupo não gênico aquele mais próximo ao grupo controle.

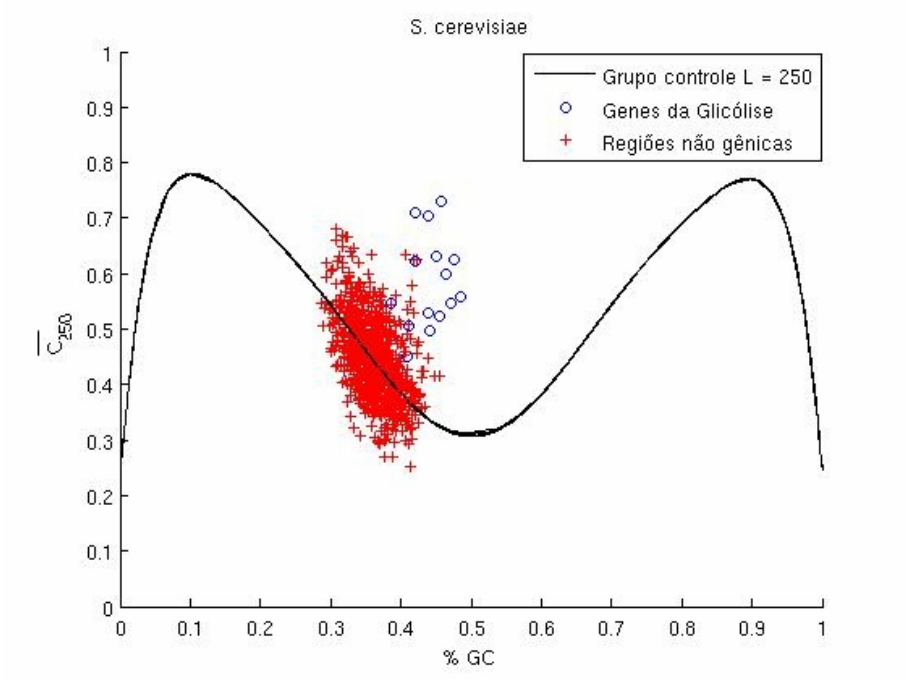


Figura 3.6 - $\bar{C}_{250}$ para todos os genes da Glicólise e todas as regiões não gênicas de tamanhos similares aos dos genes da *S. cerevisiae*.

Tabela 3.4 - Valores obtidos usando $L = 250$ para os genes da Glicólise e regiões não gênicas da *S. cerevisiae*.

S. cerevisiae	GC	$< \bar{C}_{250} >$	D_{250}
Genes da Glicólise	0.4415	0.5855 ± 0.0839	4.9507
Regiões não gênicas	0.3611	0.4433 ± 0.0709	0.0724

3.2.2 Aplicação da DFA

A Figura 3.7a apresenta os 7 genes da Glicólise da *E. coli* transformados em *DNA walk* (são selecionados apenas genes com mais de 1024*pb*) e a Figura 3.7b apresenta as 7 regiões não gênicas escolhidas arbitrariamente com 1024*pb*.

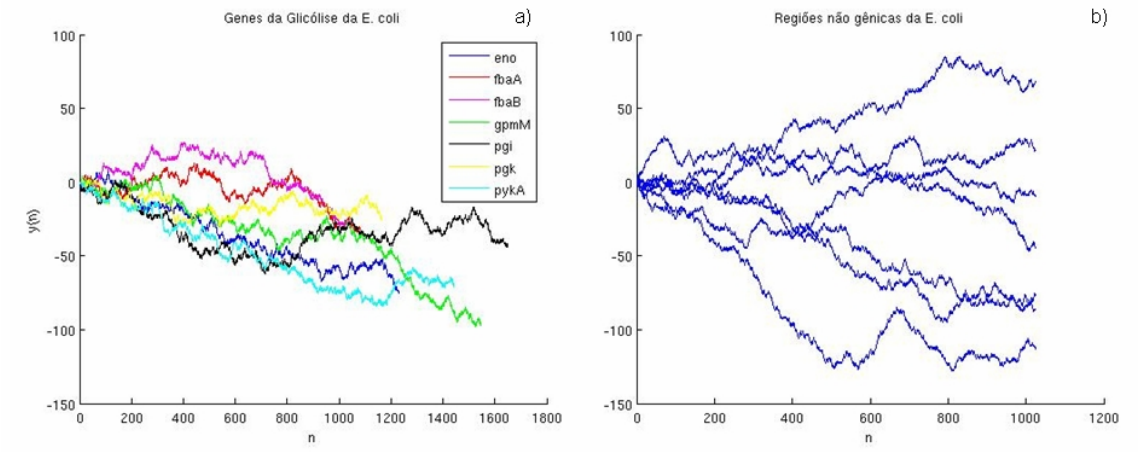


Figura 3.7 - a) *DNA walk* dos segmentos gênicos da Glicólise da *E. coli*. b) *DNA walk* dos segmentos não gênicos da *E. coli*, sendo n o tamanho das seqüências.

De uma forma geral, nota-se que o comportamento dos genes da Glicólise da *E. coli* são mais similares entre si do que o comportamento das regiões não gênicas. Este comportamento pode estar relacionado à maior presença de bases nitrogenadas purinas sendo vizinhas umas das outras, do que nas regiões não gênicas.

A Figura 3.8 mostra os valores de α para cada gene ou não gene da *E. coli*. Neste caso, verifica-se que não há separação entre os dois grupos. A Tabela 3.5 apresenta a discriminação dos valores de α e as médias para cada grupo. O alto valor de C_v (≈ 0.2) indica que não é possível fazer distinção entre os grupos.

A Figura 3.9 apresenta o *DNA walk* dos 7 genes da Glicólise e das 7 regiões não gênicas da *S. cerevisiae*. Nota-se que não é possível estabelecer diferenças entre os dois grupos. A Figura 3.10 mostra os α's obtidos com os genes da Glicólise (maiores que 1024*pb*) e regiões não gênicas com 1024*pb*. Os valores apresentados na Tabela 3.6 mostram as médias para cada grupo. Note que, estes grupos também não são separáveis e que a técnica DFA não separou os dois grupos de organismos: *E. coli* e *S. cerevisiae* ($C_v \approx 0.17$).

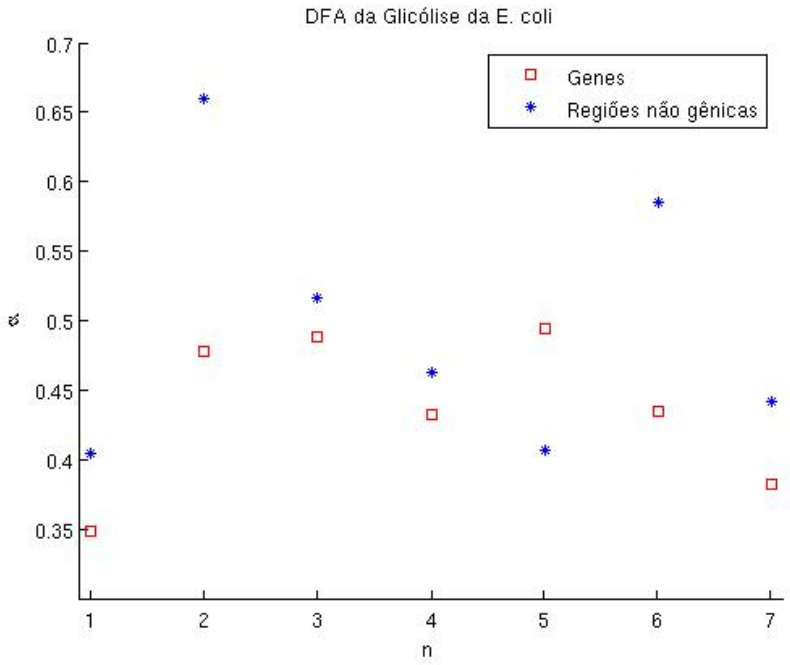


Figura 3.8 - Valores de αs obtidos dos genes da Glicólise e regiões não gênicas. Sendo n o número de seqüências analisadas de cada grupo da *E. coli*.

Tabela 3.5 - Valores de α para 7 genes da Glicólise e regiões não gênicas de tamanhos similares aos genes da *E. coli*. μ é definido como $\mu = \langle\alpha\rangle \pm \sigma_\alpha$

Genes *E. coli*	α do DFA	Regiões não gênicas	α do DFA
eno	0.3485	1	0.4040
fbaA	0.4783	2	0.6603
fbaB	0.4885	3	0.5163
gpmM	0.4324	4	0.4630
pgi	0.4940	5	0.4066
pgk	0.4346	6	0.5849
pykA	0.3822	7	0.4414
μ	0.4369 ± 0.0555		0.4966 ± 0.0965
C_v	0.1270		0.1943

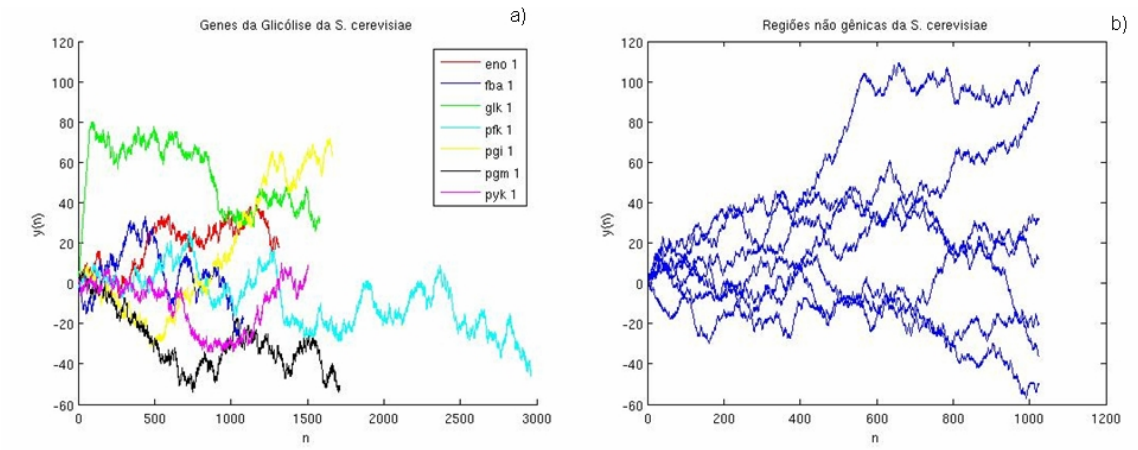


Figura 3.9 - a) *DNA walk* dos segmentos gênicos da Glicólise da *S. cerevisiae*. b) *DNA walk* dos segmentos não gênicos. Sendo n o tamanho de cada seqüência analisada.

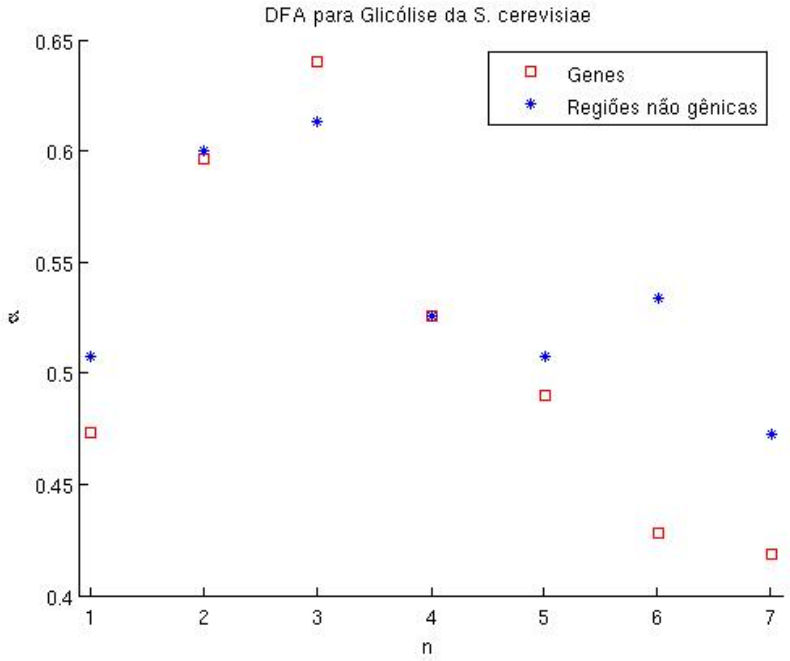


Figura 3.10 - α's de cada gene da Glicólise e região não gênica da *S. cerevisiae*. Sendo n o número de seqüências de cada grupo analisado.

Tabela 3.6 - Valores de α obtidos com a técnica DFA para 7 genes da Glicólise e regiões não gênicas de tamanhos similares aos genes da *S. cerevisiae*.

Genes *S. cerevisiae*	α do DFA	Regiões não gênicas	α do DFA
eno 1	0.4729	1	0.5077
fbaA 1	0.5962	2	0.5999
glk 1	0.6404	3	0.6130
pfk 1	0.5255	4	0.5256
pgi 1	0.4899	5	0.5077
pgm 1	0.4277	6	0.5341
pyk 1	0.4184	7	0.4726
μ	0.5101 ± 0.0833		0.5367 ± 0.0504
C_v	0.1633		0.0939

3.2.3 Aplicação da GSA

Os resultados desta Subseção são apresentados de forma comparativa, sendo avaliada a capacidade desta técnica na distinção dos organismos e dos grupos gênicos e não gênicos do mesmo genoma. A Figura 3.11 mostra o espectro-gradiente médio para os segmentos gênicos e não gênicos da *E. coli* e *S. cerevisiae*. Note que, há uma distinção entre os genes da Glicólise e os não genes, tanto para a *E. coli* como para a *S. cerevisiae*. Portanto, pode-se dizer que há um grau de assimetria diferenciado entre o grupo gênico e não gênico de cada organismo.

Verifica-se também a separação dos organismos em relação aos seus grupos representantes analisados. Na parte superior da Figura 3.11 tem-se os grupos gênico e não gênico da *E. coli*, e na parte inferior tem-se os grupos gênico e não gênico da *S. cerevisiae*, separados através do G_{POT} normalizado.

A caracterização obtida através da flutuação do coeficiente de assimetria $\langle f_{eg} \rangle$ pode ser visualizada na Figura 3.12. Novamente na parte superior têm-se os grupos gênico e não gênico da *E. coli* (distintos entre si) e na parte inferior os grupos da *S. cerevisiae* (também distintos entre si).

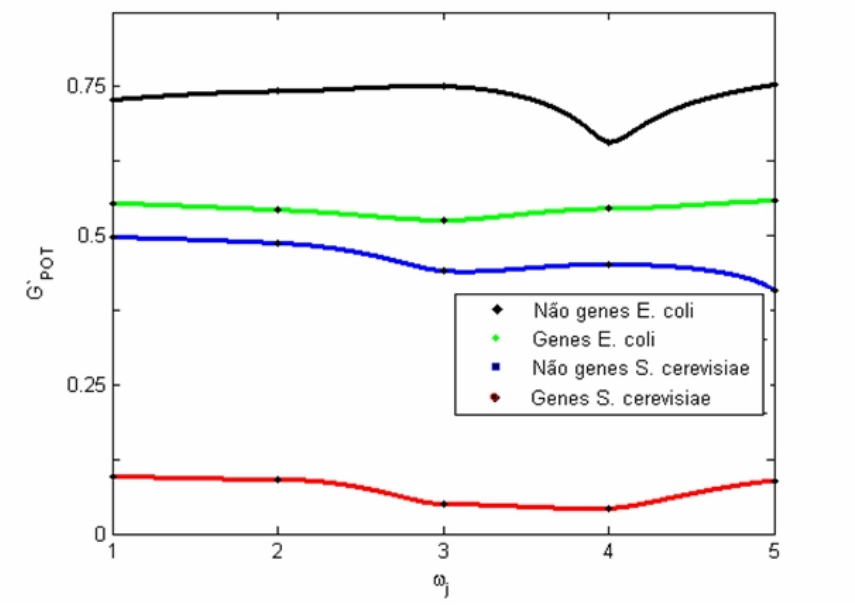


Figura 3.11 - Espectro Gradiente médio obtido a partir do grupo gênico e não gênico da *E. coli* e da *S. cerevisiae*.

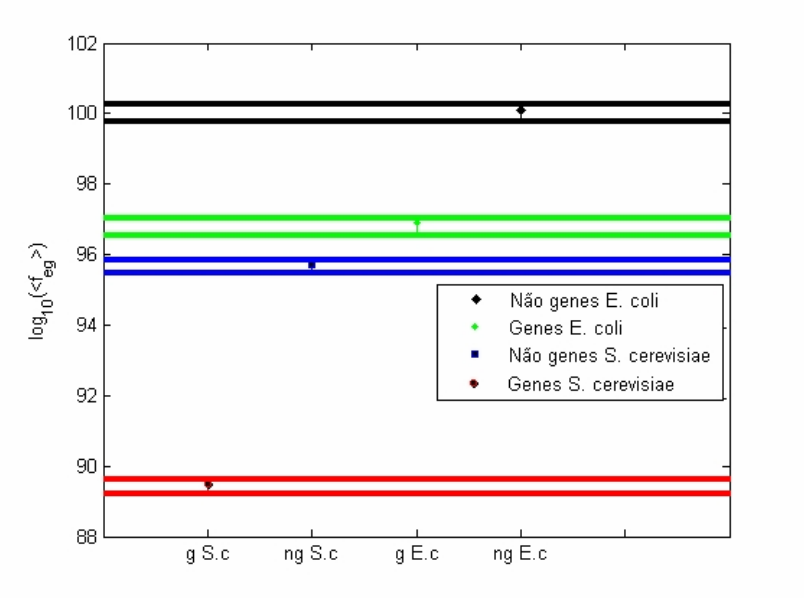


Figura 3.12 - Variação $\langle f_{eg} \rangle$ para cada conjunto de genes da Glicólise e séries não gênicas da *E. coli* e da *S. cerevisiae*.

3.2.4 Discussões dos resultados

Para os resultados das técnicas empregadas neste conjunto de dados (genes da Glicólise e regiões não gênicas com tamanhos similares aos genes da Glicólise) pode-se dizer que:

- a técnica DFA não foi capaz de caracterizar os genes da Glicólise e segmentos não gênicos, ambos com $1024pb$, apresentando um coeficiente de variação $C_v \approx 0.2$. Provavelmente este fato está relacionado a restrição da técnica ao tamanho da seqüência analisada;
- o coeficiente de dispersão caracterizou os grupos dos dois organismos. Em ambos os casos, os grupos não gênicos estão mais próximos do grupo controle (com valores $D \approx 0$) e os grupos gênicos estão mais distantes;
- a técnica GSA caracterizou os grupos de ambos organismos, diferenciando os grupos e também os organismos. De uma forma ampla, pode-se dizer que mesmo os grupos gênicos (participando do mesmo tipo de processo metabólico) são aparentemente diferentes em relação à assimetria em ambos os organismos.

3.3 Regiões gênicas e não gênicas do genoma da *S. cerevisiae*

O genoma da *S. cerevisiae* é segmentado em regiões gênicas e não gênicas, segundo informações do banco de dados GenBank. Este genoma é composto por diversos cromossomos, e por essa razão apresenta vantagens de comparação em relação a *E. coli* e *T. acidophilum*, que são constituídos apenas por um cromossomo.

Nesta Seção são apresentados os resultados obtidos, utilizando a técnica DFA e o coeficiente de dispersão, do genoma nuclear da *S. cerevisiae*.

3.3.1 Cálculo do coeficiente de dispersão

Os valores obtidos para o cálculo do coeficiente de dispersão são apresentados na Tabela 3.7, onde são diferenciados por região (gênica e não gênica) dos cromossomos (ver Figuras complementares no Apêndice H). Os valores para GC, $\langle \bar{C}_{250} \rangle$ e D_{250} são capazes de distinguir as duas regiões dos cromossomos. Destas medidas, a que mais permite distinção é D_{250}. Como o cálculo de D é baseado na aproximação em

relação a um grupo de controle, conforme abordado na Subseção 2.2.1, pode-se dizer que as regiões não gênicas dos cromossomos da *S. cerevisiae* estão localizados mais próximos do grupo controle do que o segmento codificante. Os segmentos codificantes estão localizados acima do grupo controle (apresentam valores positivos).

Tabela 3.7 - Valores de GC, $\langle \bar{C}_{250} \rangle$ e D_{250} para regiões codificantes e não codificantes para cada cromossomo da *S. cerevisiae*.

Cromossomos		GC	$\langle \bar{C}_{250} \rangle$	D_{250}	Cromossomos		GC	$\langle \overline{C}_{250} \rangle$	D_{250}
	1	0.4119	0.4604 ± 0.0852	1.5016		1	0.3586	0.4422 ± 0.0648	-0.0885
	2	0.3919	0.4607 ± 0.0764	1.0522		2	0.3580	0.4478 ± 0.0720	-0.0236
	3	0.4013	0.4682 ± 0.0805	1.3880		3	0.3465	0.4679 ± 0.0750	0.0013
	4	0.3871	0.4705 ± 0.0752	1.1091		4	0.3562	0.4517 ± 0.0687	-0.0010
	5	0.3970	0.4658 ± 0.0800	1.2685		5	0.3589	0.4512 ± 0.0680	0.0635
	6	0.3984	0.4735 ± 0.0823	1.4382		6	0.3631	0.4456 ± 0.0661	0.0792
Regiões	7	0.3891	0.4736 ± 0.0760	1.2157	Regiões	7	0.3591	0.4493 ± 0.0716	0.0267
gênicas	8	0.3943	0.4607 ± 0.0752	1.1173	não	8	0.3595	0.4628 ± 0.0691	0.2517
	9	0.3983	0.4635 ± 0.0784	1.2456	gênicas	9	0.3647	0.4455 ± 0.0618	0.1196
	10	0.3915	0.4670 ± 0.0770	1.1483		10	0.3595	0.4516 ± 0.0678	0.0758
	11	0.3900	0.4666 ± 0.0747	1.1013		11	0.3550	0.4445 ± 0.0631	-0.1404
	12	0.3916	0.4614 ± 0.0788	1.0447		12	0.3658	0.4427 ± 0.0703	0.0688
	13	0.3892	0.4674 ± 0.0735	1.1076		13	0.3601	0.4448 ± 0.0693	-0.0176
	14	0.3951	0.4612 ± 0.0768	1.1341		14	0.3600	0.4455 ± 0.0652	0.0015
	15	0.3907	0.4610 ± 0.0732	1.0325		15	0.3576	0.4510 ± 0.0705	0.0084
	16	0.3896	0.4677 ± 0.0738	1.1307		16	0.3557	0.4553 ± 0.0672	0.0395
Média		0.3942 ± 0.0062	0.4656 ± 0.0045	1.1897 ± 0.1432			0.3586 ± 0.0044	0.4500 ± 0.0071	0.0291 ± 0.0878

As Figuras 3.13, 3.14 e 3.15 apresentam os valores para GC, $\langle \bar{C}_{250} \rangle$ e D_{250} em cada um dos segmentos gênicos e não gênicos do genoma nuclear da *S. cerevisiae*. Analisando as Figuras pode-se notar que as medidas separaram as regiões gênicas e não gênicas.

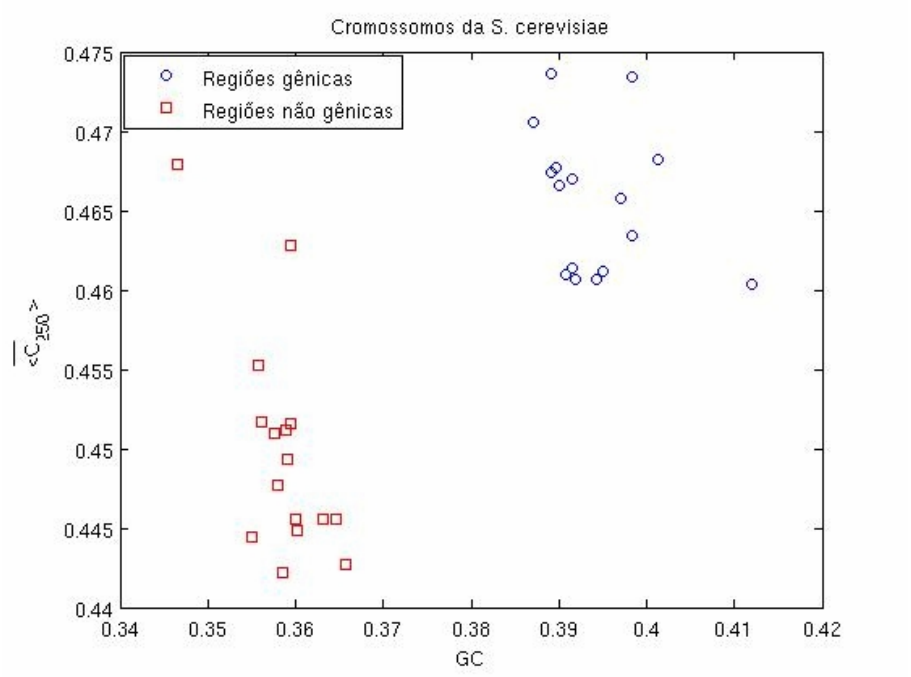


Figura 3.13 - Valores de GC versus $\langle \bar{C}_{250} \rangle$ para cada uma das regiões dos 16 cromossomos.

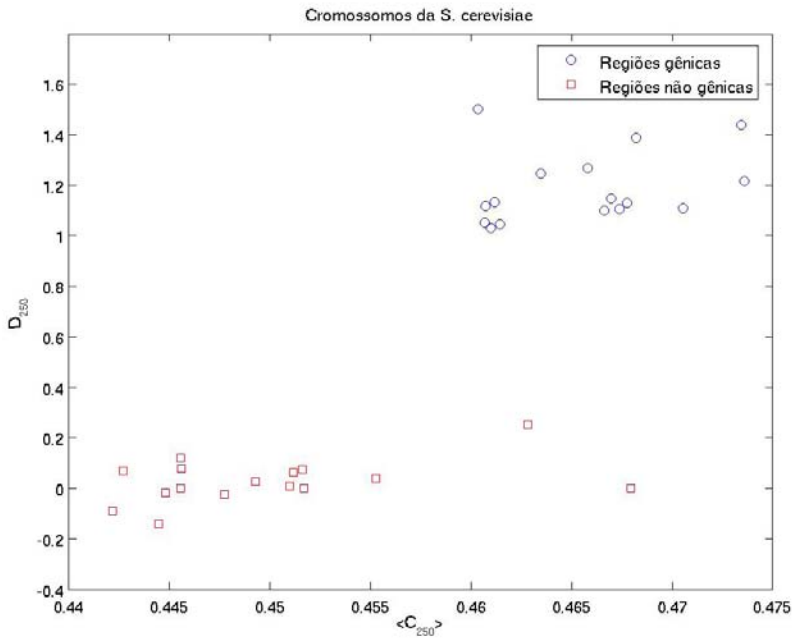


Figura 3.14 - Valores de $\langle \bar{C}_{250} \rangle$ versus D_{250} para cada uma das regiões dos 16 cromossomos.

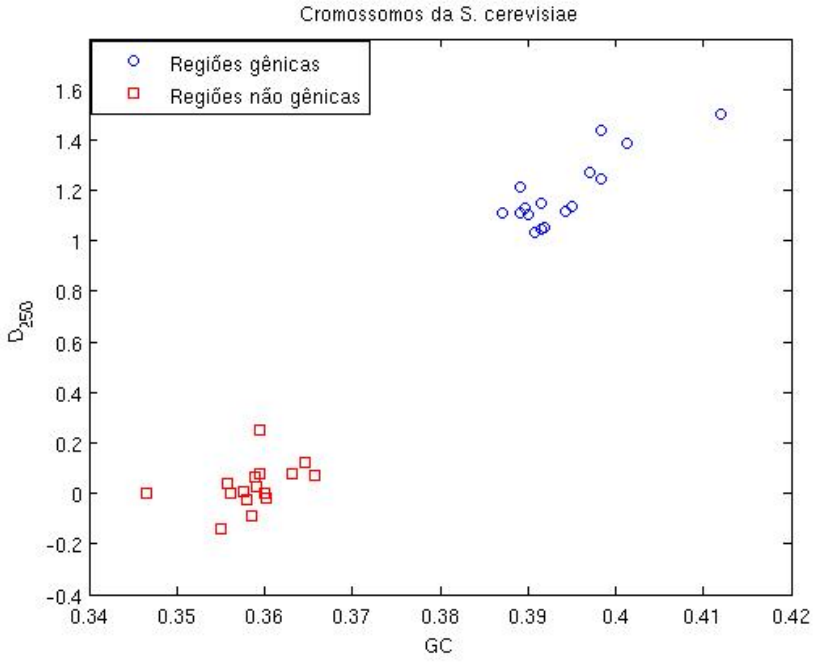


Figura 3.15 - Valores de *GC* versus D_{250} para cada uma das regiões dos 16 cromossomos.

3.3.2 Aplicação da DFA

Para a técnica DFA é utilizado o mesmo critério estabelecido na Subseção 2.2.2 para transfomação das regiões gênicas e não gênicas em um *DNA walk*. As Figuras G.1, G.2, G.3 e G.4 apresentam o *DNA walk* obtido para cada um dos segmentos gênicos e não gênicos dos cromossomos da *S. cerevisiae*. A Figura 3.16a apresenta todos os segmentos gênicos e a Figura 3.16b apresenta todos os segmentos não gênicos do genoma total. Em média este genoma tem 70.5% de regiões codificantes (BENSON et al., 2002).

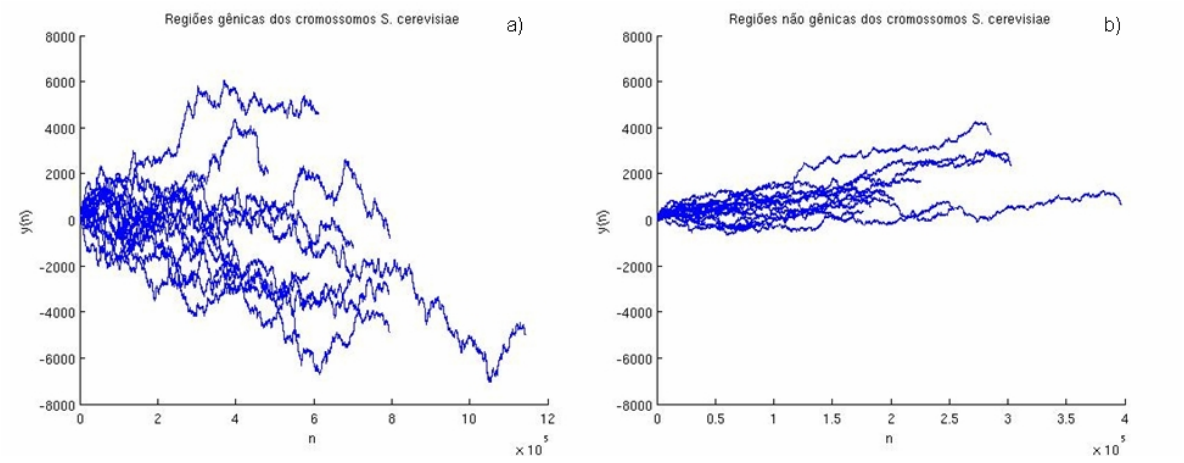


Figura 3.16 - *DNA walk* de todos os segmentos gênicos (a) e não gênicos (b) dos cromossomos da *S. cerevisiae*.

De uma forma geral pode-se verificar que em todos os cromossomos, o segmento gênico é maior que o não gênico (ver Apêndice G). Analisando a variação das bases nitrogenadas R ou Y (se uma R é vizinha ou não de Y) pode-se inferir que no *DNA walk* gênico há mais predomínio de vizinhos do mesmo tipo do que no segmento não gênico. A Figura 3.17 apresenta este resultado para um dos cromossomos da *S. cerevisiae*.

Analisando as duas áreas destacadas da seqüência gênica da Figura 3.17 pode-se verificar que ambas apresentam o predomínio de vizinhança de bases nitrogenadas do mesmo tipo (R ou Y). Esse comportamento é menos freqüente no segmento não gênico. Note que, na área destacada à esquerda na Figura, existem mais vizinhos R (A e G), e na área à direita mais vizinhos Y (T e C). Esta organização não pode ser verificada claramente no segmento não gênico na Figura 3.17.

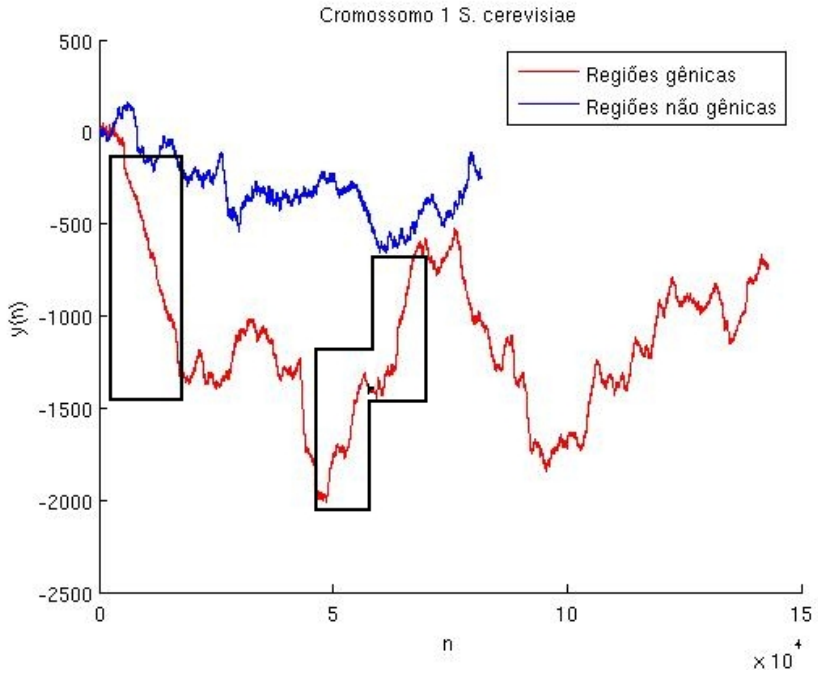


Figura 3.17 - *DNA walk* gerado a partir do pré processamento em regiões gênicas e não gênicas do cromossomo 1 da *S. cerevisiae*. As regiões destacadas pelos retângulos são exemplos de observação de predominância de vizinhos do mesmo tipo (purinas (R) ou pirimidinas(Y).

Os valores de α das regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae* podem ser visualizados na Figura 3.18. A Tabela 3.8 apresenta os valores de α e a média destes valores para cada uma das regiões gênicas e não gênicas. Analisando a Tabela 3.8, nota-se que o valor de C_v para o grupo gênico e para o grupo não gênico é menor que 0.05. A técnica DFA é capaz de caracterizar os segmentos gênicos e não gênicos de um mesmo cromossomo evidenciando uma organização diferenciada entre estes segmentos.

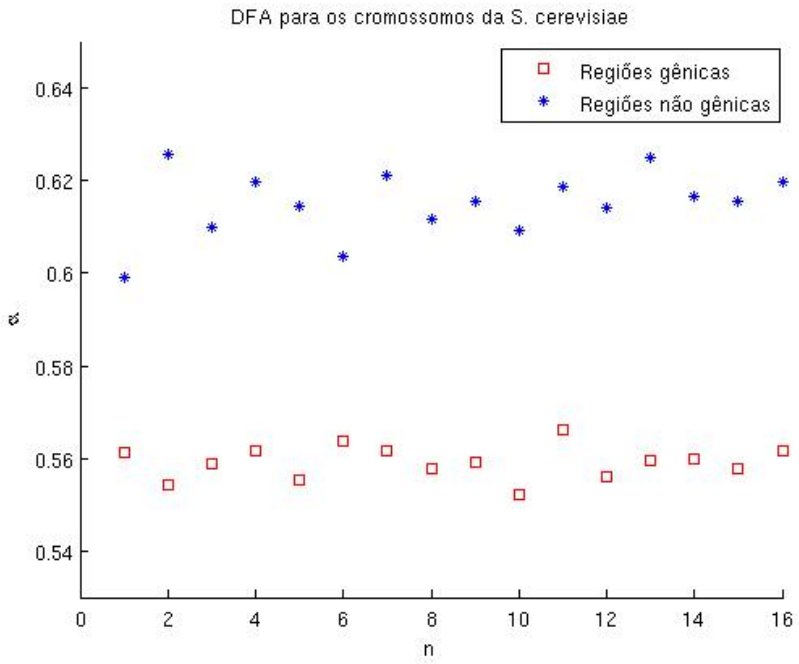


Figura 3.18 - Valores de α para cada região gênica e não gênica dos cromossomos da *S. cerevisiae*. O grupo superior ($*$) é o conjunto de segmentos não codificantes e o grupo inferior ($\square$) é o conjunto de segmentos codificantes.

Tabela 3.8 - Valores de α para as regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae*.

Cromossomo	α da região gênica	α da região não gênica
1	0.5613	0.5990
2	0.5542	0.6256
3	0.5587	0.6101
4	0.5618	0.6197
5	0.5554	0.6144
6	0.5639	0.6035
7	0.5616	0.6213
8	0.5578	0.6118
9	0.5591	0.6155
10	0.5523	0.6094
11	0.5662	0.6186
12	0.5561	0.6142
13	0.5596	0.6249
14	0.5598	0.6167
15	0.5579	0.6156
16	0.5615	0.6198
Média (μ)	0.5592 ± 0.0036	0.6150 ± 0.0071
C_v	0.0064	0.0116

3.3.3 Discussões dos resultados

Para analisar este conjunto de dados (genoma segmentado da *S. cerevisiae* em regiões gênicas e não gênicas) foram aplicadas apenas as técnicas DFA e coeficiente de dispersão. A técnica GSA foi desenvolvida para análise de séries temporais curtas[4], e por esse motivo não foi utilizada nesta parte do trabalho (sendo que o menor segmento possui $81000pb$) . Analisando os resultados com essas duas técnicas, pode-se dizer que: a técnica DFA caracterizou as seqüências em gênicas e não gênicas. O baixo valor de C_v ($C_v \approx 0.012$) evidencia esse resultado.

O coeficiente de dispersão para $L = 250$ arestas, mostrou-se uma técnica adequada para caracterizar este conjunto de dados, mostrando diferenças entre as regiões analisadas. O valor de GC dessas regiões também mostrou eficácia para caracterização do caso estudado, separando as duas regiões.

[4]Segundo (DANTAS, 2008), uma série é considerada série curta quando é constituída por $N \approx 10^3$ pontos.

4 Conclusões

A caracterização das diferentes estruturas presentes nas seqüências genéticas contribuem para estudos de origem dos organismos, sendo esta uma das áreas de interesse da Exobiologia. A relação entre os padrões estruturais e sua possível evolução nos organimos pode contribuir para a compreensão da complexidade dos sistemas biológicos. Neste trabalho foram estudadas a eficiência e robustez das técnicas análise da flutuação "destendenciada" (DFA), coeficiente de dispersão e análise espectral gradiente (GSA) para caracterizar seqüências genéticas de organismos filogeneticamente distantes.

Com relação aos resultados quando a técnica DFA é aplicada em seqüências genéticas transformadas em *DNA walk* observa-se que esta técnica caracterizou robustamente quando o *DNA walk* possui mais de três décadas de nucleotídeos (para o caso do genoma da levedura). Esta técnica caracterizou as regiões gênicas (com valores de α próximos a 0.5) e as regiões não gênicas (com valores de α próximos a 0.6) de cada cromossomo da *S. cerevisiae*.

Os resultados com relação ao cálculo do coeficiente de dispersão usando $L = 250$ tripletes conseguem caracterizar os diferentes organismos. Estes resultados, quando comparados aos resultados obtidos por Gerhardt et al. (2006), são compatíveis quanto a separabilidade dos organismos relacionados ao seu conteúdo GC e aos valores de D para as regiões gênicas e não gênicas analisadas. Os resultados significativos no emprego desta técnica são relacionados a organização das seqüências genéticas considerando-se um grupo controle. Neste contexto, os segmentos não gênicos das seqüências estão mais próximos do grupo controle, e os segmentos gênicos estão mais distantes.

No caso da GSA, os resultados têm especial interesse devido à sua sensibilidade em caracterizar pequenas mudanças de simetria (alternância de nucleotídeos ao longo da seqüência de DNA) em um padrão estrutural. Em particular, a GSA caracterizou as estruturas assimétricas dos diferentes segmentos gênicos e não gênicos das seqüências genéticas com $1024pb$ de cada organismo selecionado. Esta técnica separou os organismos procariotos e eucarioto e distinguiu os segmentos diferentes de um mesmo organismo. Estes resultados podem ser considerados inéditos, já que essa técnica só tem sido utilizada para classificação de sinais ambientais, cardiovasculares e sistemas dinâmicos (DANTAS, 2008). Dessa forma a GSA pode se tornar uma

ferramenta complementar na caracterização de estruturas genéticas, em particular, estruturas importantes para a área de Exobiologia.

Como perspectiva de trabalhos futuros pode-se citar a ampliação do conjunto de dados analisados e a aplicação destas técnicas em outros organismos para verificar sua filogenia. Em relação ao desempenho das técnicas, pode-se uniformizar a linguagem de programação utilizada integrando-as em um mesmo ambiente. E por fim, a publicação destes resultados para comunidade de Bioinformática e Exobiologia.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERT, R.; BARABáSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, n. 1, p. 47–97, 2002. 39, 107, 108

ASSIREU, A.; ROSA, R.; VIJAYKUMAR, N.; LORENZZETTI, J.; REMPEL, E.; RAMOS, F.; SA, L.; BOLZAN, M.; ZANANDREA, A. Gradient pattern analysis of short nonstationary time series: an application to lagrangian data from satellite tracked drifters. **Physica D**, v. 168, p. 397–403, 2002. 45, 111, 112, 113, 114

BAI, F.-I.; LIU, Y. zhao; WANG, T. ming. A representation of dna primary sequences by random walk. **Mathematical Biosciences**, v. 209, p. 282–291, 2007. 29, 104

BARONI, M. P. M. A.; ROSA, R. R.; SILVA, A. F. da; PEPE, I.; ROMAN, L.; RAMOS, F.; AHUJA, R.; PERSSON, C.; VEJE, E. Modeling and gradient pattern analysis of irregular sfm structures of porous silicon. **Microelectronics Journal**, v. 37, p. 290–294, 2006. 111, 113, 114

BARONI, M. P. M. A.; ROSA, R. R.; WIT, A. D. Structural differences between miscible density and viscous fingering dynamics measured by gradient pattern analysis. **submitted to PRE**, v. 37, p. 290–294, 2009. 20, 48, 112

BENSON, D.; KARASH-MIRZASCHI, L.; LIPMAN, D. Genbank. **Nucleic Acids Research**, v. 30, n. 1, p. 17–20, 2002. 33, 74, 97

BULDYREV, S. V.; DOKHOLYAN, N. V.; GOLDBERGER, A. L.; HAVLIN, S. Analysis of dna sequences using methods of statistical physics. **Physica A**, v. 249, p. 430–438, 1998. 103, 104

CRICK, F. H. C. Central dogma of molecular biology. **Nature**, v. 227, p. 561–563, Ago 1970. 19, 88, 89

CRISTEA, P. D. Genomic signal processing and statistic. In: _____. New York: EURASIP Book Series on Signal Processing and Communications, 2005. cap. Representation and analysis of DNA sequences. 20, 23, 103, 104

DANTAS, M. da S. **Análise espectral de padrões-gradiente de séries temporais curtas**. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, 2008. 18, 37, 45, 47, 48, 49, 56, 77, 79, 116, 117

DAUBECHIES, I. **Ten lectures on wavelets**. Montpelier: Capital City Press, 1992. 117

FANTONI, R. F.; MANRICH, S. Macromoléculas nos cometas. **Polímeros: Ciência e Tecnologia**, v. 18, n. 1, p. E4–E11, 2008. 96

FERREIRA, D. F. **Estatística básica**. Lavras, 1996. 54

GAO, W.; LI, B. L. Wavelet analysis of coherent structures at atmosphere-forest interface. **Journal of Applied Meteoroly**, v. 32, p. 1717–1725, 1993. 116

GARRETT, R. H.; GRISHAM, C. M. **Biochemistry**. New York: Brooks Cole, 1999. 1200 p. 35, 36, 60

GERHARDT, G. J. L.; LEMKE, N.; CORSO, G. Network clustering coefficient approach to dna sequence analysis. **Chaos Solitons & Fractals**, v. 28, p. 1037–1045, 2006. 30, 37, 38, 39, 53, 79, 108, 110

GILMOUR, I.; SEPHTON, M. A.; CONWAY, A. **An introduction to astrobiology**. Cambridge: Cambridge University Press, 2004. 358 p. ISBN 0521837367. 95, 96

GUERRA, J. de M. **Caracterização fina dos padrões de variabilidade de EGG's para validação de modelos e aplicações em microgravidade**. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, 2008. 37, 45

HAGELBERG, C. R.; GAMAGE, N. K. K. Applications of structure preserving wavelet decompositions to intermittent turbulence: a case study. In: KUMAR, E. F.-G. . P. (Ed.). **Wavelets Transforms in Geophysics**. [S.l.]: Academic Press, 1994. IX. 45, 51

HOMECK, G.; RETTBERG, P. **Complete course in astrobiology**. Weinheim: Willey-VCH Verlag GmbH & Co. KGaA, 2007. 413 p. ISBN 9783527406609. 95

HWANG, S.; SON, S.-W.; KIM, S. C.; KIM, Y. J.; JEONG, H.; LEE, D. A protein interaction network associated with asthma. **Journal of Theoretical Biology**, v. 252, p. 722–731, 2008. 109

ISHIGAMI, M.; IHARA, H.; SHINODA, H. Molecular evolution of aminoacyl trna synthetases and origin of universal genetic code. In: COSMOVICI, C. B.; BOWYER, S.; WERTHIMER, D. (Ed.). **Astronomical and Biochemical Origins and the Search for Life in the Universe**. Capri, 1996. p. 483–489. 34

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABáSI, A.-L. The large-scale organization of metabolic networks. **Nature**, v. 407, p. 651–654, 2000. 109

LESK, A. M. **Introdução à bioinformática**. Porto Alegre: Artmed, 2007. 28

LILJEROS, F.; EDLING, C. R.; AMARAL, L. A. N. Sexual networks: implications for the transmission of sexually transmitted infections. **Microbes and Infection**, v. 5, p. 189–196, 2003. 109

MALLAT, S. Multiresolution approximations and wavelets orthonormal bases of l(r). **Transactions of the American Mathematical Society**, v. 315, p. 69–87, 1989. 45

MILLER, S. L. A production of amino acids under possible primitive earth conditions. **Science**, v. 117, p. 528–529, 15 maio 1953. 27, 92

NELSON, D. L.; COX, M. M. **Lehninger principles of biochemistry**. W. H. Freeman: W. H. Freeman, 2004. 1100 p. ISBN 0716743396. 28, 87

NEWMAN, M. E. J. The structure and function of complex network. **SIAM Review**, v. 45, p. 167–256, 2003. 108

PENG, C.-K.; BULDYREV, S. V.; GOLDBERGER, A. L.; HAVLIN, S.; SCIORTINO, F.; SIMONS, M.; STANLEY, H. E. Long-range correlations in nucleotide sequences. **Nature**, v. 356, p. 168–170, 1992. 42

PENG, C.-K.; BULDYREV, S. V.; HAVLIN, S.; SIMONS, M.; STANLEY, H. E.; GOLDBERGER, A. L. Mosaic organization of dna nucleotides. **Physical Review E**, v. 49, n. 2, p. 1685–1689, 1994. 17, 29, 30, 37, 42, 43, 44, 58, 103

PODOBNIK, B.; SHAO, J.; DOKHOLYAN, N. V.; ZLATIC, V.; STANLEY, H. E.; GROSSE, I. Similarity and dissimilarity in correlations of genomic dna. **Physica A**, v. 373, p. 497–502, 2007. 29, 104

PUKKILA, P. J. Molecular biology: the central dogma. **Encyclopedia of Life Science**, 19 Abril 2001. Disponível em: <http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0000812/current/abstract>. 19, 89

RILEY, M.; ABE, T.; ARNAUD, M. B.; BERLYN, M. K.; BLATTNER, F. R.; CHAUDHURI, R. R.; GLASNER, J. D.; HORIUCHI, T.; KESELER, I. M.;

KOSUGE, T.; MORI, H.; PERNA, N. T.; PLUNKETT, G. 3rd; RUDD, K. E.; SERRES, M. H.; THOMAS, G. H.; THOMSON, N. R.; WISHART, D.; WANNER, B. L. Escherichia coli k-12: a cooperatively developed annotation snapshot - 2005. **Nucleic Acids Research**, v. 34, n. 1, p. 1–9, 2006. 98

RODRIGUES, F. A. **Caracterização, classificação e análise de redes complexas**. Tese (Doutorado) — Instituto de Física de São Carlos - USP, São Carlos, 2007. 107, 108

ROSA, R.; CAMPOS, M. R.; VIJAYKUMAR, N. L.; FUJIWARA, S.; SATO, T. Gradient pattern analysis of structural dynamics: application to molecular system relaxation. **Brazilian Journal of Physics**, v. 33, p. 605–610, 2003. 45, 111, 112, 113, 114

ROSA, R. R.; KARLICKY, M.; VERONESE, T. B.; DANTAS, M. da S. Gradient pattern analysis of short solar radio bursts. **J. Adv. Space Res.**, v. 42, n. 5, p. 844–851, 2008. 30, 45, 48

ROSA, R. R.; PONTES, J.; CHRISTOV, C.; RAMOS, F. M.; NETO, C.; REMPEL, E.; WALGRAEF, D. Gradient pattern analysis of swift-hohenberg dynamics: phase disorder characterization. **Physica A**, v. 283, n. 1–2, p. 156–159, 2000. 111, 113, 114

ROSA, R. R.; SHARMA, A. S.; VALDIVIA, J. A. Characterization of asymmetric patterns in spatially extended systems. **Int. J. Mod. Phys. C**, v. 10, n. 1, p. 147–163, 1999. 45, 48, 111, 112, 113, 114, 115

RUDD, K. E. Ecogene: a sequence genome database for escherichia coli k-12. **Nucleic Acids Research**, v. 28, n. 1, p. 60–64, 2000. 33

RUEPP, A.; GRAML, W.; SANTOS-MARTINEZ, M.-L.; KORETKE, K. K.; VOLKER, C.; MEWES, H. W.; FRISHMAN, D.; STOCKER, S.; LUPAS, A. N.; BAUMEISTER, W. The genome sequence of the thermoacidophilic scavenger termoplasma acidophilum. **Nature**, v. 407, p. 508–513, 2000. 33, 99

SANTOS, L. dos. **Relatório final da FAPESP: caracterização computacional de padrões estruturais em seqüências de DNA relacionadas a processos em redes metabólicas**. São José dos Campos, 2008. 37

SCALA, A.; AMARAL, L. A. N.; BARTHéLéMY, M. Small-world networks and the conformation space of a lattice polymer chain. **Europhys. Lett.**, v. 55, p. 594–600, 2001. 110

SCHRöDINGER, E. **O que é vida?** São Paulo: Editora Unesp, 1997. 96

SILVA, S. de Avila e. **Redes neurais artificiais aplicadas na caracterização e predição de regiões promotoras**. Dissertação (Mestrado) — Universidade do Vale do Rio dos Sinos, 2006. 29

STRASSER, B. J. A world in one dimension: Linus pauling, francis crick and the central dogma of molecular biology. **Hist. Phil. Life Sci.**, v. 28, p. 491–512, 2006. 88

TRIFONOV, E. N. Glycine clock: Eubacteria first, archaea next, protoctista, fungi, planta and animalia at last. **Gene Therapy and Molecular Biology**, v. 4, p. 313–322, Dez 1999. 27, 28, 92

WATSON, J. D.; CRICK, F. H. C. Molecular structure of amino acids. **Nature**, v. 171, n. 4356, p. 737–738, 25 abril 1953a. 87

_____. Genetical implications of the structure of deoxyribonucleic acid. **Nature**, v. 171, n. 4361, p. 964–967, 30 maio 1953b. 19, 87, 89

A APÊNDICE A - A Estrutura dos Ácidos Nucléicos

A.1 As moléculas de Ácidos Nucléicos

Os organismos vivos e os vírus possuem em sua constituição moléculas denominadas Ácidos Nucléicos. Essas moléculas podem ser o DNA (*Deoxyribonucleic Acid*) e o RNA (*Ribonucleic Acid*). Com exceção dos vírus, que são formados por um dos tipos, os demais organismos contêm ambos os Ácidos Nucléicos. Desde o começo do século XX já era de conhecimento dos cientistas que os Ácidos Nucléicos eram o "material genético" das células, mas foi só em 25 de abril de 1953 que James D. Watson e Francis Crick publicaram na *Nature* a estrutura do DNA (WATSON; CRICK, 1953a).

Neste artigo, Watson e Crick afirmam que a estrutura do DNA tem duas cadeias helicoidais girando sobre o mesmo eixo, sendo que a seqüência de átomos de cada é formada em direções opostas. As unidades formadoras dessa molécula são fosfato, desoxirribose (açúcar de cinco carbonos) e as bases nitrogenadas (purinas: A (adenina) e G (guanina); pirimidinas: T (timina) e C (citosina)) (Figura A.1). Diferentemente dos modelos propostos anteriormente, neste, as bases nitrogenadas estão do lado interno da hélice e as moléculas de fosfato do lado de fora (WATSON; CRICK, 1953a). Afirmaram que as bases nitrogenadas de uma cadeia estariam conectadas as outras bases nitrogenadas da cadeia complementar da seguinte forma: A com T e G com C (ver Figura A.2).

Em 30 de maio de 1953 os autores da estrutura do DNA publicaram também na *Nature* um novo artigo explicando as implicações genéticas da estrutura do DNA. Neste, afirmavam que a molécula era uma cadeia longa constituída de grupos alternados de açúcar (desoxirribose) e fosfatos, onde cada açúcar estava ligado a uma das bases nitrogenadas (ver Figura A.2). A unidade fundamental, formada por fosfato, açúcar e base foi chamado de nucleotídeo, os nucleotídeos em seqüência no DNA poderiam ser copiados, desde que obedecidos as regras de pariamento das bases (WATSON; CRICK, 1953b).

Esses dois artigos contribuiram muito para o que sabe-se hoje sobre estrutura dos Ácidos Nucléicos. Sobre o pareamento das bases, sabe-se que, as bases G e C são conectadas por três pontes de hidrogênio e as bases A e T duas. Sendo assim, é requerido maior energia para dissociação da ligação GC (NELSON; COX, 2004). Em relação a justaposição dos nucleotídeos (que recebem os mesmos nomes das

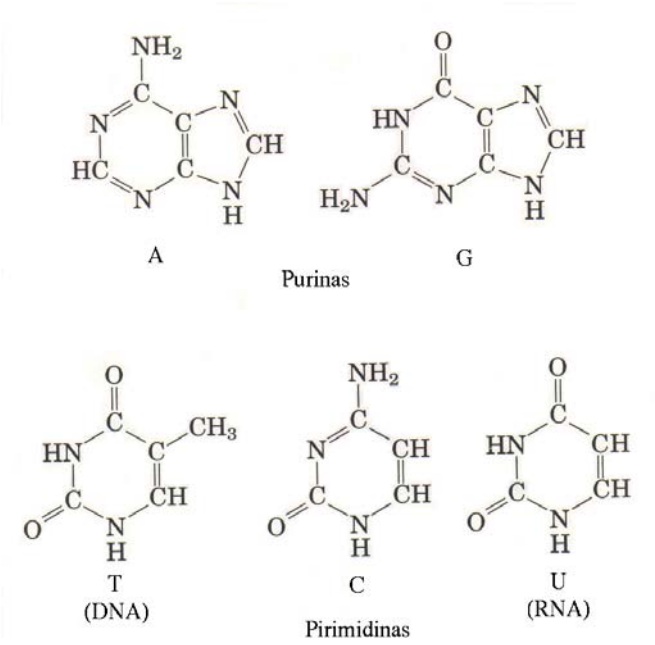


Figura A.1 - Bases nitrogenadas encontradas nos Ácidos Nucléicos. As bases nitrogenadas purinas são *A* e *G* e as bases nitrogenadas pirimidinas são *T* (encontrada apenas em DNA), *U* (encontrada apenas em RNA) e *C*.

bases nitrogenadas) formarem um ácido nucléico, é conhecido que, conforme sua disposição, refletirá em uma propriedade de funcionabilidade de um determinado organismo vivo, ou seja, conforme a seqüência dos nucleotídeos em uma região terá uma proteína distinta.

A molécula de DNA possui mecanismos que permitem sua replicação ou autoduplicação, sua transcrição para RNA e a tradução para proteínas. A replicação ou autoduplicação do DNA ocorre durante a formação de novas células. A transcrição para RNA mensageiro[1] que por sua vez será traduzido para aminoácidos, cuja seqüência será responsável pela proteína formada. Esses mecanismos constituem o Dogma Central da Biologia, (ver Figura A.3) proposto por Crick em 1958 (CRICK, 1970) e (STRASSER, 2006).

[1]Na célula são transcritos três tipos de RNA. RNA ribossômico - que origina o ribossomo, RNA transportador - carrega os aminoácidos e se liga ao RNA mensageiro que traz os códons que serão traduzidos. Os três juntos desempenham o papel de tradução de proteínas.

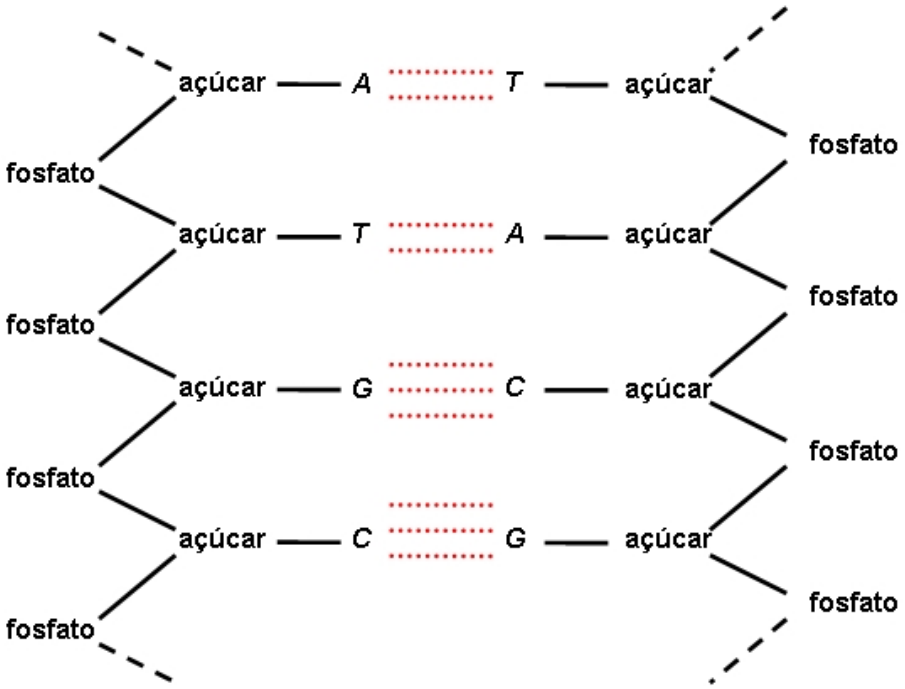


Figura A.2 - Representação da formulação química da molécula de DNA. As pontes de hidrogênio são simbolizadas pelas linhas pontilhadas vermelhas. Verifica-se que para o pareamento das bases *A* e *T* há duas pontes de hidrogênio e para as bases *G* e *C* são necessárias três pontes de hidrogênio. Adaptado de (WATSON; CRICK, 1953b).

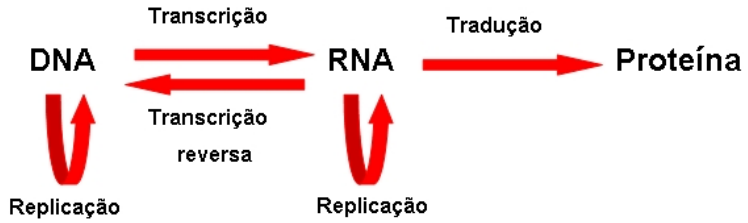


Figura A.3 - Representação do Dogma Central da Biologia. Baseado em (CRICK, 1970) e (PUKKILA, 2001).

A.1.1 Como a informação genética passa do DNA para as proteínas?

Basicamente, a informação genética passa do DNA para as proteínas através do processo de síntese do RNA mensageiro e depois de aminoácidos, conforme citado na Seção A.1. A síntese de RNA é realizada por uma das fitas de DNA. No RNA não há a base nitrogenada *T*, no lugar desta há a base pirimidina chamada uracila (*U*) (ver Figura A.1). O conjunto de três nucleotídeos dá-se o nome de códon ou triplete. É conhecido a existência de 64 códons que traduzem apenas 20 aminoácidos, por-

tanto, diz-se que o código genético é degenerativo, já que, o número de aminoácidos traduzidos é menor que o número possível de códons (ver Tabela A.1 e Tabela A.2).

Esses 20 aminoácidos estão presentes nas proteínas conhecidas dos organismos. É importante salientar que, não foram quaisquer segmentos de DNA responsáveis por esse processo de síntese. Nesta seqüência, há segmentos capazes de serem transcritos e traduzidos para uma determinada proteína e outros não. Os segmentos que participam desse processo chamam-se éxons (regiões gênicas) e os que não participam chamam-se íntrons (regiões não gênicas).

A quantidade de éxons em um genoma depende da quantidade de moléculas de DNA presentes nesse genoma, do seu tamanho e de qual tipo de organismo. Por exemplo, quando considerado um procarioto[2] como a bactéria *Eschechiria coli* seu genoma é constituído por um único cromossomo circular composto por 4.639.675 pares de bases nitrogenadas (*pb*) e destas 85% são codificantes, ou seja, 3.943.723*pb*. Quando considerado um eucarioto[3] como a levedura *Saccharomyces cerevisiae* que tambéma é unicelular como a bactéria, mas seu genoma é constituído de 16 cromossomos nucleares, na sua totalidade com 12.070.901*pb*. Sendo destes aproximadamente 72% codificantes, ou seja, 8.632.491*pb* e 28% da região codificante corresponde a 3.438.410*pb*. A região não codificante desse organismo é equiparável com a região codificante da bactéria, portanto, pode-se dizer que nesse contexto, a bactéria aproveita mais seu genoma em relação as regiões codificantes que a levedura.

A Tabela A.1 mostra a tradução dos códons em aminoácidos e a degeneração do código genético, pois verifica-se que em muitos casos mais de um códon pode obter o mesmo aminoácido:

- Um códon: Met/começo, Trp.
- Dois códons: Asn, Asp, Cys, Gln, Glu, His, Lys, Phe, Tyr.
- Três códons: Ile, parada.

[2]Organismo geralmente unicelular que não possui núcleo celular, como bactérias e archaeas.
[3]Organismo unicelular ou pluricelular que possui núcleo celular e membranas internas formando organelas citoplasmáticas, como fungos, animais, plantas e protozoários.

Tabela A.1 - Os 64 tripletes de RNA mensageiro que codificam os 20 aminoácidos existentes nos organismos vivos. Conforme as combinações possíveis dos códons (nucleotídeos na 1^a, 2^a e 3^a posição), tem-se o aminoácido cuja sigla está na coluna da 2^a posição. O códon de início de uma região codificadora (*AUG*) também é o mesmo códon que é traduzido para o aminoácido metionina. Os códons de parada (*UAA*, *UGA* e *UAG*) indicam onde a tradução de proteínas termina.

1^a posição do códon (extremidade 5′)	2^a posição do códon				3^a posição do códon (extremidade 3′)
	U	C	A	G	
U	Phe	Ser	Val	Cys	U
	Phe	Ser	Val	Cys	C
	Leu	Ser	parada	parada	A
	Leu	Ser	parada	W	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/começo	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Tabela A.2 - Tabela com as siglas utilizadas para denotar os 20 aminoácidos.

Sigla	Aminoácido	Sigla	Aminoácido
Ala	Alanina	Leu	Leucina
Arg	Arginina	Lys	Lisina
Asp	Ácido Aspártico	Met	Metionina
Asn	Asparagina	Phe	Fenilalanina
Cys	Cisteina	Pro	Prolina
Gln	Glutamina	Ser	Serina
Glu	Ácido Glutâmico	Thr	Treonina
Gly	Glicina	Trp	Triptofano
His	Histidina	Tyr	Tirosina
Ile	Isoleucina	Val	Valina

- Quatro códons: Ala, Gly, Pro, Thr, Val.

- Cinco códons: nenhum.

- Seis códons: Arg, Leu, Ser.

Observa-se que, outra característica denotada é que uma pequena alteração na segunda posição do triplete, por exemplo, traduz diferentes aminoácidos. Veja que, a diferença entre a tradução de Gly, Ala ou Asp é apenas a segunda posição do códon, sendo que G traduz Gly, C Ala e um A traduz Asp.

A.1.2 A evolução da estrutura dos Ácidos Nucléicos e o experimento de Stanley Miller

Outro aspecto referente à codificação dos códons em aminoácidos é o aparecimento destes durante a evolução e origem do código genético. Segundo Trifonov (1999) a ordem decrescente de surgimento dos aminoácidos é:

Gly, Ala, Asp, Val, Pro, Ser, Glu, Leu, Thr, Ile, Asn, Phe, His, Lys, Arg, Gln, Cys, Met, Tyr e Trp.

Miller (1953) obteve os três primeiros aminoácidos (Gly, Ala e Asp) em seu experimento que simulou a atmosfera planetária primitiva (Figura A.4). Esse experimento foi uma tentativa de confirmar que compostos orgânicos (como primeiros passos para origem de vida terrestre) poderiam surgir com os gases que, acreditavam comporem a atmosfera planetária primitiva: metano (CH_4), amônia (NH_3), vapor de água (H_2O) e hidrogênio (H_2). Miller construiu um aparato onde a água era aquecida e o vapor misturava-se com os gases que sofriam descargas elétricas, eram condensados, voltando para o frasco que continha água. Após alguns dias desse circuito, observou a formação de uma substância túrbida no frasco, que com cromatografia descobriu a presença desses aminoácidos.

Após esse experimento, sabe-se que para a tradução desses aminoácidos pelo RNA mensageiro, é necessário que o primeiro nucleotídeo do códon seja G e que o segundo, no caso de Gly e Ala seja G ou C respectivamente, como já mencionado anteriormente. Portanto, a posição desses nucleotídeos no códon infere maior presença desses aminoácidos primordiais em proteínas dos organismos, ressaltando assim a importância desses nucleotídeos na seqüência genética.

A quantidade de GC presente nos Ácidos Nucléicos e consequentemente a quantia de Gly em um organismo pode contribuir para determinar a separação entre duas espécies ou até mesmo a origem (TRIFONOV, 1999). Segundo Trifonov (1999), a es-

Figura A.4 - Stanley Miller no Laboratório de Harold C. Urey na Universidade de Chicago. Disponível em *http : //www.accessexcellence.org/WN/NM/miller.php*. Acessado em 20 de novembro de 2008.

timativa de percentagem de Gly no momento da separação dos reinos de organismos vivos, durante a evolução é:

- Eubacteria (13.5%) - organismos procariotos como as bactérias;
- Archaea (11.5%) - organismos procariotos como as archaeas;
- Protista (10.5%) - organismos eucariotos como os protozoários;
- Fungi (9%) - organismos eucariotos como os fungos e leveduras;
- Plantae/Animalia (8%) - organismos eucariotos como as plantas e os animais.

A Figura A.5 infere como os seres vivos podem ser separados e classificados por sua filogenia, fornecendo exemplos de cada reino citado acima.

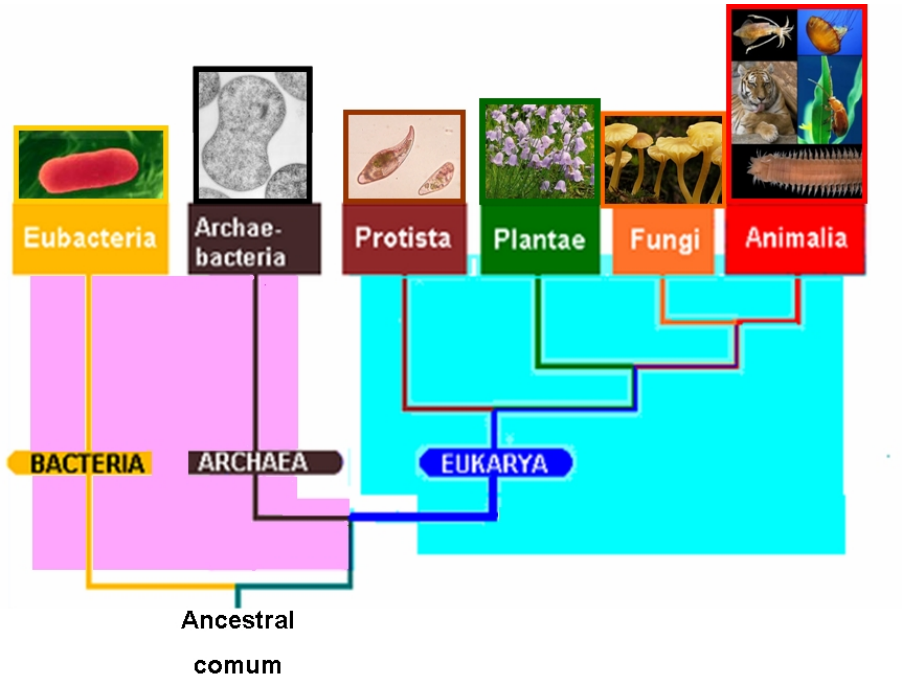


Figura A.5 - Representação dos reinos de organismos vivos. Os reinos Eubacteria e Archaebacteria são constituídos de organismos procariotos e os reinos Protista, Plantae, Fungi e Animalia são constituídos de organismos eucariotos. Adaptado de $http$: $//www.marcobueno.net/resumos/resumo.asp?f\_id\_resumo = 47$. Acessado em 29 de outubro de 2008.

B APÊNDICE B - Exobiologia ou Astrobiologia

Essa ciência, juntamente com outras áreas do conhecimento, por exemplo, Biologia Molecular, Astrofísica e Ecologia procura responder questões: "Como a vida surgiu?", "Estamos sozinhos no Universo?" ou ainda "Qual o futuro da vida na Terra e no Universo?" (HOMECK; RETTBERG, 2007).

Para responder tais questões, inicialmente é necessário contextualizar a palavra vida. A palavra vida tem infinitas definições (HOMECK; RETTBERG, 2007). Segundo Gilmour et al. (2004) são necessárias duas características para ser atribuído vida a um determinado sistema. A primeira diz respeito à capacidade de auto-replicação e o segundo à capacidade de evolução darwiniana, ou seja, quando imperfeições ou mutações ocasionadas no processo de replicação ou cópia estão sujeitas a seleção natural (incorporando-se às características da estrutura ou organismo).

Questões sobre a origem da vida sempre desafiaram a curiosidade dos cientistas. Inicialmente, considerava-se que a vida surgiu espontaneamente na Terra (HOMECK; RETTBERG, 2007). Essa convenção era explicada pela Geração Espontânea, simples hipótese que baseava-se na observação e, consequentemente, na suposição sobre o aparecimento de vermes e moscas na carne em putrefação, de peixes surgindo da lama de rios, sapos e ratos da umidade do solo, etc. O primeiro passo para que esta hipótese não fosse mais usada ocorreu em 1668 quando Francesco Redi (1627 - 1697) demonstrou que, se a carne em putrefação fosse mantida em local adequado e fechado, sem contato com moscas, os vermes não apareceriam (GILMOUR et al., 2004). Outra importante colaboração ocorreu em 1676 através de Anthony van Leeuwenhoek (1632 - 1723) que detectou microrganismos através do primeiro microscópio rudimentar.

A hipótese da Geração Espontânea foi ultrapassada totalmente quando Louis Pasteur (1822 - 1895) apresentou uma série convincente de experimentos na metade do século XIX. Ele mostrou que um caldo ou solução devidamente esterilizado e excluído do contato com microrganismos poderia manter-se assim até que fosse estabelecido um novo contato.

Pasteur conseguiu responder uma importante questão: que a vida não surgira da Geração Espontânea, mas sim, surgira de outra já existente. Mas ao mesmo tempo uma nova questão apareceu: se toda vida veio de uma outra pré existente, de onde

veio a primeira vida? E ironicamente, esta resposta admitia uma origem diretamente relacionada aos intrínsecos processos físicos e químicos da matéria inanimada presente no Universo (GILMOUR et al., 2004; SCHRöDINGER, 1997).

Após algumas missões espaciais e análise de material oriundo de cometas descobriu-se que há matéria orgânica no espaço. Dentre eles, os Cometas são os objetos mais ricos em componentes orgânicos (observações já detectaram cianeto de hidrogênio e formaldeído) (GILMOUR et al., 2004). Sabe-se por exemplo que o Cometa Halley possui uma quantia substancial de material orgânico. Em média, as partículas de poeira ejetadas do núcleo do Halley contém 14% de carbono.

A presença de purinas, pirimidinas, componentes do DNA, e polímeros de formaldeído tem sido inferidos dos fragmentos analisados pelos espectômetros de massa das sondas *Giotto* e *Vega*. Entretanto, não há direta identificação de moléculas complexas orgânicas provavelmente presentes nos grãos de poeira cósmica e no núcleo de cometas. Muitas espécies químicas de interesse para a Exobiologia, por exemplo, foram detectadas no Cometa Hyakutake em 1996, incluindo amônia, metano, etileno e cianeto de metila (FANTONI; MANRICH, 2008).

É possível que, grãos provenientes de Cometas tenham sido importantes fontes de moléculas orgânicas para a Terra primitiva. Cometas com órbitas instáveis poderiam ter colidido com planetas, incluindo a Terra. A colisão do Cometa Shoemaker-Levy 9 com Júpiter em julho de 1994 é um notável exemplo desse tipo de evento (FANTONI; MANRICH, 2008). Provavelmente esses eventos seriam mais comuns a 4 bilhões de anos atrás, quando o número de Cometas orbitando em torno do Sol era maior que o número atual.

C APÊNDICE C - Descrição dos Genomas

O banco de dados utilizado para obtenção dos genomas é o GenBank que é de domínio público sendo amplamente utilizado (BENSON et al., 2002). Cada genoma é registrado por um número (NC) e são usados arquivos do tipo fasta que fornecem a seqüência de nucleotídeos da fita *sense* do DNA (responsável pela tradução dos RNAs).

Para cada cromossomo tem-se diversas informações, o banco de dados fornece uma tabela com diversas informações dos genes presentes neste e a descrição de sua função, ou seja, da proteína traduzida. Na Figura C há uma parte da tabela listada no GenBank para o cromossomo da *E. coli* com a descrição do produto transcrito e onde o gene começa e termina entre outras informações não consideradas aqui.

Product Name	Start	End	Strand	Length	Gi	GeneID	Locus
thr operon leader peptide	190	255	+	21	16127995	944742	thrL
fused aspartokinase I and homoserine dehydrogenase I	337	2799	+	820	16127996	945803	thrA
homoserine kinase	2801	3733	+	310	16127997	947498	thrB
threonine synthase	3734	5020	+	428	16127998	945198	thrC
predicted protein	5234	5530	+	98	16127999	944747	yaaX
conserved protein	5683	6459	-	258	16128000	944749	yaaA
predicted transporter	6529	7959	-	476	16128001	944745	yaaJ
transaldolase B	8238	9191	+	317	16128002	944748	talB
predicted molybdochelatase	9306	9893	+	195	16128003	944760	mog
conserved inner membrane protein associated with acetate transport	9928	10494	-	188	16128004	944792	yaaH
conserved protein	10643	11356	-	237	16128005	944771	yaaW
predicted protein	11382	11786	-	134	16128007	944751	yaaI
chaperone Hsp70, co-chaperone with DnaJ	12163	14079	+	638	16128008	944750	dnaK
chaperone Hsp40, co-chaperone with DnaK	14168	15298	+	376	16128009	944753	dnaJ
IS186/IS421 transposase	15445	16557	+	370	16128010	944754	insL

Figura C.1 - Exemplo de tabela de regiões codificantes dos primeiros nucleotídeos da *E. coli* retirado do GenBank. Na primeira coluna tem-se o nome do produto transcrito, na segunda e terceira colunas tem-se o nucleotídeo inicial e o nucleotídeo final do gene respectivamente.

C.1 Genoma da *E. coli*

A bactéria usada é a *Escherichia coli* tipo $K-12$ e subtipo MG1655 sob o registro no GenBank NC 000913. Essa subespécie de *E. coli* possui uma particular importância pois pode-se determinar experimentalmente propriedades de seus produtos gênicos

provendo fundamental relevância para a anotação de inúmeros genes de outros organismos (RILEY et al., 2006). O conteúdo GC deste cromossomo circular é de 50%, sendo que 85% deste DNA é composto por regiões codificantes. Na análise dos genes envolvidos na Glicólise são:

- Gene **glk**: traduz a enzima Glucokinase que é responsável pela transformação unidirecional da glicose em glicose 6-fosfato;
- Gene **pgi**: traduz a enzima Fosfoglicose Isomerase (Fosfohexose Isomerase ou Glicose-6-fosfato Isomerase) que utiliza a glicose-6-fosfato para transformar em frutose-6-fosfato;
- Gene **pfkA**: traduz a Fosfofrutokinase ou 6-Fosfofrutokinase-1 responsável pela formação da frutose 1,6-bisfosfato a partir da frutose-6-fosfato.
- Gene **pfkB**: traduz a enzima Fosfofrutokinse ou 6-Fosfofrutokinase-2. Este gene é chamado de isogene do gene pfkA pois é responsável pela mesma etapa metabólica;
- Gene **fbaA**: traduz a enzima Frutose bis-P aldolase ou Frutose Bisfosfato Aldolase Classe II utiliza a 1,6-bisfosfato como substrato e obtém duas moléculas - uma gliceraldeido 3-fosfato e uma dihidroxiacetona fosfato (DHAP);
- Gene **fbaB**: traduz a enzima Frutose Bis-P Aldolase ou Frutose Bisfosfato Aldolase Classe I, isogene do gene fbaA;
- Gene **tpiA**: traduz a enzima Triose P Isomerase que transforma a dihidroxiacetona fosfato em gliceraldeido 3-fosfato;
- Gene **gapA**: traduz a enzima Gliceraldeído Fosfato Dehidrogenase ou Gliceraldeído 3-P Dehidrogenase A que é responsável pela formação de 1,3-bisfosfoglicerato com a utilização do gliceraldeido-3-fosfato;
- Gene **pgk**: traduz a Fosfoglicerato Kinase que catalisa a reação 1,3-bisfosfoglicerato para 3-fosfoglicerato;
- Gene **gpmA**: traduz a enzima Glicerol P Mutase (Fosfoglicerato Mutase 1, 2,3-Bisfosfoglicerato-dependente) que realiza a reação 3-fosfoglicerato para 2-fosfoglicerato;

- Gene **gpmM**: traduz a enzima Glicerol P Mutase ou Fosfoglicerato Mutase, Mn-dependente) que é isogene do gene gpmA;

- Gene **eno**: traduz a enzima Enolase metabolisa fosfoenolpiruvato de 2-fosfoglicerato;

- Gene **pykA**: traduz a enzima Piruvato Kinase ou Piruvato Kinase II responsável pela formação do piruvato a partir do fosfoenolpiruvato;

- Gene **pykF**: traduz a enzima Piruvato Kinase ou Piruvato Kinase I ou Frutose 1,6-Bisfosfato-ativado. Esse gene é isogene do pykA.

C.2 Genoma da *T. acidophilum*

O genoma da *Thermoplasma acidophilum* DSM 1728 oriundo do GenBank tem o registro NC 002578. Seu genoma possui 45% de *GC* e 87% é codificante. Este organismo tem 1564905*pb*, sendo considerado o menor genoma (RUEPP et al., 2000).

C.3 Genoma da *S. cerevisiae*

O genoma da *Saccharomyces cerevisiae* é composto de 16 cromossomos nucleares. Na Tabela C.3 há o registro no GenBank, quantidade de pares de bases de nucleotídeos e percentagens de *GC* e de regiões gênicas para cada cromossomo nuclear da levedura.

Para estabelecer os genes envolvidos na Glicólise da *S. cerevisiae* é feito uso do banco de dados SGD (*Saccharomyces Genome Database* - www.yeastgenome.org) que traz informações a respeito da levedura. Os genes selecionados são:

- Gene **glk** 1: traduz a enzima Glicokinase, que fosforila a glicose. Gene encontrado no cromossomo 3 da levedura;

- Gene **pgm** 1: traduz a enzima Fosfoglicomutase que cataliza a glicose 1-fosfato em glicose 6-fosfato. Encontra-se no cromossomo 11;

- Gene **pgm** 2: desempenha a mesma função que o gene pgm 1. Encontra-se no cromossomo 13;

- Gene **pgi** 1: traduz a enzima Fosfoglicose Isomerase que cataliza a glicose 6-fosfato em frutose 6-fosfato e vice-versa. Encontra-se no cromossomo 2;

Tabela C.1 - Registro do GenBank para cada cromossomo da *S. cerevisiae*, $\%GC$ e % de região codificante.

Cromossomo	Registro no GenBank	Quantidade de nucleotídeos	GC	Região codificante
1	NC 001133	230208pb	39%	61%
2	NC 001134	813178pb	38%	73%
3	NC 001135	316617pb	38%	68%
4	NC 001136	1531918pb	37%	73%
5	NC 001137	576869pb	38%	67%
6	NC 001138	270148pb	38%	67%
7	NC 001139	1090946pb	38%	71%
8	NC 001140	562643pb	38%	71%
9	NC 001141	439885pb	38%	70%
10	NC 001142	745745pb	38%	74%
11	NC 001143	666454pb	38%	71%
12	NC 001144	1078175pb	38%	72%
13	NC 001145	924429pb	38%	74%
14	NC 001146	784333pb	38%	73%
15	NC 001147	1091289pb	38%	71%
16	NC 001148	948062pb	38%	72%
Total médio			38%	70.5%

- Gene **pfk** 1: traduz a enzima Fosfofrutokinase que converte a frutose 6-fosfato em frutose 1,6-bisfosfato. Seqüência presente no cromossomo 7;
- Gene **fba** 1: traduz a enzima Aldolase que cataliza a conversão da frutose 1,6-fosfato em gliceraldeido 3-fosfato e uma dihidroxiacetona fosfato (DHAP). Presente no cromossomo 11;
- Gene **tpi** 1: traduz a Triose Fosfato Isomerase que cataliza a conversão da DHAP em gliceraldeido 3-fosfato. Encontrado no cromossomo 4;
- Gene **tdh** 1: traduz a Gliceraldeido 3-fosfato Dehidrogenase que cataliza a reação de gliceraldeido 3-fosfato em 1,3 bis-fosfoglicerato. Encontra-se no cromossomo 10;
- Gene **tdh** 2: isoenzima do gene tdh 1. Presente no cromossomo 10;
- Gene **tdh** 3: izoenzima do gene tdh 1. Encontra-se no cromossomo 7;
- Gene **pgk** 1: traduz a 3-Fosfoglicerato Kinase converte a 1,3 bis-fosfoglicerato em 3-fosfoglicerato. Encontra-se no cromossomo 3;

- Gene **gpm** 1: traduz a Fosfloglicerato Mutase que cataliza a reação de 3-fosfoglicerato para 2-fosfoglicerato. Presente no cromossomo 11;
- Gene **gpm** 2: homólogo do gene gpm 1. Presente no cromossomo 4;
- Gene **gpm** 3: homólogo do gene gpm 1. Presente no cromossomo 15;
- Gene **eno** 1: traduz a enzima Enolase I que cataliza a conversão da 2-fosfoglicerato em fosfoenolpiruvato. Presente no cromossomo 7;
- Gene **eno** 2: traduz a enzima Enolase II que cataliza a conversão da 2-fosfoglicerato em fosfoenolpiruvato. Encontra-se no cromossomo 8;
- Gene **pyk** 1: traduz a Piruvato Kinase que cataliza a conversão da fosfoenolpiruvato para piruvato terminando a Glicólise. Presente no cromossomo 1.

D APÊNDICE D - Transformação da representação de seqüências

Dependendo da análise da estrutura dos Ácidos Nucléicos desejada, torna-se necessário, inicialmente, a transformação da informação genômica contida em uma seqüência genética. A representação simbólica padrão da informação genômica - pelos símbolos dos nucleotídeos na seqüência de moléculas de DNA ou RNA ou pela seqüência simbólica de aminoácidos nas cadeias polipeptídicas correspondentes (para regiões codificantes) - tem bastante relevância na procura por específica informação, mas isso pode limitar o tipo de análise a ser realizada (CRISTEA, 2005).

Segundo Cristea (2005) há três principais dicotomias das propriedades bioquímicas das bases nitrogenadas que permitem arranjá-las nas seguintes classes (ver Figura D.1):

a) Estrutura Molecular: as bases A e G são bases purinas (R) e as bases T e C são bases pirimidinas (Y);

b) Força de ligação entre as bases: entre a base A de uma fita e a base T de outra fita do DNA há duas pontes de hidrogênio, configurando uma ligação fraca - *weak bond* (W). Entre a base G de uma fita e a base C de outra fita do DNA há três pontes de hidrogênio, sendo uma ligação mais estável e mais forte quando comparada a ligação das bases A - T - *strong bond* (S);

c) Conteúdo radical: A e C contêm o grupo amino (CH_3) - M *class*, enquanto T e G contêm o grupo keto ($C = O$) - K *class*.

Baseado nessas classes pode-se inferir outras formas de representar uma seqüência genética do que simplesmemte representação simbólica padrão. É possível, por exemplo, representação tridimensional através da combinação dessas classes, apresentando o DNA de uma outra maneira além das formas usuais (PENG et al., 1994; BULDYREV et al., 1998; CRISTEA, 2005). Através da redução da dimensionalidade de representação pode-se mapear os nucleotídeos, códons ou aminoácidos usando um mapa real de uma dimensão. Os dígitos (0, 1, 2 e 3) podem ser os quatro nucleotídeos e os dígitos (0, 1, 2,..., 63) os 64 códons.

Para Cristea (2005) há 24 possibilidades de atribuições dos dígitos de 0 a 3 para as bases A, T, G e C, sendo que a escolha mais adequada é a apresentada na Tabela D.1.

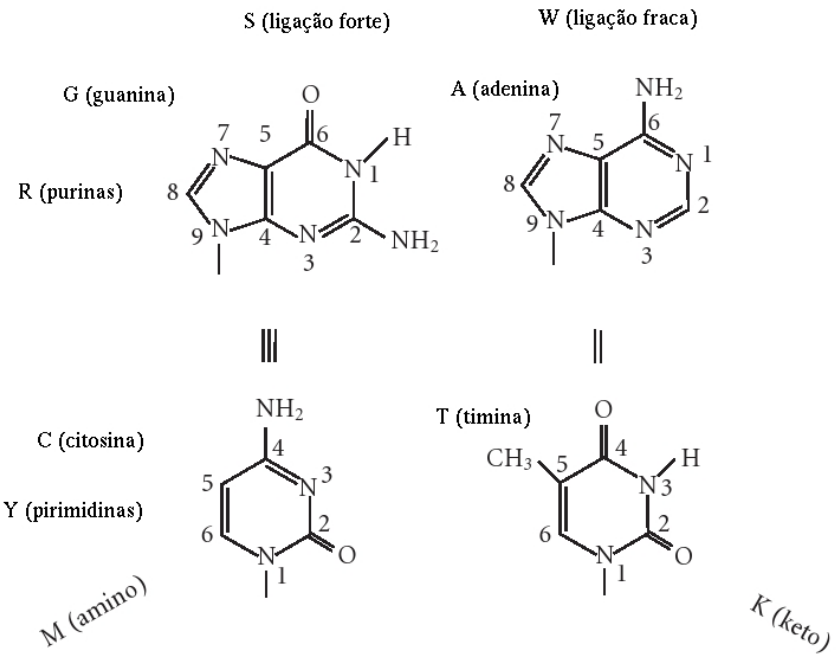


Figura D.1 - As possíveis classificações das bases nitrogenadas. Verifica-se as três pontes de hidrogênio exisitentes entre as bases *G* e *C* (ligação forte) e as duas pontes de hidrogênio existentes entre as bases *A* e *T* (ligação fraca). Adaptada de (CRISTEA, 2005).

Tabela D.1 - Representação dos nucleotídeos em dígitos nas quatro bases nitrogenadas. (CRISTEA, 2005)

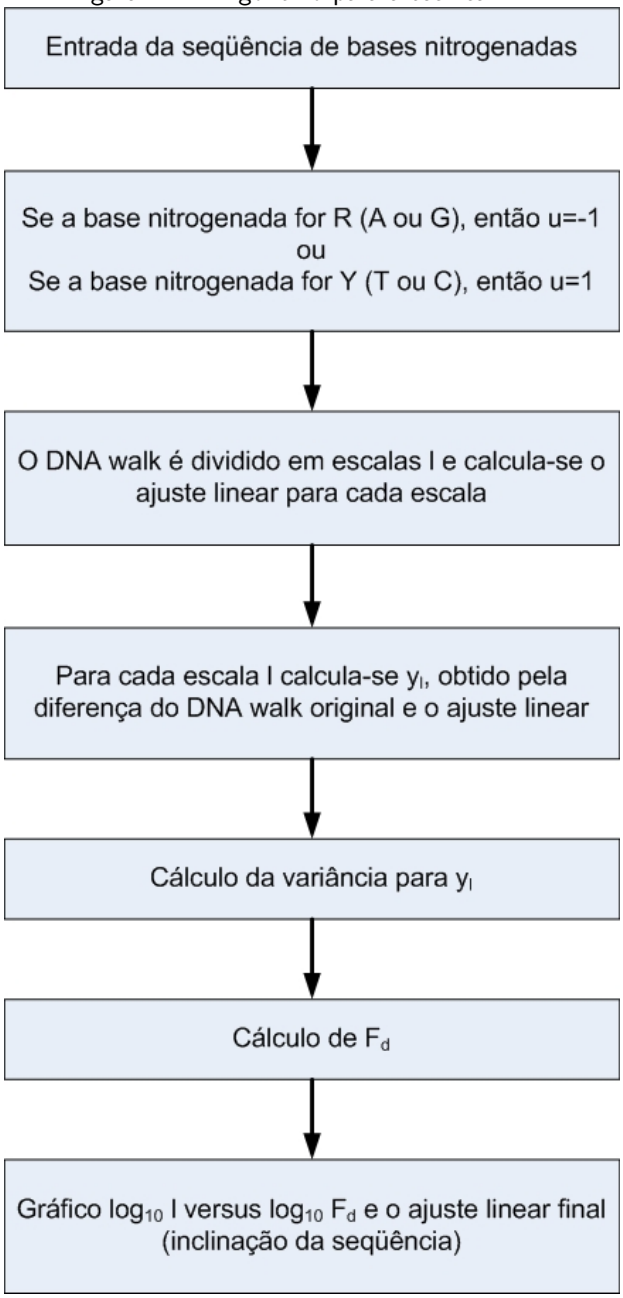
Pirimidinas	Purinas
Timina $(T) = 0$	Adenina $(A) = 2$
Citosina $(C) = 1$	Guanina $(G) = 3$

Aplicações dessas transformações de seqüências podem ser verificadas em Buldyrev et al. (1998), Podobnik et al. (2007) e Bai et al. (2007).

D.1 Algoritmo da técnica DFA

A técnica DFA descrita no Capítulo 2 foi implementada em Matlab®. O algoritmo está descrito no esquema a seguir:

Figura D.2 - Algoritmo para a técnica DFA.



E APÊNDICE E - Redes complexas

Muitos sistemas na natureza e na sociedade podem ser representados através de redes complexas, alguns exemplos podem ser: representação do metabolismo celular, a sociedade, a internet, reações químicas, rede de computadores ligados fisicamente, entre outros exemplos. No estudo de redes complexas usa-se o formalismo da teoria de grafos (ALBERT; BARABáSI, 2002). A estrutura da rede pode ser definida como nó ou vértice e a ligação entre esse vértices como arestas. Em uma relação social, por exemplo, os vértices podem ser definidos pelas pessoas e a ligação entre elas, as arestas, podem ser a amizade ou a comunicação existente.

Enquanto teoria de grafos incialmente foi focada em grafos regulares (cujo número de ligações incidente em um vértice - grau, é o mesmo em todos os vértices), a área de redes complexas estuda desde grafos *randômicos* (primeiramente estudados por Paul Erdös e Alfred Rényi 1959), proposto como o modelo mais simples e grafos irregulares (com diferentes graus dos vértices). De acordo com o modelo de grafo *randômico*, um grafo com N vértices está conectado por L arestas, que são escolhidos aleatoriamente por $N(N-1)/2$ por possíveis arestas. Na Figura E.1 tem-se um exemplo de grafo com 5 vértices e o número máximo possível de arestas para um grafo totalmente conectado.

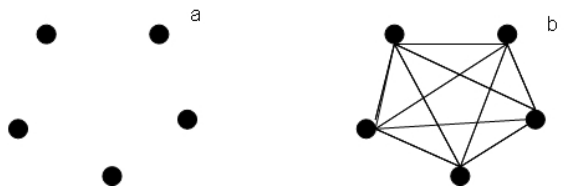


Figura E.1 - (a) Grafo com $N = 5$ vértices isolados. (b) Número máximo de arestas ($L = 10$), para que nesse exemplo, seja um grafo regular totalmente conectado.

Segundo Albert e Barabási (2002) há alguns fenômenos essenciais para compreensão sobre redes complexas: *small world* ou mundo pequeno, *clustering* ou aglomeração e grau de distribuição.

O conceito de *small world* em termos simples descreve o fato que, apesar de frequentemente as redes serem maiores, na maioria dessas há um caminho relativamente curto entre dois vértices. Uma vez que, a distância de separação entre dois vértices cresce mais lentamente do que o tamanho da rede (RODRIGUES, 2007). A

distância entre dois vértices é definido como o número de arestas ao longo do menor caminho conectando eles. Um exemplo de manifestação popular de *small worlds* é o conceito de "seis graus de separação" revelado pelo psicólogo social Stanley Milgram (1967), que concluiu que há um comprimento típico de seis entre pares de pessoas nos Estados Unidos. As propriedades do *small world* caracteriza a maioria das redes complexas, por exemplo, os produtos químicos em uma célula são tipicamente separadas por três reações (considera-se a reação química como sendo uma aresta). O conceito de *small world* não é um indicativo de princípio particular de organização mas é uma evidência que mesmo em redes maiores há uma distância relativamente pequena entre dois vértices (ALBERT; BARABáSI, 2002; NEWMAN, 2003).

A aglomeração ou *clustering* é uma propriedade comum de redes sociais onde há círculos de amigos ou conhecidos no qual todos os membros se conhecem, ou seja, um vértice 1 está conectado a um vértice 2 e este a um vértice 3. Há uma grande chance que o vértice 3 também esteja conectado ao 1, formando assim um ciclo de ordem três (RODRIGUES, 2007). Essa inerente tendência para formação de *cluster* é quantificada pelo coeficiente de aglomeração (*clustering coefficient*) (NEWMAN, 2003). Esse coeficiente verifica a razão entre o número de arestas entre os vizinhos de um vértice i (E_i) e o número máximo de arestas do vértice i conectado a seus vizinhos que é dado por $k_i(k_i - 1)/2$ (ALBERT; BARABáSI, 2002; GERHARDT et al., 2006; RODRIGUES, 2007; NEWMAN, 2003).

$$c_i = \frac{2E_i}{k_i(k_i - 1)} \tag{E.1}$$

O coeficiente de aglomeração da rede total ($\bar{C}$) é a média de todos os c_i's. Uma definição alternativa para $\bar{C}$ é abordada em (GERHARDT et al., 2006). A Figura E.2 apresenta a variação de c_i e $\bar{C}$ para diferentes configurações de um grafo.

O terceiro conceito essencial de redes complexas é o grau de distribuição, que pode ser definido como o número de arestas incidentes em um vértice. Pode-se definir $P(k)$ para ser a fração de vértices em uma rede que tem grau k, ou a probabilidade que um vértice é selecionado uniformemente e aleatoriamente tem grau k. Em um grafo *randômico* onde as arestas são dispostas ao acaso, a maioria dos vértices possuem aproximadamente o mesmo grau, próximo ao grau médio $\langle k \rangle$ da rede (ALBERT; BARABáSI, 2002).

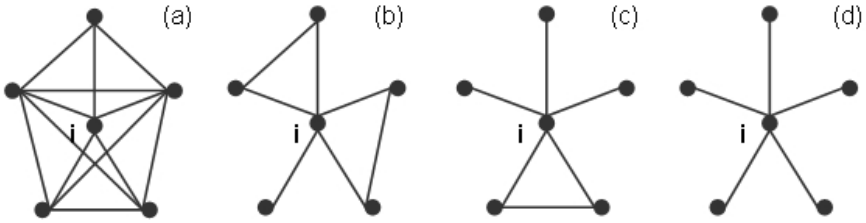


Figura E.2 - Quatro exemplos de configurações possíveis de um grafo com 6 vértices. (a) Grafo totalmente conectado, $c_i = 1$ e $\bar{C} = 1$. (b) $c_i = 0.2$ e $\bar{C} = 0.7$. (c) Para um vértice i com as mesmas conexões mas com número de arestas menor entre seus vizinhos tem-se $c_i = 0.1$ e $\bar{C} = 0.35$. (d) Vértice i onde seus vizinhos possuem grau 0 ou 1, $c_i = 0$ e $\bar{C} = 0$.

E.1 Redes complexas em sistemas biológicos

A seguir é listado alguns exemplos de redes complexas aplicados em sistemas biológicos:

- Jeong et al. (2000) representou redes metabólicas de 43 organismos através de grafos. Neles os vértices são representados pelos substratos, por exemplo ATP e AMP - moléculas relacionadas a energia, e as arestas são representadas pelas reações químicas geradas por enzimas que agem sobre os substratos originando outros. Nesse caso, os grafos são direcionados pelas reações químicas.

- Redes de interação de proteínas associadas a asma estudadas por Hwang et al. (2008) mostra que há vértices altamente conectados (*hubs*) e são responsáveis por muitas propriedades particulares das redes analisadas. Que a descoberta de genes associados a essas proteínas é importante pois pode auxiliar na produção de fármacos que regulam sua expressão.

- A interação entre espécies através de redes ecológicas também é um tipo de representação de redes complexas. Nela as espécies são os vértices e as arestas as próprias interações entre espécies, podendo ser do tipo predação ou relações de mutualismo.

- A dissiminação de doenças infecciosas graves através de redes sexuais também pode ser estudada através de redes complexas e suas propriedades. Os vértices são as pessoas e as relações sexuais as arestas (LILJEROS et al., 2003).

- O dobramento de proteínas também pode ser representado por um tipo de rede complexa onde os vértices são os distintos estados de conformação e há uma aresta entre eles se são obtidos um do outro, apresentando propriedades de *small worlds*. O coeficiente de aglomeração encontrado é maior que para grafos aleatórios, uma diferença que aumenta com o tamanho da rede (SCALA et al., 2001).

- Uma forma de representar uma seqüência de DNA, abordada particularmente nesse estudo, é através de um grafo de tripletes (trios de nucleotídeos). Os vértices são os tripletes e há uma aresta entre dois vértices quando esses são justapostos na seqüência original (GERHARDT et al., 2006).

F APÊNDICE F - Análise Espectral Gradiente

F.1 Análise de padrões gradientes

O principal objetivo da GPA é quantificar assimetrias em escalas locais e globais de um dado perfil temporal, espacial ou espaço-temporal, por meio de uma operação computacional que caracteriza padrões - através das medidas de pequenas e grandes amplitudes em tais padrões - como grades gradientes (ou uma seqüência de matrizes) (ROSA et al., 1999; ROSA et al., 2000; ASSIREU et al., 2002; ROSA et al., 2003; BARONI et al., 2006). Essa grade gradiente é representada por uma matriz denominada, matriz das amplitudes:

$$M_A = M_A(1,1), ..., M_A(i,j), ..., M_A(\sqrt{N}, \sqrt{N}) | i,j \in I \quad e \quad M_A \in \Re \tag{F.1}$$

onde N é o tamanho da série temporal. A matriz quadrada M_A, possui dimensões espaciais (x,y), discretizadas em $\sqrt{N} \times \sqrt{N}$ pontos, com $i \geq i \geq \sqrt{N}$ e $i \geq j \geq \sqrt{N}$.

Usualmente, cada intensidade da amplitude $M_A(i,j)$, na matriz de amplitudes, representa uma medida local de energia espacialmente distribuída (por exemplo: velocidade relativa, taxa de concentração, intensidade de emissão, temperatura, etc.). A flutuação espacial do padrão global $M_A(i,j)$, para um dado instante t, pode ser caracterizada através do campo vetorial gradiente $G_t = \nabla[M_A(x,y)]_t$. Uma flutuação espacial local, entre um par instantâneo de intensidades e pertencentes ao padrão global, é caracterizada por seu vetor gradiente, definido entre cada par de pontos da grade bi-dimensional. Nesta representação, os valores relativos entre as amplitudes (que determinam a norma e a orientação de cada vetor) são mais relevantes do que os seus valores absolutos.

O conceito de simetria usado neste trabalho é o descrito em (ASSIREU et al., 2002). O conceito de simetria gradiente é ilustrado na Figura F.1. A Figura F.1a mostra um perfil totalmente simétrico em relação ao eixo vertical. O perfil é composto por 100 pontos de modo que a sua matriz quadrada tem a forma 10×10 (a Figura F.2 mostra como é feito o mapeamento dos valores da série temporal ou espacial para o formato da matriz quadrada). O padrão mostrado na Figura F.2 corresponde ao perfil mostrado na Figura F.1d.

Levando em consideração um eixo diagonal nesse campo gradiente, podemos observar que para cada vetor local $\underline{v}$ da grade vai existir um vetor correspondente $-\underline{v}$ com

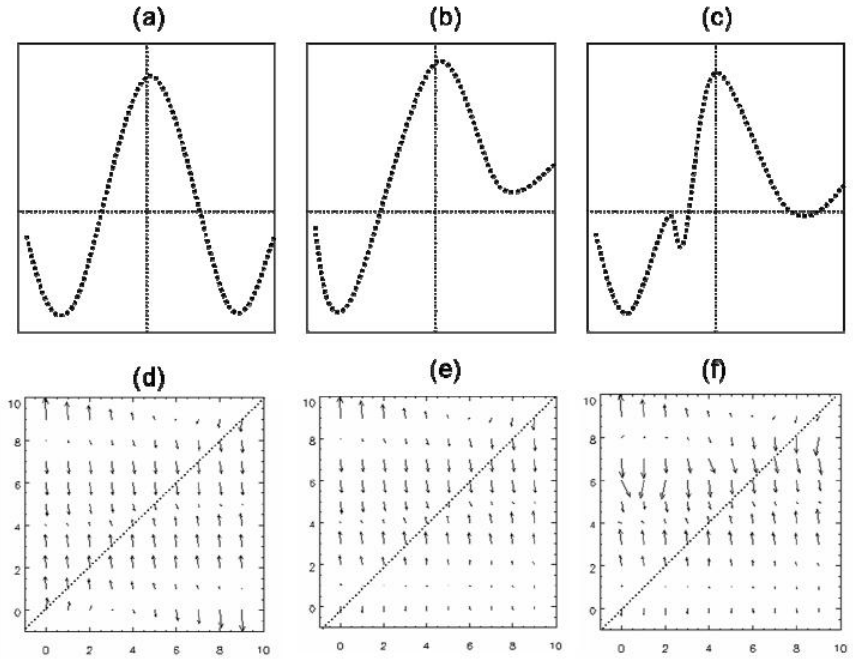


Figura F.1 - Três exemplos de perfis de amplitudes compostas de 100 pontos e seus respectivos padrões-gradientes (BARONI et al., 2009).

o mesmo módulo, mas com fase oposta - denominados vetores simétricos. Assim, se removido os pares simétricos para quantificar assimetria, ao final da operação, não haverá vetores remanescentes no padrão gradiente. As Figuras Figura F.1b e Figura F.1c são exemplos de perfis assimétricos e seus padrões-gradientes podem ser visualizados nas Figuras Figura F.1e e Figura F.1f, respectivamente.

A Figura Figura F.2 demonstra como é feito o mapeamento de uma série temporal ou espacial para a matriz correspondente e o seu respectivo padrão-gradiente. Como exemplo, uma série contendo 100 pontos corresponde a uma matriz de tamanho, onde cada linha da matriz é um grupo de 10 pontos seqüenciais da série analisada, tomados da esquerda para a direita. Como mostrado em Assireu et al. (2002), os valores de G_A calculados a partir dessas matrizes não dependem da direção da série temporal ou espacial como são tomados (da direita para a esquerda ou vice-versa).

F.1.1 Coeficiente de assimetria gradiente

O coeficiente de assimetria (G_A) é parte da técnica conhecida como análise de padrões gradientes, *Gradient Pattern Analysis* (GPA) (ROSA et al., 1999; ROSA et al., 2003). Essa técnica é sensível ao ponto de detectar pequenas diferenças presentes

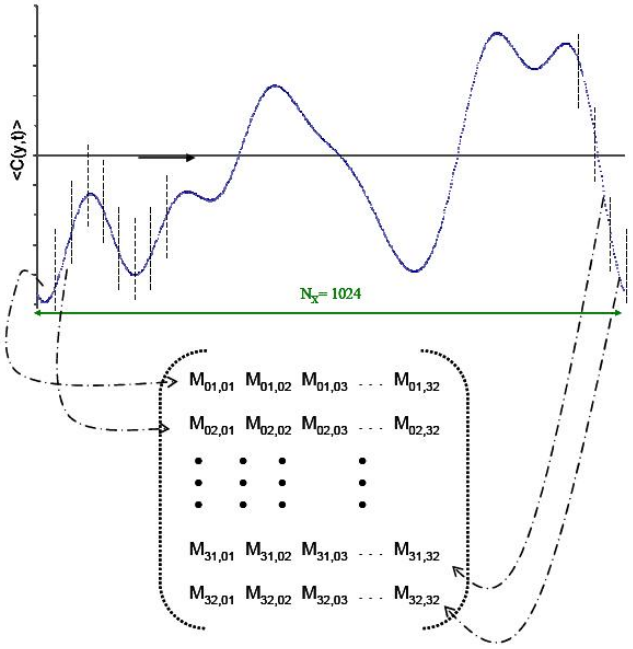


Figura F.2 - Metodologia para mapear uma série temporal ou espacial de tamanho N numa matriz de tamanho $\sqrt{N} \times \sqrt{N}$. No exemplo tem-se uma série com 1024 pontos distribuídos em uma matriz 32×32 pontos.

em uma série espaço-temporal para séries curtas, vantagem que pode ser explorada na análise de segmentos de DNA. Já que para detectar diferenças ou padrões característicos em séries usando-se métodos tradicionais, como Espectro de Lei de Potência, são necessários 10^4 pontos em uma série.

O coeficiente de assimetria gradiente[1] G_A pode ser definido por:

$$G_A = \frac{N_c - N_v}{N_v} \tag{F.2}$$

onde N_v é o número total de vetores assimétricos remanescentes após a remoção dos pares simétricos e N_c é o número de conexões entre esses vetores (gradientes locais). Como o perfil mostrado na Figura F.1a é totalmente simétrico, o gradiente assimétrico não existe e seu coeficiente de assimetria gradiente é nulo ($G_A \equiv 0$). Esse operador computacional mede a quebra de simetria de uma dada grade de flutuação e tem sido usado em várias aplicações (ROSA et al., 1999; ROSA et al., 2000; ASSIREU

[1]Esse é um dos quatro momentos gradientes apresentados por Rosa et al. (2003). É importante ressaltar que esses momentos possuem a propriedade de serem invariantes globalmente em relação à rotação da grade e modulação da amplitude.

et al., 2002; ROSA et al., 2003; BARONI et al., 2006).

Cada gradiente local é obtido pela computação da seguinte forma:

$$\frac{d(M_A(i,j))}{di} \quad e \quad \frac{d(M_A(i,j))}{dj} \tag{F.3}$$

Cada gradiente local da matriz representa a diferença de amplitude do perfil nas direções i (linha) e j (coluna), respectivamente. O espaçamento entre os pontos em cada direção é assumido como um. Uma rotina protótipo desse tipo de gradiente e a rotina clássica do campo de triangulação de Delaunay podem ser encontradas em Assireu et al. (2002).

A conexão geométrica entre os vetores é gerada pela triangulação de Delaunay, tomando o ponto final de cada vetor assimétrico como vértice. Devido às possíveis mudanças nas fases de cada gradiente local (um vetor na grade gradiente), a quantidade N_c é muito sensível para detectar flutuações assimétricas locais na grade (ROSA et al., 1999) e, conseqüentemente, no perfil original. Essa triangulação e sua sensibilidade podem ser observadas na Figura F.3.

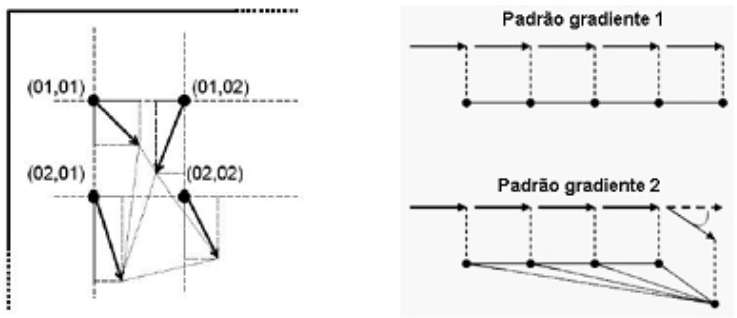


Figura F.3 - (a) Exemplo arbitrário de uma triangulação local de Delaunay entre quatro vetores locais em sua grade gradiente correspondente; (b) exemplo da sensibilidade da triângulação para detectar mudanças na fase do padrão gradiente.

Vários testes em padrões randômicos têm mostrado que o coeficiente G_A quantifica de maneira eficiente o nível de assimetria do perfil. Além disso, G_A é muito mais sensível e preciso para caracterizar estruturas complexas e irregulares do que medidas de correlação (ROSA et al., 1999). Quando não há flutuação assimétrica num padrão gradiente, o número total de vetores assimétricos é zero, e por definição G_A é nulo.

Para um padrão gradiente randômico e totalmente desordenado, G_A tem o valor mais alto e seu o valor cresce assintoticamente até 2 (ROSA et al., 1999), ou seja G_A é máximo quando todas as fases e normas são diferentes na grade gradiente. Para um padrão complexo composto de flutuações assimétricas locais, G_A é não nulo e define classes de flutuação complexa.

A Figura F.4 mostra a sensibilidade da triangulação em detectar qualquer incremento da assimetria local em grades gradientes. As Figuras F.4a e F.4c são as grades gradientes assimétricas obtidas depois de remover todos os pares simétricos dos padrões mostrados nas Figuras F.4e e F.4f, respectivamente. Os respectivos padrões da triangulação são mostrados nas Figuras F.4b e F.4d.

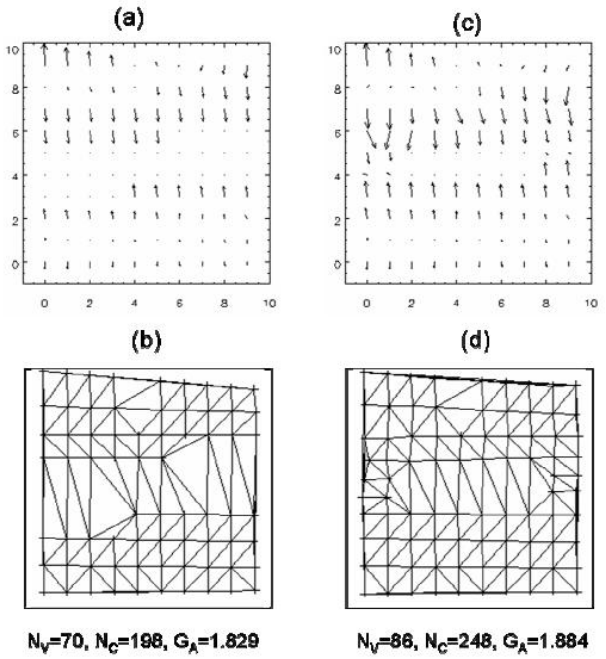


Figura F.4 - Grades gradientes assimétricas e padrão de triangulação respectivos dos perfis de amplitude mostrados na Figuras F.1b e F.1c.

F.2 Escala de Máxima Coerência

Segundo Dantas (2008), a escala máxima de coerência, λ_{mc}, determina qual a máxima escala de dilatação a para uma dada ondeleta-mãe de referência. Com o objetivo de uniformizar a análise para todas as seqüências de DNA utiliza-se a ondeleta-mãe Chapéu Mexicano, de acordo com o método introduzido por Gao e Li (1993):

$$\psi(t) = (1 - t^2)e^{-t^2/2} \tag{F.4}$$

O diagrama tempo-escala é obtido a partir da transformada contínua de ondeleta (CWT - *Continuous Ondeleta Transform*) do sinal $A(t)$:

$$W_\Psi\left[A(t)\right](a,b) = a^{-1/2}\int_{-\infty}^{+\infty} A(t)\Psi(\frac{t-b}{a})dt \tag{F.5}$$

Sobre cada escala que define o diagrama tempo-escala, determina-se a variância da energia ao longo do tempo. A escala máxima de coerência é obtida a partir da propriedade da máxima variância em um gráfico $var[\Psi(a)] \times a$ quando:

$$\frac{\partial var[\Psi(a)]}{\partial a} = 0 \tag{F.6}$$

F.3 Representação Multirresolução

Segundo Dantas (2008), a representação multirresolução de uma série temporal ou espacial (as seqüências genéticas são consideradas neste trabalho como séries espaciais) é realizada através da sua decomposição e reconstrução por meio de uma wavelet que mantenha as características estruturais do sinal em todas as suas componentes , ou seja, é usada a técnica da WMA, para obter as componentes de aproximação da série espacial analisada e uma ondeleta simétrica. A análise realizada é a por aproximação, pois, há interesse na variabilidade das seqüências genéticas, incluindo suas tendências. A decomposição e reconstrução das seqüências genéticas fornece, a cada nível, uma versão aproximada da seqüência relacionada com a escala da wavelet aplicada, que também está intimamente relacionada com uma freqüência-componente do dado analisado.

Para gerar as componentes de aproximação com o mínimo de distorção, é utilizado

a ondeleta biortogonal com ordem de reconstrução igual a seis e ordem de decomposição igual a oito (bior6.8) (DANTAS, 2008). Nesta abordagem os coeficientes da ondeleta biortogonal têm valores discretos, em que as classes de decomposição e reconstrução seguem uma escala diádica. Para maiores detalhes ver (DAUBECHIES, 1992). Após obter as componentes das seqüências genéticas é possível calcular o coeficiente de assimetria para cada série analisada.

G APÊNDICE G - *DNA walk* para éxons e íntrons do genoma nuclear da *S. cerevisiae*

A seguir estão listadas as Figuras que mostram cada cromossomo nuclear da levedura dividido em *DNA walk* gênico e *DNA walk* não gênico.

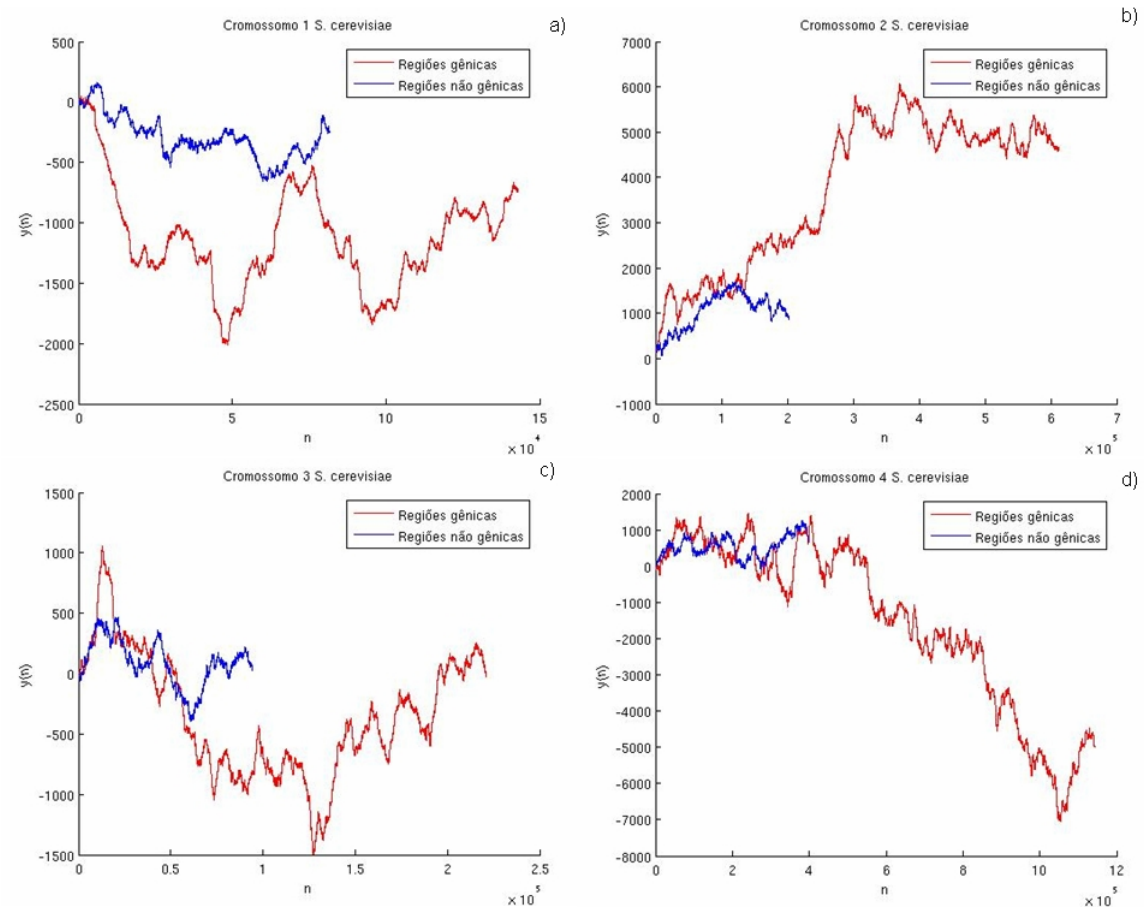


Figura G.1 - *DNA walk* gerado a partir do pré processamento em regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae*. (a) Cromossomo 1, (b) cromossomo 2, (c) cromossomo 3 e (d) cromossomo 4.

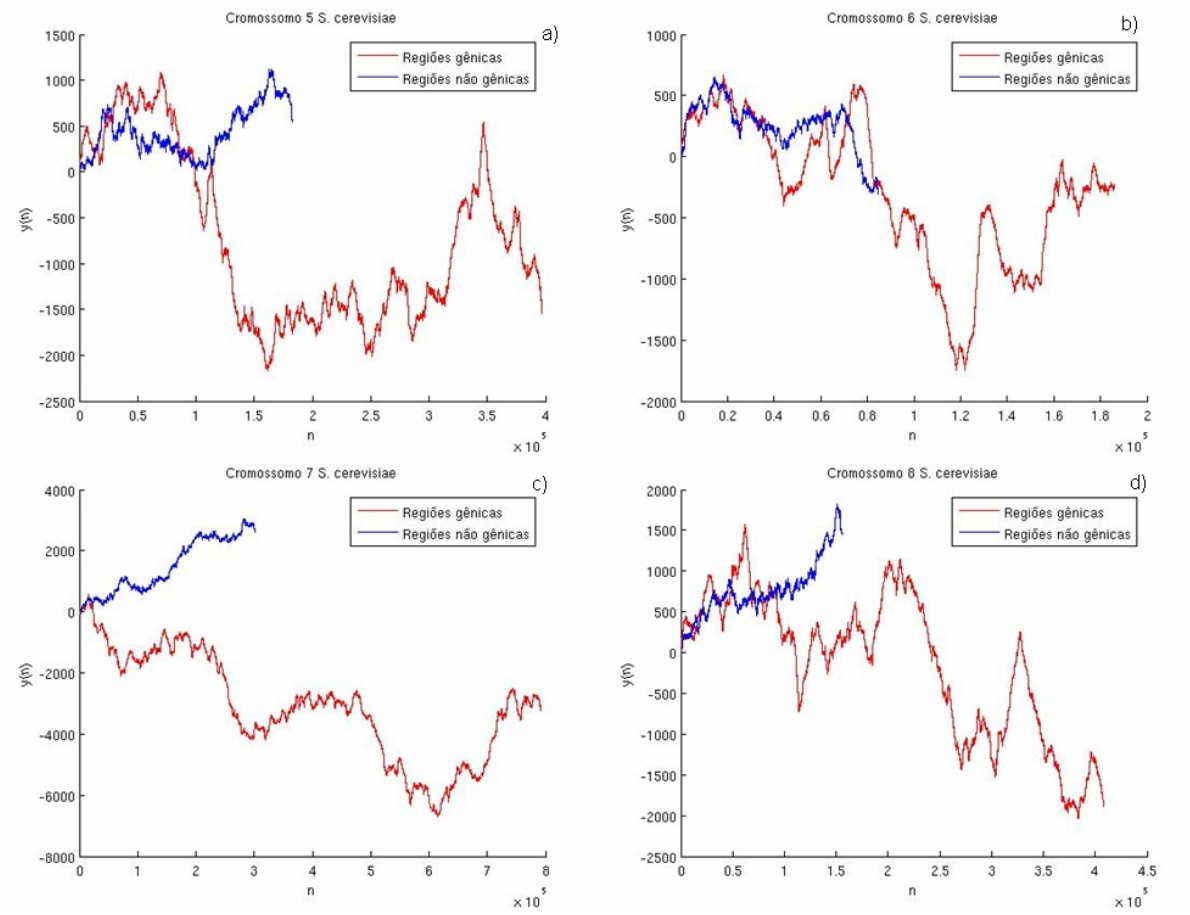


Figura G.2 - *DNA walk* gerado a partir do pré processamento em regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae*. (a) Cromossomo 5, (b) cromossomo 6, (c) cromossomo 7 e (d) cromossomo 8.

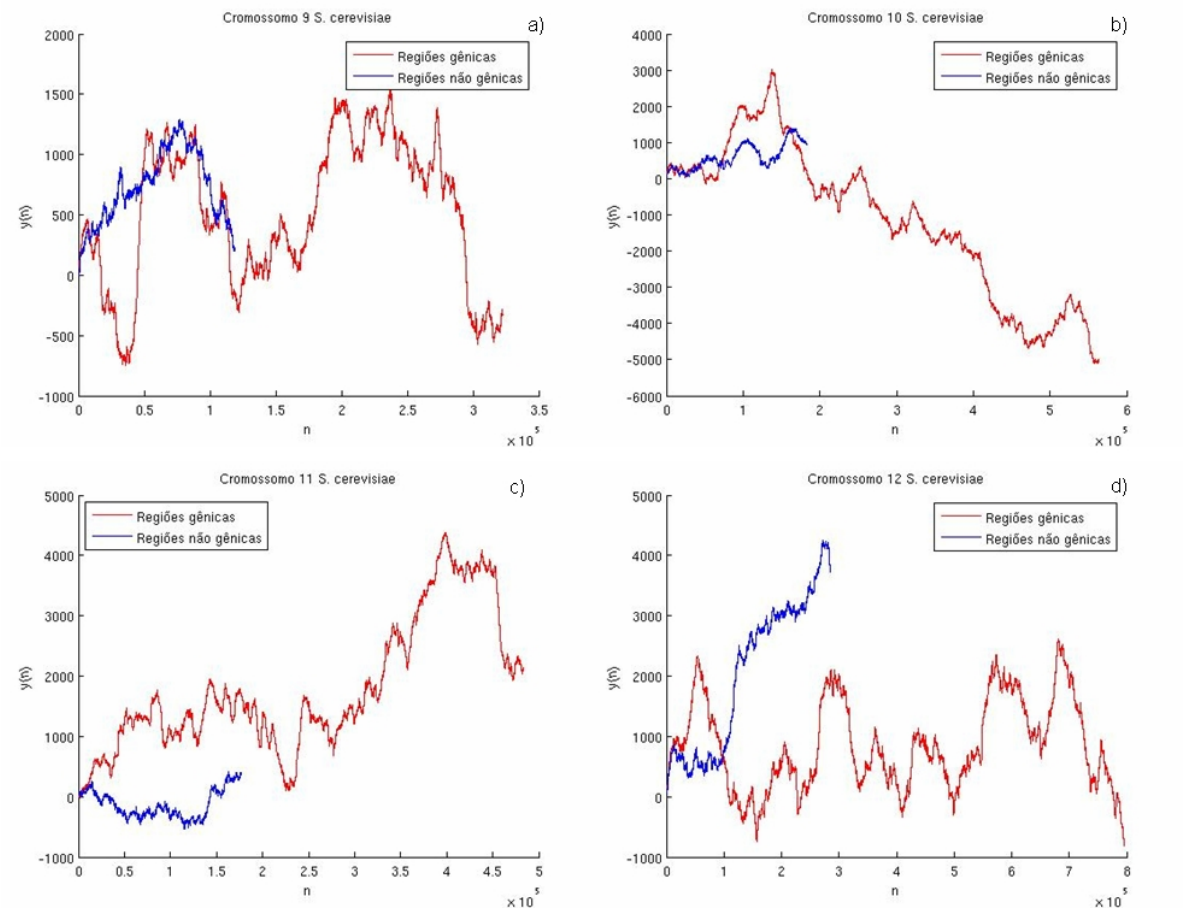


Figura G.3 - *DNA walk* gerado a partir do pré processamento em regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae*. (a) Cromossomo 9, (b) cromossomo 10, (c) cromossomo 11 e (d) cromossomo 12.

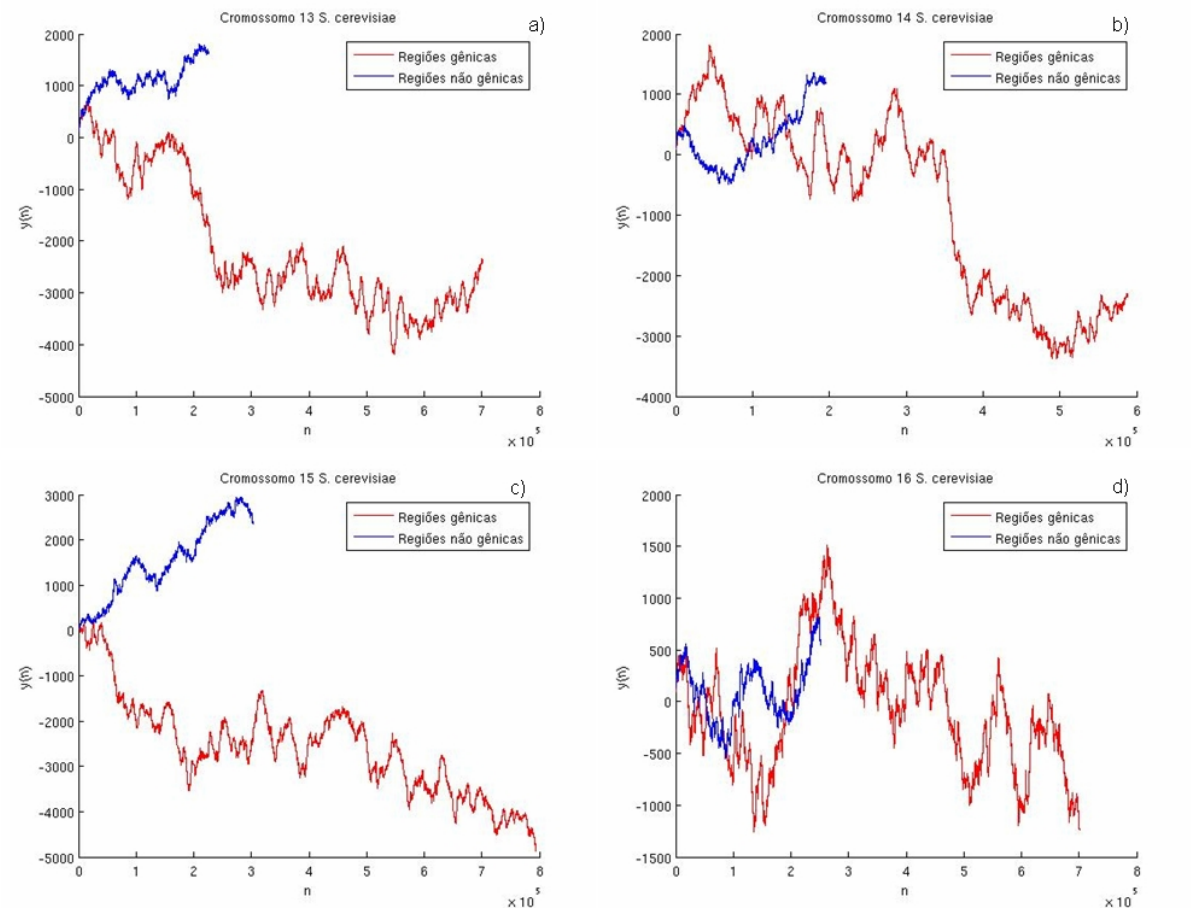


Figura G.4 - *DNA walk* gerado a partir do pré processamento em regiões gênicas e não gênicas de cada cromossomo da *S. cerevisiae*. (a) Cromossomo 13, (b) cromossomo 14, (c) cromossomo 15 e (d) cromossomo 16.

H APÊNDICE H - Coeficiente de Aglomeração com $L = 250$ para *S. cerevisiae*

A seguir é listado os gráficos referentes a Tabela 3.7:

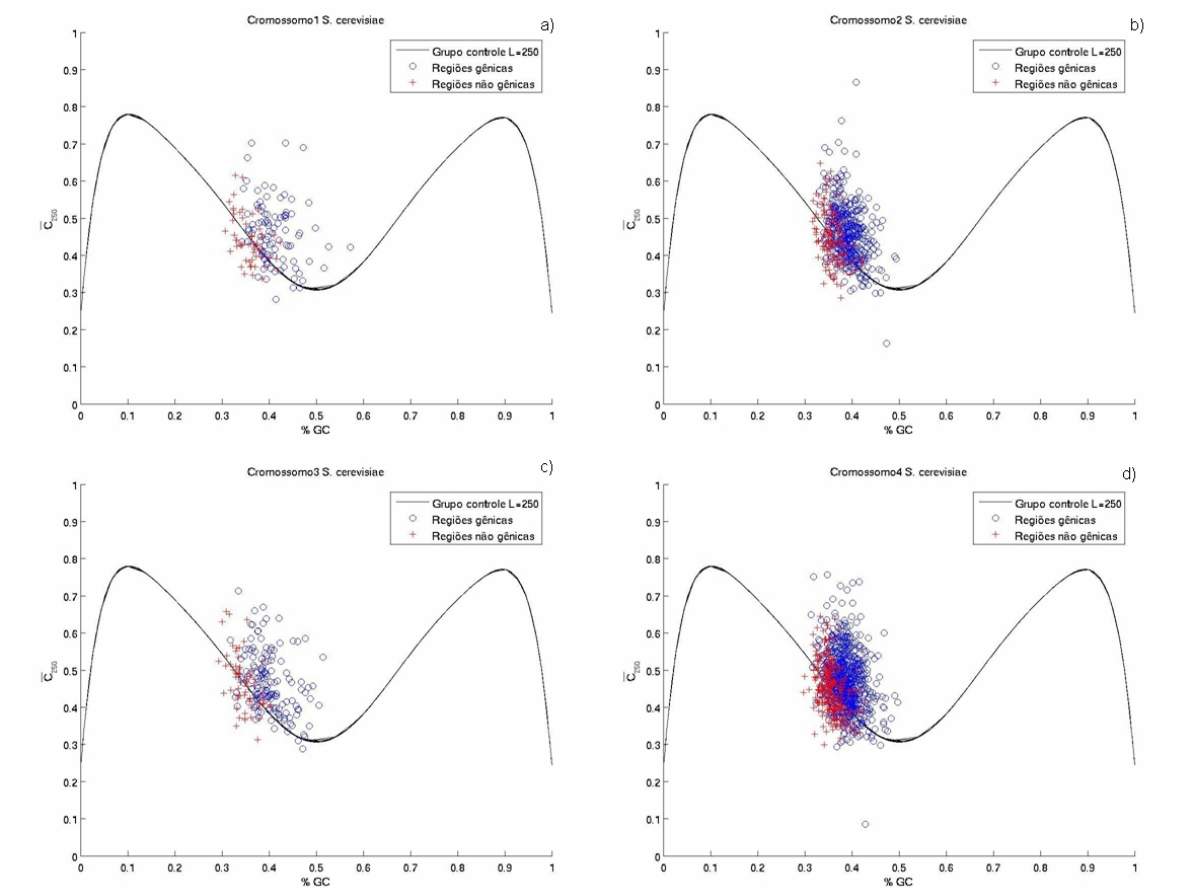


Figura H.1 - $\bar{C}_{250}$ para os grupos gênicos ($\circ$) e não gênicos ($+$) dos cromossomos 1 a 4 da *S. cerevisiae*, respectivamente denominados (a) até (d).

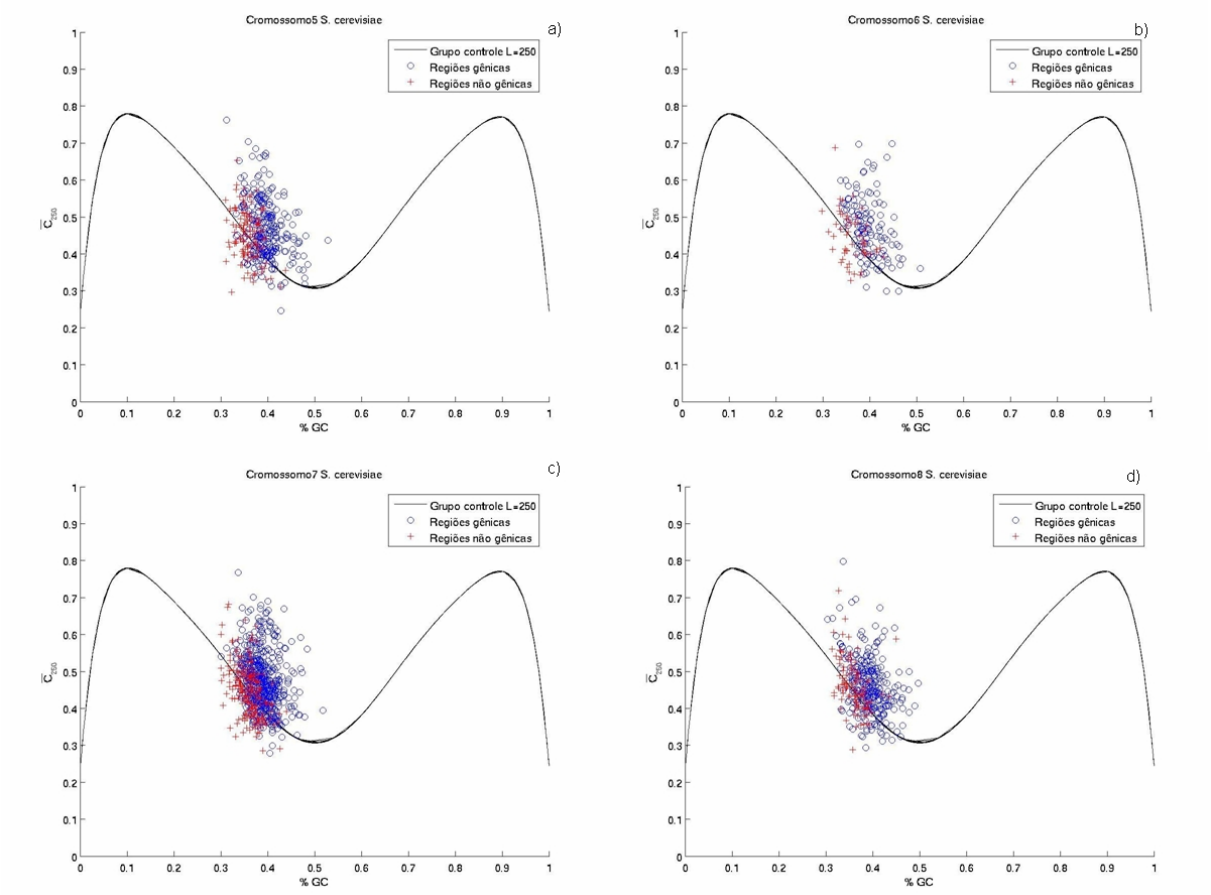


Figura H.2 - $\bar{C}_{250}$ para os grupos gênicos ($\circ$) e não gênicos ($+$) dos cromossomos 5 a 8 da *S. cerevisiae*, respectivamente denominados (a) até (d).

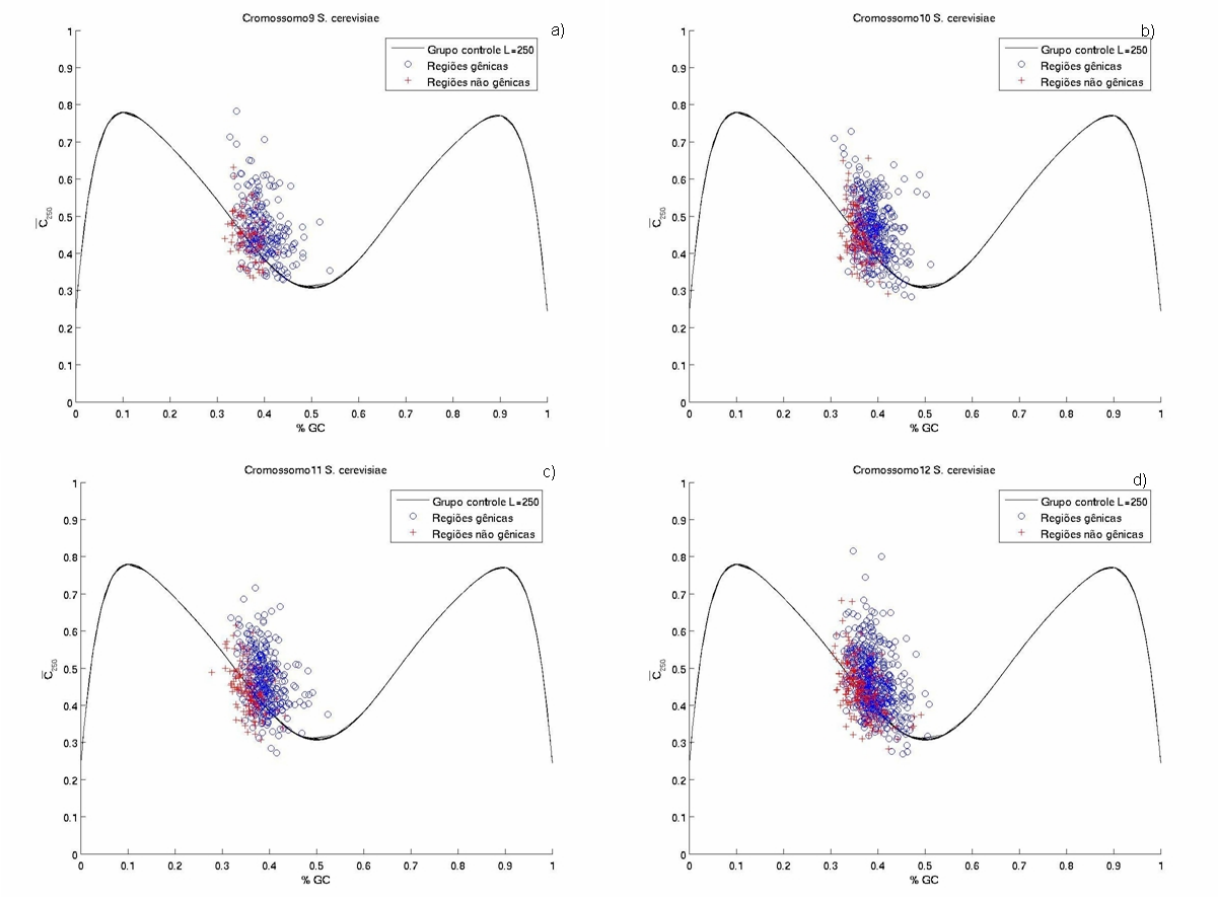


Figura H.3 - $\bar{C}_{250}$ para os grupos gênicos ($\circ$) e não gênicos ($+$) dos cromossomos 9 a 12 da *S. cerevisiae*, respectivamente denominados (a) até (d).

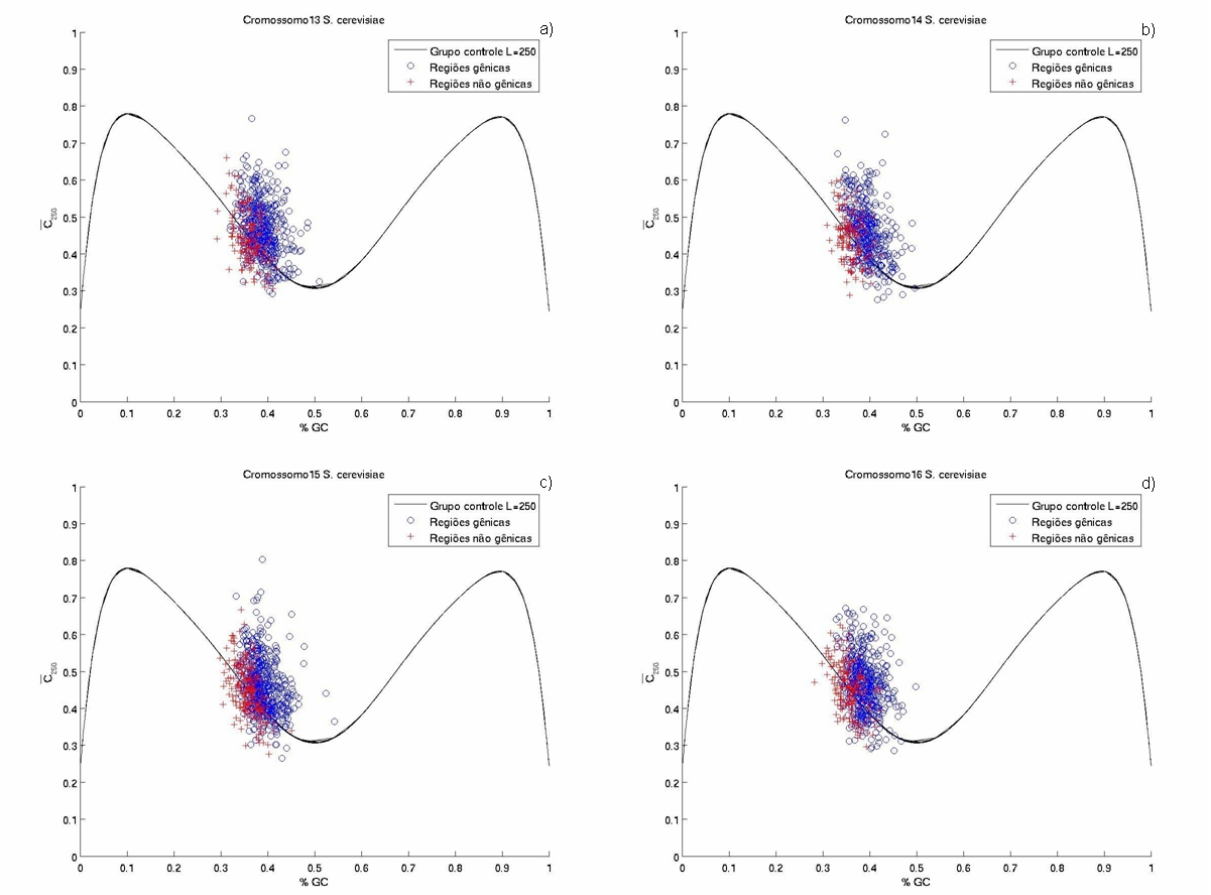


Figura H.4 - $\bar{C}_{250}$ para os grupos gênicos ($\circ$) e não gênicos ($+$) dos cromossomos 13 a 16 da *S. cerevisiae*, respectivamente denominados (a) até (d).

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.