APLICAÇÃO DE ALGORITMOS GENÉTICOS NO ESTUDO DA RADIAÇÃO CÓSMICA DE FUNDO

Cristiane Loesch de Souza
Carlos Alexandre Wuensche
Divisão de Astrofísica
Instituto de Pesquisas Espaciais – INPE
São José dos Campos, S. P.
E-mails: cristiane@das.inpe.br
alex@das.inpe.br

RESUMO

Os Algoritmos Genéticos pertencem a uma classe de algoritmos de otimização numérica que incorporam em um algoritmo computacional a noção de seleção natural. Este trabalho propõe a utilização dos Algoritmos Genéticos na estimativa de parâmetros cosmológicos primários (H_0, λ, Ω_0, Ω_b), a partir do espectro de potência das flutuações de temperatura da Radiação Cósmica de Fundo em Microondas (RCFM). Para isso foi feita a implementação de uma função de ajuste composta por uma combinação de n gaussianas que será comparada a diversas simulações de um espectro de potência, correspondente a diferentes modelos cosmológicos, gerados pelo código CMBFAST.

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

RELATÓRIO DE BOLSA DE INICIAÇÃO CIENTÍFICA

CNPQ /PIBIC

APLICAÇÃO DE ALGORITMOS GENÉTICOS NO ESTUDO DA RADIAÇÃO CÓSMICA DE FUNDO

Cristiane Loesch de Souza
Carlos Alexandre Wuensche
E-mails: cristiane@das.inpe.br
alex@das.inpe.br

Agosto 2000/ Junho 2001

AGRAD

Agradecemos ao CNPQ e ao PIBI(na
realização da Iniciação Científica, e ao IN   am
muito a realização de nosso trabalho.

INDICE

SUMÁRIO

Os Algoritmos Genéticos pertencem a uma classe de algoritmos de otimização numérica que incorporam em um algoritmo computacional a noção de seleção natural. Este trabalho propõe a utilização dos Algoritmos Genéticos na estimativa de parâmetros cosmológicos primários (H_0, λ, Ω_0, Ω_b), a partir do espectro de potência das flutuações de temperatura da Radiação Cósmica de Fundo em Microondas (RCFM). Para isso foi feita a implementação de uma função de ajuste composta por uma combinação de n gaussianas que será comparada a diversas simulações de um espectro de potência, correspondente a diferentes modelos cosmológicos, gerados pelo código CMBFAST.

1. INTRODUÇÃO

Este trabalho propõe a utilização de Algoritmos Genéticos na otimização de problemas físicos, relacionados à Astrofísica e à Cosmologia.

1.1- Algoritmos Genéticos

Algoritmos Genéticos são procedimentos computacionais para procura e otimização, baseados em modelos simplificados de populações genéticas. De uma forma geral, são mecanismos de otimização que incorporam em uma montagem computacional a noção de seleção natural.

A função otimização é um bom exemplo pela qual avaliamos a eficiência ao utilizar um algoritmo, pois este, mesmo com um grande espaço de busca (domínio - conjunto de valores que satisfazem a função) , irá recorrer ao mínimo de recursos de procura para obter uma resposta satisfatória ao problema que o algoritmo se propõe solucionar, ou seja, ele codifica a representação de uma função e procura em seu domínio original as respostas mais prováveis para a solução final do problema.

Os Algoritmos Genéticos combinam as noções de evolução e sobrevivência, e procuram a solução de um dado problema depois de definir uma representação adequada para ele (Goldberg, 1989; Mitchell, 1998).

Eles empregam uma solução candidata (um ponto no espaço de busca) propriamente codificada, denominada *cromossomo*. Estes cromossomos são agrupados em conjuntos denominados *populações*. Uma população definida num intervalo de tempo é denominada uma *geração*. Neste contexto, os Algoritmos Genéticos funcionam através da combinação de pedaços de soluções de uma população, cujas aptidões são determinadas por uma função de avaliação apropriada, de modo que soluções melhores sejam obtidas a cada geração. Em toda geração, uma nova população é criada a partir de uma geração velha e uma parte nova aleatória é experimentada. O Algoritmo Genético explora informações históricas para encontrar novos pontos de procura com uma performance melhorada, ou seja, através das informações da geração anterior é que ele irá se adaptar para produzir uma nova geração mais evoluída, e que melhor se adapte ao sistema.

Uma população de possíveis soluções para o problema em questão evolui de acordo com operadores probabilísticos concebidos a partir de metáforas biológicas, de modo que há uma tendência de que, na média, os indivíduos representem soluções cada vez melhores à medida que o processo evolutivo continue.

Os Algoritmos Genéticos necessitam da informação do valor de uma função objetiva para cada membro da população, que deve ser um valor não-negativo. Nos casos mais simples, utiliza-se justamente o valor da função que se quer maximizar. A função objetiva dá, para cada indivíduo, uma medida de quão bem adaptado ao ambiente ele está, ou seja, quanto maior o valor da função objetiva, maiores são as chances do indivíduo sobreviver no ambiente e reproduzir-se, passando parte de seu material genético às gerações posteriores.

A avaliação da adaptabilidade de cada indivíduo resulta num valor que é denominado “fitness” .

A manipulação de soluções vem da necessidade de testar as probabilidades de evolução, de modo a evitar simultaneamente, locais inadequados no espaço de busca e o aumento do tempo de convergência para a solução ótima. Ela é feita por elementos, denominados Operadores Genéticos, entre os quais os principais são *Seleção, Cruzamento* e *Mutação*, cujos conceitos computacionais assemelham-se a seu uso na Biologia. Suas definições seguem abaixo:

• *Seleção*: mecanismo estimulador dos processos de reprodução assexuada (um indivíduo) e seleção natural. Segundo a Teoria de Darwin a seleção é o processo pelo qual os seres melhor se adaptam ao meio ambiente, produzindo mais descendentes do que seres menos adaptados. Em termos computacionais, gera-se uma população temporária de N indivíduos, extraídos com probabilidade proporcional à adequabilidade relativa de cada indivíduo na população. Neste processo, indivíduos pouco adequados terão alta probabilidade de desaparecer da população nas gerações seguintes, ou seja, serem extintos, ao passo que indivíduos adequados terão grandes chances de sobreviver.

• *Cruzamento:* processo sexuado, envolvendo dois indivíduos. Este processo ocorre entre genes ou sem genes e quando ocorre entre dois cross sites alternativos chama-se segregação cromossômica, a qual ocorre entre parâmetros de Algoritmos Genéticos. A troca de fragmentos entre pares de cromossomos trata-se de um processo aleatório que ocorre com probabilidade fixa de cruzamento e que deve ser especificada pelo usuário do algoritmo.

• *Mutação:* é equivalente à busca aleatória. Basicamente, seleciona-se uma posição aleatória num cromossomo e muda-se o valor do gene correspondente aleatoriamente para outro alelo possível. O processo de mutação é geralmente controlado por um parâmetro fixo que indica a probabilidade de um gene sofrer mutação.

Quando utiliza-se um Algoritmo Genético deve-se definir o tamanho da população, N, além das probabilidades de recombinação, mutação, e outros operadores. Caso ele seja mais sofisticado não há regras claras para a escolha de parâmetros, devendo o usuário definir os que melhor se adaptam ao problema em estudo.

Os Algoritmos Genéticos devem adaptar o sucesso de ações anteriores para melhorar a probabilidade de alcançar suas metas, através da interpretação do modelo, e então executar ações que levarão à solução esperada:

- Devem mapear situações e respostas associadas;
- Devem monitorar gerações a fim de otimizar o "fitness" da população e minimizar o número de iterações;
- Devem escolher os métodos de avaliação que melhor se adaptarem à geração, a partir de resultados de gerações anteriores, corrigindo os parâmetros ainda insatisfatórios daqueles que não tiverem um bom resultado, e;
- Devem selecionar os indivíduos mais aptos.

Os Algoritmos Genéticos imitam a vida, são complexos e impossíveis de se predizer. Seu trabalho de procura, através da seleção das soluções mais plausíveis, funciona como o processo de seleção natural, em relação a evolução, nos quais só quem se adapta sobrevive. O intuito do nosso trabalho é demonstrar a versatilidade e poder dos Algoritmos Genéticos em Astrofísica e Cosmologia, como uma classe forte de técnicas de otimização.

2. MOTIVAÇÃO

Os Algoritmos Genéticos serão utilizados na otimização de métodos de procura de parâmetros cosmológicos. Para isso, será analisado um conjunto de medidas da Radiação Cósmica de Fundo em Microondas (RCFM) para estimar valores dos parâmetros cosmológicos primários, utilizando-se Algoritmos Genéticos.

Um dos objetivos da Cosmologia é estimar o tamanho, a idade, a composição química e a geometria do Universo, ou seja, traçar um perfil da evolução do Universo desde o Big Bang.

Uma questão fundamental é entender como, de um composto de uma mistura de gás e radiação em equilíbrio térmico ele passou para o estado extremamente complexo e diversificado que vemos hoje, com galáxias, estrelas e planetas concentrados em partes do céu e regiões vazias em outras.

Não se sabe em que época as galáxias e as primeiras estrelas se formaram, sabe-se que isso deve ter acontecido quando o Universo tinha entre 10 milhões e 1 bilhão de anos de idade. Entretanto, nada se sabe também sobre a taxa de formação ou a distribuição de galáxias naquela época. O que se conhece é o Universo mais jovem, com cerca de 300 mil anos de idade, um período que é explorado a partir das informações que se extrai do estudo da RCFM. Através da comparação dos resultados destes estudos, com outros dados coletados sobre a distribuição de galáxias e aglomerados é que poderemos entender como o Universo evoluiu até o seu estado atual.

2.1- Radiação Cósmica de Fundo em Microondas

A RCFM é uma emissão em rádio freqüência, com o pico na faixa de microondas (~115 GHz), que foi detectada em 1965 por Arno e Robert, ELA, é uma radiação altamente isotrópica, com um espectro de corpo negro com temperatura de aproximadamente 2,726 ± 0,0010, distribuída de forma praticamente uniforme por todo céu (Milone,1999; Oliveira,2000).

A RCFM pode ser estudada através de três características, são elas:

- Através do espectro de potências;
- Através da distribuição angular, e;
- Através da polarização.

Em nosso trabalho nos deteremos apenas nas duas primeiras.

A RCFM é uma forte evidência de que o Universo um dia foi mais denso e quente do que é hoje. Para produzir uma radiação com as características da RCFM, o Universo deveria apresentar uma forma completamente diferente, pois, nessa época, não existiam planetas, estrelas e galáxias. Ele devia estar completamente preenchido por um "plasma" composto de radiação e partículas elementares: o "plasma primordial".

Essa radiação é considerada como um resíduo da época do Big Bang, onde a história do Universo começa, cerca de 10 a 15 bilhões de anos atrás.

A RCFM, formada cerca de 300 mil anos após a criação, foi se resfriando por causa da expansão e, hoje, sua temperatura é de cerca de 2,726 K. Esta temperatura corresponde à faixa de microondas no espectro eletromagnético e a radiação vem de todas as regiões do céu. E a potência dessa emissão é distribuída num grande intervalo de freqüências, ao invés de estar concentrada numa única freqüência, como acontece com um transmissor de rádio comum.

O estudo da RCFM é feito medindo-se a emissão ao longo desse intervalo de freqüência em diferentes regiões do céu. Acredita-se, hoje, que a evolução do Universo, durante o processo de formação de estruturas, não afetou a RCFM, de modo que ela possui praticamente as mesmas características de quando foi criada, exceto pela temperatura, que vem abaixando por causa da expansão.

A RCFM é considerada um resíduo cósmico praticamente intocado, e nos oferece uma excelente oportunidade de estudar detalhes do Universo jovem. Para entender melhor a origem e a evolução das estruturas que vemos hoje, é essencial que se saiba como eram as condições físicas existentes naquela época e as características da RCFM são diretamente dependentes dessas condições, de modo que estudar a RCFM nos permitirá entender melhor a física do Universo jovem.

Hoje, as observações da RCFM indicam que nos primeiros instantes do Universo a matéria e a radiação encontravam-se quase que uniformemente distribuídas por todo espaço, mas que apresentava também pequenas distorções (de 1 parte em 10^5), a partir das quais acredita-se que as gigantescas estruturas que vemos, hoje, se formaram.

Embora a intensidade da RCFM seja extremamente uniforme em todo o céu, a distribuição de galáxias é extremamente irregular, com flutuações na densidade de galáxias por volume sendo extremamente alta. Mapas dessa distribuição revelam um padrão notável de estruturas com regiões vazias e aproximadamente esféricas. Essa distribuição não uniforme de galáxias, filamentos e vazios é conhecida como estrutura em grande escala.

Ao estudar as estruturas em grande escala (1.000.000 anos) junto à RCFM (300.000 anos) existem perguntas a serem respondidas:

- Qual a natureza e a distribuição da matéria do Universo?
- Qual é a densidade média da matéria do Universo?
- Essa densidade é maior, menor ou igual à densidade critica?
- Onde e quando as primeiras estruturas se formaram?

A combinação de medidas das estruturas em grande escala e das flutuações de temperatura na RCFM são o conjunto de informações mais detalhado e importante, existente atualmente para estudar a origem e evolução do Universo.

2.2- Parâmetros Cosmológicos

Parâmetro de Hubble

$$H(t) = \frac{\dot{R}}{R}$$

Parâmetro de aceleração

$$q(t) = -\frac{\ddot{R}R}{\dot{R}^2} = -\frac{\ddot{R}}{RH^2}$$

Densidade crítica

$$\rho_c(t) = \frac{3H^2}{8\pi G}$$

Parâmetro de densidade

$$\Omega = \frac{\rho(t)}{\rho_c(t)}$$

O subscrito 0 (H_0, q_0, Ω_0,$\rho_{c,\,0}$) denotará o valor atual destas quantidades. Por exemplo, H_0= 100h Km s^{-1} Mpc^{-1} e $\rho_{c,\,0} \sim 2 * 10^{-29} h^2$ g cm^{-3}.

Qual a importância destes parâmetros?

A *Constante de Hubble* (H_0 = 100h Km s^{-1} Mpc^{-1} , $0,4 < h < 1,0$) permite determinar as distâncias das galáxias e estimar a idade do Universo. Ela pode ser obtida através de medidas das mais variadas observáveis cosmológicas, como por exemplo, as velocidades de aglomerados de galáxias, os estudos de redshift, etc.

A *Constante Cosmológica* surgiu a partir de um conflito entre o conceito de um Universo imutável e o de um Universo evolutivo. Esta constante, vem sendo estudada e, quando considerada não nula, possibilita a existência dos modelos mais variados de Universo (aberto, fechado, ...).

A *Densidade de Matéria Bariônica* é a composição química da matéria "ordinária" do Universo constituído essencialmente de prótons e nêutrons. O valor de Ω_b é determinado pela época e duração da fase da nucleosíntese primordial.

A Densidade de Matéria do Universo Ω_0 : $\Omega_0 = \Omega_b + \Omega_\Omega + \Omega_{DM} + \ldots$

2.3- Utilização

Existem diferentes métodos de minimização de funções e extração de parâmetros em espectro de potência de RCFM que obtiveram sucesso, como é o caso do Simulated Annealing. O Simulated Annealing é uma rotina de otimização que quando comparada a outras rotinas, apresenta maior eficiência e rapidez, embora muitas dessas rotinas se pareçam, como é o caso dos Algoritmos Genéticos.

E um dos métodos mais proveitosos para explorar o potencial de muitos dos algoritmos novos é encontrar dados sobre a RCFM.

Procura-se, então, mostrar a eficiência da "Procura" (que é atingida se o número de soluções encontradas é pequeno comparado às possibilidades de solução), com a utilização dos Algoritmos Genéticos no processo de determinação de parâmetros cosmológicos a partir do espectro de potência de flutuações de temperatura da RCFM, visando provar sua total eficiência, além da possibilidade de uma maior compreensão sobre a evolução do Universo jovem.

3. DESENVOLVIMENTO

Foi necessário um estudo de introdução à programação de computadores, para que fosse adquirido o domínio da linguagem Fortran. Esse estudo permitiu um contato com o conceito genérico de algoritmos e com a "gramática" das linguagens de programação e do compilador Fortran. Após a familiarização com os comandos e variáveis, trabalhamos durante algum tempo num código da passagem meridiana de objetos celestes, utilizado no planejamento de vôo do experimento o BEAST, lançado a bordo de balão estratosférico.

Foi, então, elaborado um estudo sobre Algoritmos Genéticos e otimização de funções, a partir do qual pôde-se iniciar a adaptação de um código de otimização baseado em Algoritmos Genéticos, denominado PIKAIA, utilizado em nosso trabalho para a implementação e testes de algumas funções para estimar os parâmetros cosmológicos citados anteriormente.

Pikaia (Charboneau,1995) é um Algoritmo Genético de otimização bastante flexível, podendo ser modificado de acordo com as necessidades do usuário.

Essencialmente, o código PIKAIA:

- associa o número de indivíduos da população,
- determina o número de gerações para essa população,
- predestina a população a um número de gerações com valores parametrizados,
- parametriza as características da população para cada geração,
- coloca o número de dígitos (cromossomos) retidos no código do fenótipo e no genótipo, e,
- define uma relação linear com a posição e a criação de probabilidades.

Paralelamente ao conceito de Algoritmo Genético, estudamos o conceito de otimização de funções (Press et al., 1992), com o objetivo de implementar, no código PIKAIA, uma função adequada à estimativa dos parâmetros cosmológicos mencionados na seção anterior.

Compararemos a saída de cada geração do Pikaia com um espectro sintético gerado por outro código de domínio público: CMBFAST.

CMBFAST calcula o espectro de potência linear da anisotropia da RCFM a partir da integração de fontes da anisotropia dentro do cone de luz dos fótons.

Foram feitas diversas simulações com CMBFAST dos parâmetros cosmológicos, para estudar a dependência da forma do espectro com a variação dos parâmetros. Foram realizadas variações:

- da Constante de Hubble (H_0);
- dos Neutrinos;
- das Flutuações (adiabáticas e isocurvatura);
- da Composição "química" (densidade):

• Ω_b = Densidade de Matéria Bariônica,

• Ω_c = Densidade de Matéria Escura Fria,

• Ω_v = Constante Cosmológica,

• Ω_n = Densidade de Matéria Escura Quente; e

- do Índice espectral das flutuações.

Após a geração dos espectros, implementamos a função, anteriormente definida, em PIKAIA para a comparação com as estimativas feitas com os diferentes valores apresentados na literatura.

Paralelamente a este trabalho foi também realizado um estudo orientado de Astronomia e Astrofísica Fundamental (Oliveira, K. S.; Saraiva, M. F. O., 2000), com o objetivo de proporcionar um maior conhecimento do assunto, auxiliar na compreensão dos termos utilizados no decorrer do desenvolvimento da Iniciação Científica e introduzir conceitos fundamentais relativos às diversas áreas de Astronomia e Astrofísica.

4. METODOLOGIA

4.1 - Fortran 77

Com o estudo de introdução à linguagem de programação, testou-se e otimizou-se um código, em Fortran 77, que calculava a hora do trânsito de uma determinada fonte no céu, em função das coordenadas da fonte (ascensão reta, declinação), latitude e longitude para um certo local desejado. Um exemplo do arquivo de entrada possui o seguinte formato (sem as especificações escritas, como Local, Data, etc):

Local: Fort Sumner

Data: 5 1 1998

Latitude: 34 28 12

Longitude: -104 13 12

Fonte 1: GUM Ascensão reta: 3.5 15. 20.749 Declinação: 71. 50. 01.

Fonte 2: Alkaid Ascensão reta: 3.5 13. 47.51 Declinação: 49. 18. 47.

Fonte 3: Alfa Leonis Ascensão reta: 3.5 10. 08.37 Declinação: 11. 58. 01

A Figura 1 mostra um dos resultados obtidos, de algumas fontes de interesse (UT x elevação), para o dia 22/10/2000.

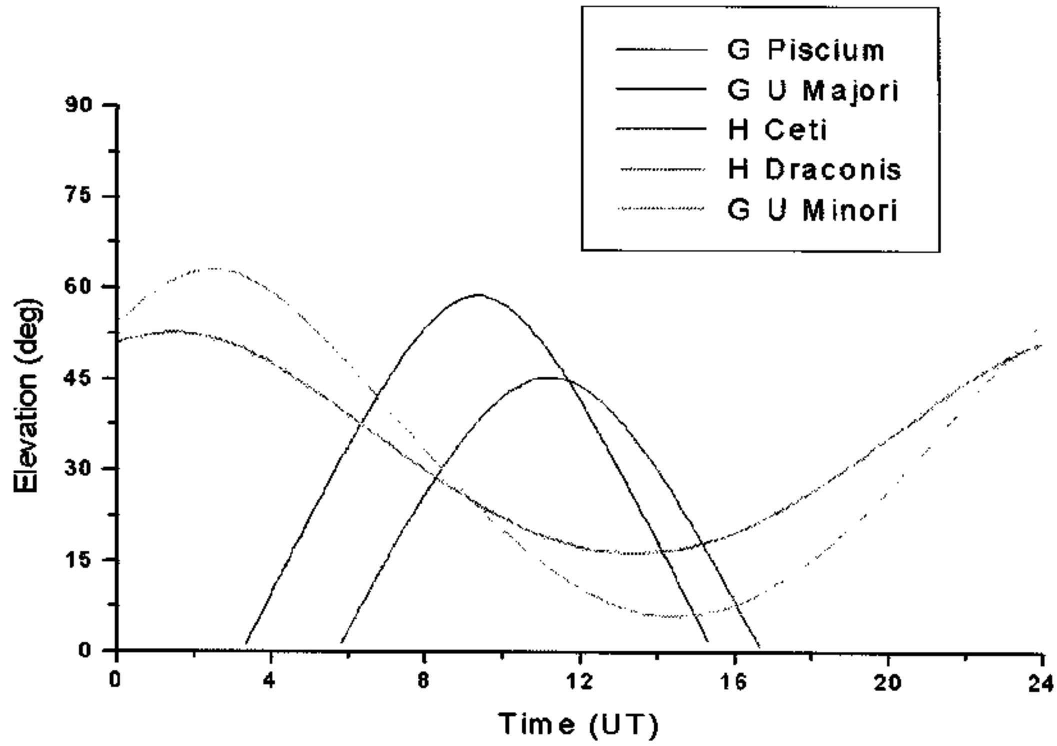


Figura 1: Horário de trânsito de algumas fontes de interesse, cujo máximo fica próximo ao campo de visada do experimento BEAST, para o dia 22/10/2000

4.2 – Algoritmos Genéticos

O passo seguinte foi estudar Algoritmos Genéticos, como uma categoria de algoritmos de procura. Suas principais características são:

- minimizar o número de iterações, do problema ou da função, necessárias para localizar uma solução satisfatória;

- seu método de otimização é semelhante à evolução biológica, e;

- adapta-se ao sistema buscando resolvê-lo diante das diversas dificuldades para encontrar uma melhor solução, ou seja, codifica a representação de uma função e procura em seu domínio original as respostas mais prováveis para a solução final do problema;

A seguir apresenta-se uma ilustração de como funciona na prática um modelo de Algoritmo Genético, para um determinado problema (Wahde,1997).

Dada $f(x,y)=e^{xy}$, um problema para o qual a solução deve ser $f(x,y)=e$, onde $x,y\in[0,1]$, sabe-se que a resposta deve ser x =1 e y =1. Além disso, o código aqui utilizado excluirá o par $(1,1)$, justamente a solução que estamos procurando! Esse exemplo servirá para ilustrar como o código genético escolhido pode influenciar de forma negativa ou positiva na eficiência de um GA.

Devem-se, então, estabelecer as regras que serão utilizadas para sua realização:

1º - Nossa *população* será composta por pares ordenados (x, y), tal que x, y sejam números reais no intervalo fechado $[0,1]$, que possuirão três algarismos significativos após a vírgula.

2º - Os *indivíduos* serão os elementos que irão compor a população.

3º - O *código genético* é necessário para representar todo e qualquer indivíduo da população, e será estabelecido como uma relação unívoca que associe a cada par ordenado um número natural n $(000000\leq n\leq 999999)$, tal que os três primeiros algarismos correspondam à representação do número x e os três últimos à do número y.

4º - O *cromossomo* será um número n qualquer e, finalmente, qualquer algarismo entre zero e nove constituirá um *gene*.

5º- Regras de seleção serão estabelecidas para distinguir candidatos à solução daqueles que não se encaixam nela, utilizar-se-á três critérios:

a) perfil p,

$$p(x,y) = e^{xy},$$

(O perfil do candidato a ser solução de nosso problema deve ser próximo de $p(x,y) = e$.)

b) desvio d, e,

$$d(x,y) = e - p(x,y),$$

c) adaptação a

$$a(x,y) = 1/[1 + d(x,y)].$$

Assim, deveremos ser capazes de maximizar a adaptação e minimizar o desvio de um indivíduo para que ele se torne candidato em potencial à solução de nosso problema.

6º- Deve-se, então, selecionar aleatoriamente um número limitado de indivíduos ao qual denominaremos de *geração mãe*. A partir dessa geração, vamos obter gerações de descendentes até que, um ou vários de seus indivíduos, possuam o perfil desejado. Para tanto, é necessário selecionar e combinar entre si esses indivíduos. Visando esse fim, vamos definir três operações básicas que denominaremos de *reprodução, acasalamento e mutação.*

a) A *reprodução* de um indivíduo será feita mediante a cópia exata de seu cromossomo, o qual será transmitido sem modificações para a geração seguinte.

b) O *acasalamento* será realizado após feita a seleção de dois indivíduos (casal doador), os quais trocarão material genético entre si mediante a quebra de seus cromossomos e permuta de seus genes. Os cromossomos surgidos desse acasalamento serão transmitidos para a geração seguinte.

c) Na *mutação* selecionamos um indivíduo e alteramos um ou mais genes em seu cromossomo. O cromossomo mutante será transmitido para a geração seguinte.

As operações reprodução, acasalamento e mutação, tal como foram definidas, impedem que, de geração em geração, haja aumento no número de indivíduos, isto é, sempre teremos um número predeterminado de indivíduos, candidatos à solução, competindo entre si.

As operações:

- Para operar a reprodução, podemos dispor os N cromossomos de uma dada geração em ordem crescente de adaptação, e determinar que os k indivíduos que possuam as melhores adaptações sejam reproduzidos para a próxima geração.
- Para selecionar casais para acasalar, calculamos o somatório total da adaptação a e obtemos um número aleatório r entre zero e esse somatório. A seguir, comparamos r com os vários somatórios parciais da adaptação. O primeiro indivíduo j que satisfizer a condição

$$\sum_{i=1}^{j} a_i \geq r$$

será escolhido para acasalamento. O número de casais escolhidos pode ser predeterminado ou não.

- Para operar a mutação, deveremos estipular um número p $(0 < p < 1)$, denominado probabilidade mutante. De posse de p, escolhemos um certo número de cromossomos e para cada um dos seus genes obtemos um r aleatório. Sempre que a condição $r < p$ for satisfeita, o referido gene será substituído por um outro gene qualquer.

7º - Ilustração do problema:

n°	Geração mãe	$p = e^{xy}$	$d = e - p$	$a = 1/(1+d)$
1	895419	1.45499	1.26329	0.44183
2	361786	1.32809	1.39019	0.41837
3	622771	1.61536	1.10292	4.47553
4	126305	1.03917	1.67911	0.37325
5	141010	1.00141	1.71687	0.36807
6	321292	1.09826	1.62002	0.38167

O indivíduo três será reproduzido para a geração 1, por ser o que possui a maior adaptação.

Agora, obtemos $\sum_{i=1}^{4} a_i = 2.45872$ e sorteamos r tal que $0 \leq r \leq 2.45872$.

Portanto, o cromossomo que satisfizer a condição tabelada a seguir será escolhido para acasalar. Temos então $r_1 = 1.6$ e $r_2 = 1.9$,

nº	Geração mãe	Somatório parcial das adaptações	Condição
1	895419	0.44183	0<= r <= 0.44183
2	361786	0.86020	0.44183<= r <=0.86020
3	622771	1.33573	0.86020<= r <=1.33573
4	126305	1.70898	1.33573<= r <=1.70898
5	141010	2.07705	1.70898<= r <=2.07705
6	321292	2.45872	2.07705<= r <=2.45872

Em nossa tabela os indivíduos que apresentam r_1 e r_2 em sua condição são: o 4 e o 5, estes então serão nosso casal doador, e o gene selecionado para a quebre é o gene de número 4.

Casal doador	Quebra do cromossomos	Acasalamento	Cromossomos transmitidos
126305	1263 05	1263 10	126310
141010	1410 10	1410 05	141005

Para aplicar a mutação, deve-se estipula uma probabilidade mutante que em nosso caso será pmut = 0.05. Escolhe-se então um dado indivíduo, utilizaremos o de número 6 por ser pouco adaptado, e para cada um de seus genes sorteia-se um r, tal que $0 \leq r \leq 1$, se $r < p\,\text{mut}$ o gene é alterado aleatoriamente

Gene	3	2	1	2	9	2
r associado	0.52	0.69	0.83	0.17	0.73	0.04

Logo, o gene na posição 6 poderá ser substituído por qualquer um dos genes existentes, inclusive por ele próprio, em nosso caso será substituído pelo gene 7 o cromossomo mutante será dado por 321297.

Obs.: Como *pmut* é pequeno, o mais provável era que o cromossomo não sofresse mutação alguma.

Então, finalizando, substitui-se a geração mãe pelos novos cromossomos adquiridos e pelos não utilizados obtendo, assim, nossa geração 1.

Geração mãe		Geração 1
895419		895419
361786		361786
622771	→	622771
126305		126310
141010		141005
321292		321297

O GA torna-se agora repetitivo. Aplicamos os mesmos critérios e as mesmas operações aos cromossomos da geração 1, visando obter a geração 2 e assim sucessivamente. Com um número predeterminado de iterações ou após a satisfação de alguma condição dada, deveremos obter uma taxa elevada de indivíduos bem adaptados.

É muito importante perceber que a necessidade de gerar números aleatórios, em qualquer GA, é tão grande que torna-se indispensável a utilização de alguma ferramenta eficiente, e isenta, na obtenção desses números. Caso contrário, o GA pode estar sujeito a algum tipo de comportamento vicioso, onde certos resultados são privilegiados em detrimento de outros.

4.3 - Funções de distribuição

DISTRIBUIÇÕES DE FREQUÊNCIA

Quando se resumem grandes massas de dados brutos, costuma-se freqüentemente distribuí-los em classes ou categorias e determinar o número de indivíduos pertencentes a uma mesma classe, denominada *freqüência de classe*. Um arranjo tabular dos dados por classes, juntamente com as freqüências correspondentes, é denominado *Distribuição de Freqüência* (Mandel, 1964; Spiegel, 1978).

No conceito básico de estatística temos Distribuição de Probabilidades associada a uma Variável Aleatória.

Mas qual é o conceito de Probabilidade e de Variáveis Aleatórias?

Probabilidade de um evento é considerada como a freqüência relativa, quando o n.º de observações é muito grande. A probabilidade propriamente dita é o limite da freqüência relativa quando o número de observações cresce indefinidamente.

Variáveis Aleatórias são aquela cujos valores são obtidos por um experimento aleatório e aos quais podemos associar probabilidades. Lembrando que a soma das probabilidades de todos os valores que a variável aleatória pode assumir é igual a 1(um).Estas podem ser:

discretas: assumem apenas um número limitado de valores em qualquer escala de medida e é obtida mediante alguma forma de contagem, e, também na interpretação de seus valores é dada exatamente por esse mesmo valor,

contínuas: assumem, teoricamente, qualquer valor numa escala de medida e resultam freqüentemente de uma medição, sendo dado em geral, em alguma unidade de medida, que se trata geralmente de um valor aproximado.

Por esse motivo, a Distribuição de Probabilidade encontra-se dividida em dois tipos:

Distribuição de probabilidade discreta

Se uma variável x pode assumir um conjunto discreto de valores X1, X2,...,Xk, cuja soma das probabilidades seja igual a Pi, diz-se que esta definida uma distribuição de probabilidade discreta de X.

A função P(x) assume valores p1, p2, ..., pk, respectivamente para X= X1, X2, ...,Xk, é denominada função de probabilidade ou de freqüência de X. Como x pode assumir certos valores com dadas probabilidades, ele é freqüentemente denominado variável aleatória discreta. É análogo uma distribuição de freqüências relativas, com estas substituídas pelas probabilidades. Por essa razão, podem-se imaginar as distribuições de probabilidade como uma forma teórica ou de limite ideal das distribuições de freqüências relativas, quando o número de observações feitas tornar-se muito grande. Por esta razão , pode-se imaginar que as distribuições de probabilidade referem-se a populações, ao passo que as de freqüência relativa referem-se a amostras delas extraídas.

Para variações discretas, o valor numérico x resultado de uma experiência pode assumir qualquer uma das seqüências dadas da possibilidade de saída:

x1, x2, x3, x4, ... a distribuição de probabilidade é então simplesmente um valor associado com cada possibilidade de saída xi.

Distribuição de probabilidade contínua

Neste caso, X pode assumir um conjunto continuo de valores e P(x) é denominada função de densidade e quando dada diz-se que foi definida a distribuição de probabilidade contínua para X, que fica denominada como variável aleatória continua.

Para variáveis continuas:

- não é possível escrever uma seqüência de possibilidades de saída, porque qualquer valor é uma possibilidade.

Então, considera-se um pequeno intervalo de x a x+dx e define a freqüência de distribuição em termos que a probabilidade do experimento x esteja neste intervalo

$$\text{Prob} \quad [x_i \leq x \leq x_i + dx] = f_i * .dx$$

Se for selecionada, ao acaso, uma amostra de tamanho N de uma população, é possível, então, mostrar que o valor <u>m</u> esperado para a media da amostra é igual à media μ da população, porém não se pode concluir que o valor esperado de qualquer quantidade calculada da amostra seja igual à correspondente da população.

O valor esperado para a variância de uma amostra, como foi definida, não é igual à da população e sim (n-1)/n vezes essa variância.

Não se pode esperar que um único valor reflita toda informação requerida pela freqüência de distribuição ou pela função distribuição.7

A variância de uma distribuição de freqüência pode ser descrita como o segundo momento sobre a média.

O desvio padrão de uma freqüência de distribuição é definido como a raiz quadrada da variância

DISTRIBUIÇÃO NORMAL OU GAUSSIANA

Normalmente em estatística trabalhamos com uma amostra representativa de determinado fenômeno. Na coleta de dados referentes a estes fenômenos, obtém-se séries de distribuição, cujos valores tendem a concentrar-se em torno de um determinado valor, e, quando plotados em um plano cartesiano, eles assumem a forma de uma curva chamada distribuição.

Existem casos em que a curva de distribuição se apresenta em forma de um sino, perfeitamente simétrica em relação à ordenada principal (x = Mo = Me), curva esta comumente chamada de " Os Erros Acidentais", ou ainda de "Gauss", isto porque os erros de dada observação, quando devidos ao acaso, e as medições são em número bastante elevados, eles dispõem-se simetricamente em torno de um valor.

Devemos lembrar contudo, que não são apenas os erros acidentais que assumem forma de sino, existindo outros fenômenos nos mais variados campos de atuação, cujas distribuições apresentam curvas do tipo chamado "Normal".

Resumindo, a Distribuição Normal é uma distribuição teórica, podendo ser aplicada em grande número de fenômenos.

A Distribuição Normal apresenta em seu histograma funções de distribuição simétricas, mas é destinada a adquirir a posição central na teoria estatística e prática.

Neste tipo de distribuição realmente compreende-se uma família de curvas diferenciadas entre si em sua média ou sua variância, ou em ambos e se duas distribuições tem a mesma média ou a mesma variância são completamente idênticas.

$$Y = \left(\frac{1}{\sigma - \sqrt{2\pi}} \right) * e^{-\frac{1}{2} * \left(\frac{(x-\mu)^2}{\sigma^2} \right)}$$

Isso mostra que a curva normal é simétrica sobre sua média μ e seu desvio padrão σ é a distância entre o local da média e o ponto de inflexão da curva. A probabilidade para uma variação continua é medida pelos elementos de área sob a curva representando a distribuição de função de probabilidade.

A distribuição de probabilidade do desvio padrão normal contém parâmetros desconhecidos quando observa-se que a média e o desvio foram eliminados, considerando como um caso particular, quando $\mu = 0$ e $\sigma = 1$. Todas as distribuições normais são relacionadas ao desvio padrão normal e pode ser reduzido a isso por uma simples transformação de escalas. Se uma tabela apresenta os valores de probabilidade do desvio padrão normal, estes valores podem ser utilizados por qualquer distribuição normal.

Exemplo:

Dada uma variável normal x encontrar a probabilidade sabendo μ , σ, e a localização.

- 1° - construir o intervalo reduzido, encontrando onde o desvio padrão provavelmente esta situado
- 2° - tabular a probabilidade de área da função de distribuição contínua: Pode-se selecionar um conjunto de valores xi, e tabular as probabilidades cumulativas, ou, selecionar uma seqüência de probabilidades de interesse particular e listar os valores variáveis correspondentes (este método é utilizado principalmente para responder perguntas do tipo: " encontre aquele valor cujo desvio padrão normal excederá apenas uma vez em cem")

Quando tem-se um valor de x conhecido no segundo tipo de tabulação, é chamado de quantiles, se este é expresso em por cento é chamado percentage, e um quantile de importância particular é a mediana de uma distribuição. Em qualquer distribuição normal a mediana coincide com a média , que para o desvio padrão normal é nula.

Em nosso trabalho estes conceitos têm grande importância, pois a função utilizada para estimar os parâmetros cosmológicos é composta por três gaussianas, das quais serão extraídas as informações sobre seus desvios padrão e suas médias, que irão auxiliar na solução do problema.

4.4 - Geração de Números Aleatórios

Na confecção dos Algoritmos Genéticos um fator de grande importância, principalmente para os casos em que deve haver mutação, é a geração de números aleatórios.

Denominamos números aleatórios, números escolhidos ao acaso (com exatamente a mesma probabilidade de escolha para todos) dentro de um espaço de parâmetro através de um gerador. O programa que gera os aleatórios recebe um valor arbitrário, e então retorna uma seqüência aleatória, sendo que uma mesma semente irá sempre retornar a mesma seqüência aleatória, a qual para ser realmente aleatória não deve apresentar periodicidade, para que durante o processo um determinado valor não seja privilegiado diante dos demais, o que faria com que o a seqüência deixasse de ser aleatória. (Press et al., 1992).

No caso dos Algoritmos Genéticos os geradores de números aleatórios apresentam grande importância, por exemplo, na realização das mutações nas quais selecionamos um indivíduo e alteramos um ou mais genes em seu cromossomo. O cromossomo mutante será transmitido para a geração seguinte. Para que isso ocorra é estipulada uma condição, relacionada à probabilidade mutante, e são gerados valores aleatórios que serão relacionados a cada gene do cromossomo. Se algum desses valores não satisfaz a condição aquele gene ao qual ele corresponde é alterado também de forma aleatória e este poderá ser substituído por qualquer um dos genes existentes, inclusive por ele próprio.

4.5 - Métodos de Minimização

Métodos de minimização, que são utilizados seguindo o critério de otimização Dada uma função F que depende de uma ou mais variáveis independentes para solucioná-la deve-se encontrar o valor das variáveis onde F alcança o seu máximo ou mínimo, utilizando um programa que restringirá seu cálculo nas vizinhanças de F e, as vezes nas de suas derivadas, respeitando a todas a suas variáveis, que possam ser necessárias (Press et al., 1992).

Infelizmente, não existe um algoritmo perfeito de otimização. Por isso, deve-se tentar mais de um método para comparação. Deve-se, então escolher entre métodos que precisam apenas de cálculo da função a ser minimizada e métodos que também requerem cálculos das derivadas das funções.

Algoritmos que utilizam a derivação são muitas vezes mais poderosos do que aqueles que usam apenas uma função, mas muitas vezes não compensa realizar o cálculo de derivadas adicionais, devido a este requerer muito tempo.

No tratamento de problemas de otimização, a execução do código deve terminar assim que a solução ótima for encontrada. Já no caso de funções multimodais, um ponto ótimo pode ser o suficiente, mas pode haver situações nas quais todos ou o maior número possível de pontos ótimos sejam desejados. E, na maioria dos casos, não se pode afirmar com certeza se um dado ponto ótimo corresponde a uma solução global. Como conseqüência, normalmente utiliza-se o critério do número máximo de gerações ou um tempo limite de processamento, para a decisão da "convergência" da solução. Um outro critério plausível é a idéia de estagnação, ou seja, quando não se observa melhoria da população depois de várias gerações consecutivas.

Os critérios devem ser considerados de acordo com o problema, a convergência para uma solução ótima deve estar sujeita a um agente de convergência que indicará quando se alcançou o melhor resultado, e que não há como melhorar este resultado de acordo com os critérios acima.

Os Algoritmos Genéticos são relativamente fáceis de codificar e modificar. Em condições práticas, podem ser complementados com técnicas de otimização convencionais (Goldberg, 1989; Mitchell, 1998).

Enquanto tais técnicas podem acelerar a convergência facilmente, muitos requerem precisão na modelagem de parâmetros e executam bem somente uma subdivisão de classe especifica do problema.

Vários são os mecanismos que utilizam os Algoritmos Genéticos (aplicados para uma representação mais flexível e uma maior qualidade), adaptando-os a sua necessidade de procura. Eles utilizam representações dinâmicas dirigidas por uma parte de novos operadores, os operadores de reprodução, associados à informação extraída da performance atual e então explicitamente aplicados aos mecanismos primários. O princípio que rege essa aplicação dos Algoritmos Genéticos junto a outros mecanismos de procura baseia-se na teoria de Lamark, que afirma que características de adaptação podem ser adquiridas por herança.

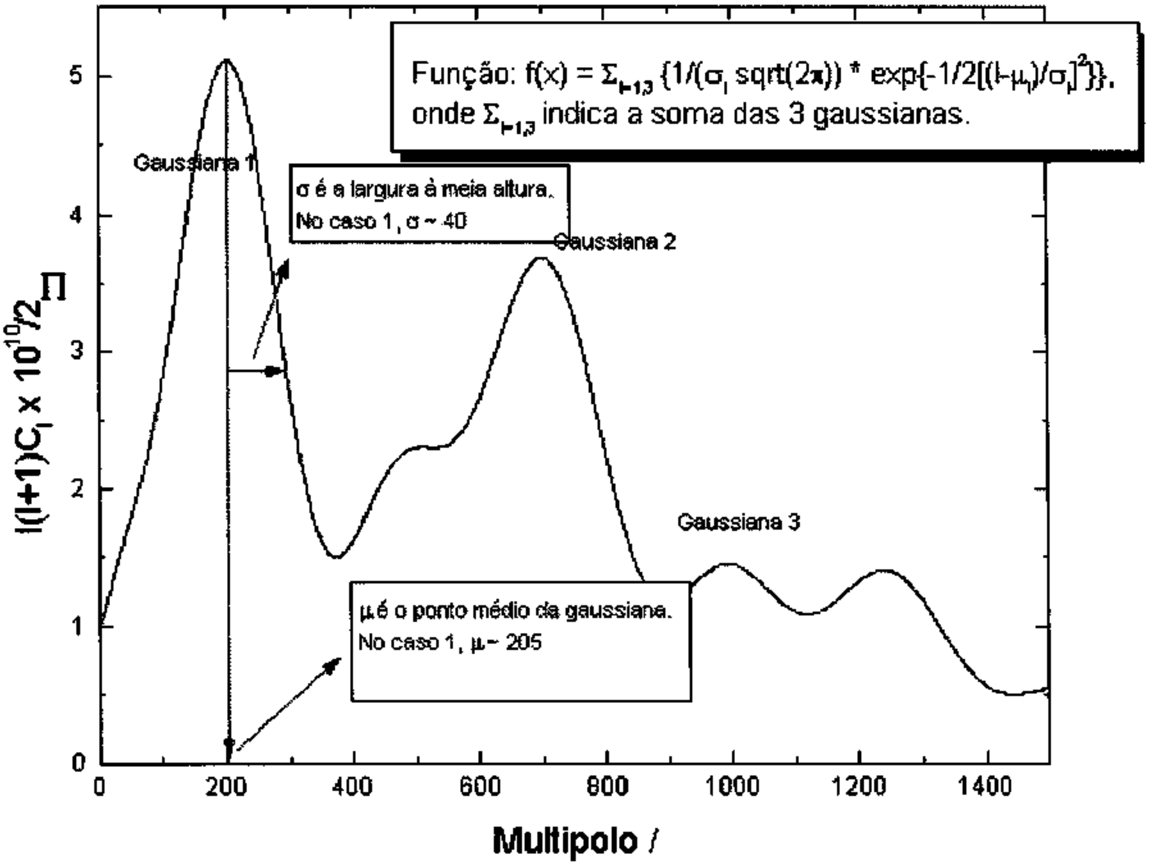
LS-I, ARGOT, Genes, Messy Genetic Algorithm, são alguns exemplos de implementações que podem ser utilizadas.

Na prática, os Algoritmos Genéticos são utilizados, muitas vezes, para complementar técnicas convencionais de otimização em outros métodos de procura devido a sua maior eficiência e facilidade de trabalho. Conceitualmente os Algoritmos Genéticos encarnam os mesmos mecanismos que conduziram à nossa própria existência e oferecem um modo extremamente robusto e diferente de fazer otimização, sem usar as técnicas convencionais e seus princípios básicos.

Utilizou-se o código Pikaia para implementar uma função cujas características se assemelham ao espectro de potência da RCFM. Devido à sua estrutura, pode-se ajustar o espectro à uma função do tipo:

$$F(x) = ax + G_1(x) + G_2(x)$$

em que G1 e G2 são as funções gaussianas. Com esse ajuste, pretendemos fazer estimativas de parâmetros cosmológicos primários mencionados na seção 2.2. Como mostra a figura a seguir:



4.6 – CMBFAST

CMBFAST gera espectros de potência da RCFM, a partir da combinação de diversos ingredientes, incluindo os parâmetros cosmológicos mencionados na seção 2.2, que são extremamente úteis na comparação com resultados experimentais. Parâmetros cosmológicos são importantes na determinação do espectro de potência da RCFM. Uma forma de determiná-los é considerar uma região do espaço de parâmetros e compará-la a resultados já validados, procurando delimitar a região de validade dos mesmos.

O CMBFAST é um código computacional utilizado para calcular o espectro de potência da RCFM, a polarização e as funções de transferência. Ele é baseado na integração sobre as fontes do cone de luz de fótons, no qual a anisotropia da temperatura é escrita como uma integral temporal de um termo geométrico (dado por auto-funções radiais que não dependem do modelo cosmológico) e um termo de fonte (que pode ser expresso em termos de fótons, barions ou de perturbações métricas, que podem ser calculadas utilizando um número pequeno de equações diferenciais), separando assim, os efeitos dinâmicos e geométricos da anisotropia da RCFM. Este método permite um tempo de processamento significativamente mais curto do que métodos precedentes.

Embora seja um código mais usado para computar a anisotropia e a polarização da RCFM, ele pode também ser utilizado para calcular e normalizar funções de transferência para espécies diferentes (fótons, barions, etc). A função de transferência da matéria é utilizada no cálculo do espectro de potência da matéria no Universo. Este espectro de potência da matéria é a complementação ao espectro de potência bidimensional da RCFM. Na prática, a distribuição da matéria no Universo é encontrada traçando a posição e o redshift das galáxias e computando a potência desse mapa.

No CMBFAST são estudados os picos acústicos que foram congelados no tempo na Superfície de Último Espalhamento (SUE). A taxa de expansão do Universo tem o efeito de alongar estas características em todas as escalas. Uma característica particular é dada pelo tamanho dentro do "horizonte" que depende do raio de expansão do Universo e da velocidade do som local, porque é a velocidade limite com que uma oscilação da pressão pode percorrer.

A "densidade crítica" do Universo é a densidade que precisamente balancearia a energia da expansão e a energia gravitacional de toda a matéria encontrada. Atualmente as densidades de várias espécies são expressas em termos da densidade crítica enquanto analisam-se as abundâncias relativas de cada espécie: barions, matéria escura, neutrinos, fótons, e "lambda" , esperam-se que todas as frações adicionadas cheguem a uma unidade no caso do Universo plano.

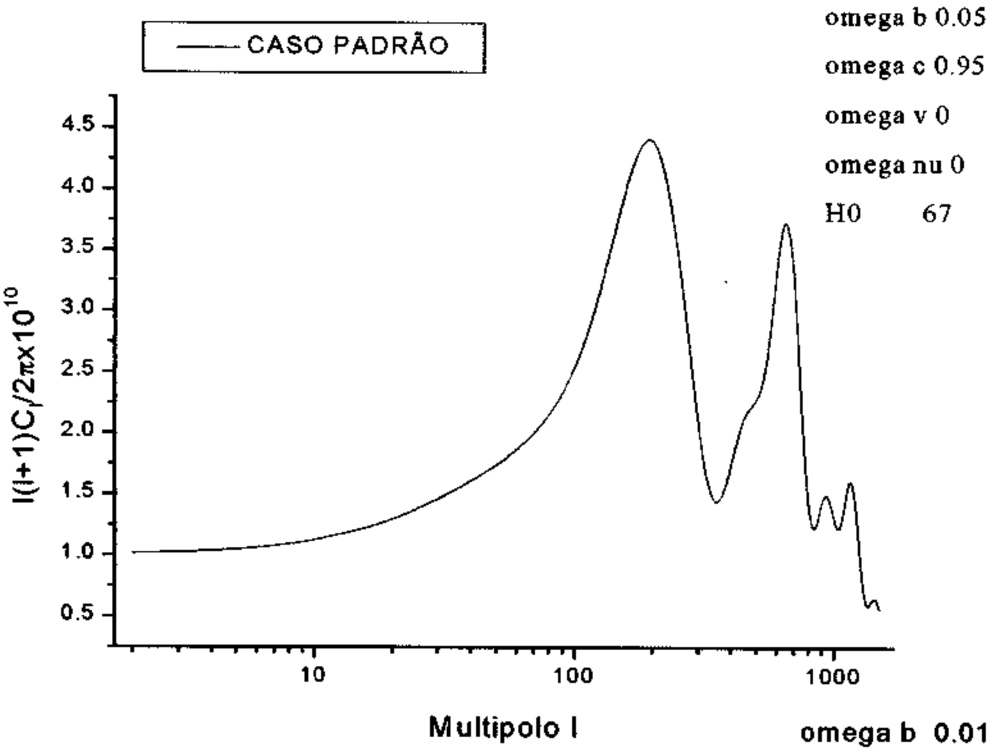
A solução numérica da integral e da equação de transporte de fótons difere na geometria de Universos planos e não-planos. No caso da geometria plana o termo geométrico é dado pela função esférica de Bessel, e na não-plana utilizam-se as funções ultra-esféricas de Bessel, dependentes da distância radial e do vetor de onda k.

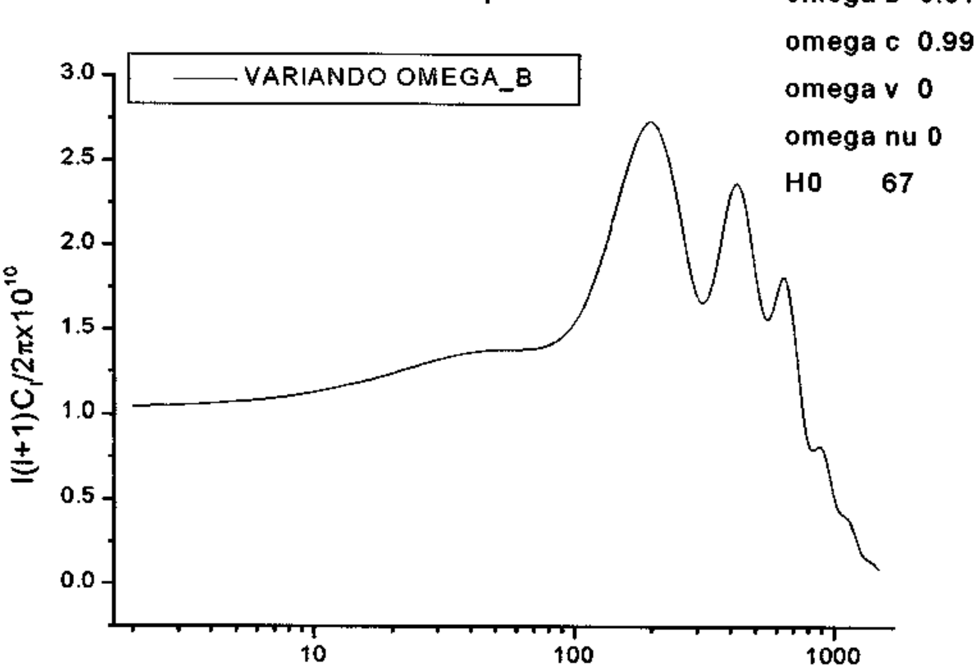
Embora em nosso trabalho nos detenhamos aos Universos Planos, é preciso salientar que o CMBFAST pode calcular o espectro para Universos planos, abertos e fechados.

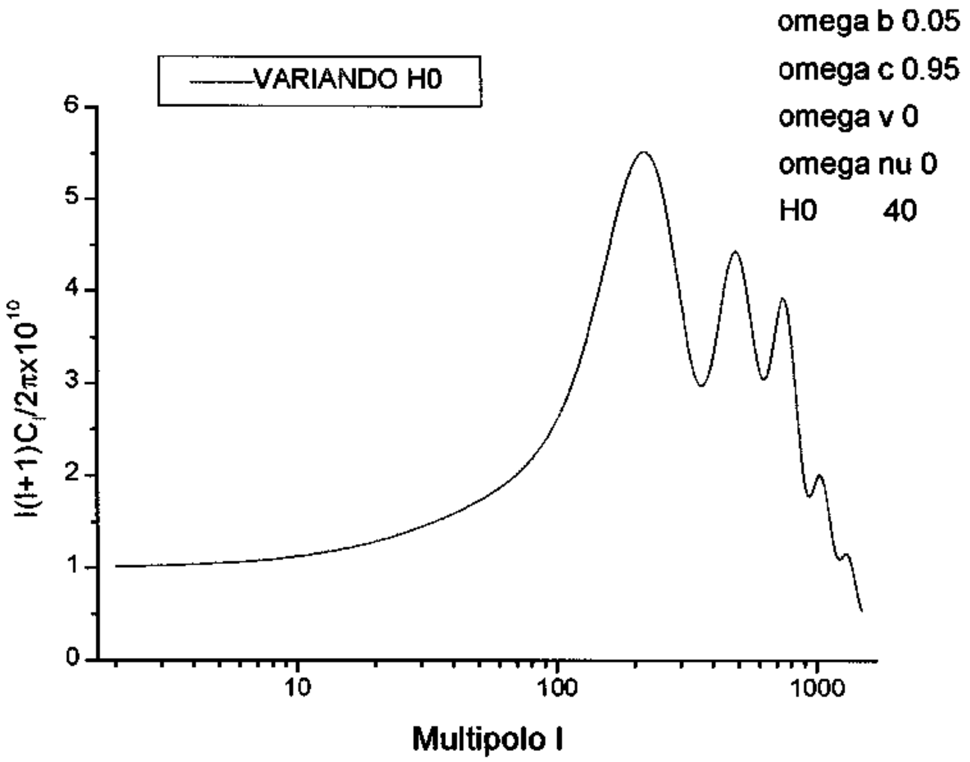
No caso dos Universos Fechados a principal preocupação deve ser a amostragem l, pois quando a distância angular do diâmetro SUE é reduzida, característica deste tipo de universo, o espectro de potência tem sua amostragem l comprimida, significando que l não é bom o bastante e deve ser aumentado.

Nos casos dos Universos plano e aberto a diferença está somente na maneira em que o termo geométrico é calculado. Nos Universos Planos as funções esféricas de Bessel são interpoladas em tabelas pré-computadas e nos Abertos as funções ultra-esféricas são obtidas integrando uma equação diferencial, e suas perturbações escalares podem somente ter condições iniciais de isentropia.

A seguir temos exemplos de gráficos calculados a partir da utilização do CMBFAST:







5. RESULTADOS

Após desenvolvido todo o estudo, acima citado, foi feita a implementação da função definida na seção 4 em Pikaia. Mas, ainda não foi possível a obtenção dos espectros de potência que devem ser gerados em Pikaia para comparação aos gerados pelo CMBFAST, pois o programa vem apresentando problemas de convergência.

A solução deste problema será atribuída à segunda fase da realização da Iniciação Científica, a partir de agosto.

6. DISCUSSÃO

O estudo da linguagem de programação Fortran, permitiu as adaptações necessárias ao código Transito, e ajudou a implementação da função descrita na seção 4 em Pikaia. A implementação de Pikaia permitiu o melhor domínio da linguagem, além de um teste importante da utilização dos Algoritmos Genéticos numa área, que tem se mostrado desprovida de alternativas eficientes para a análise de dados dos experimentos mais recentes.

O CMBFAST, utilizado em nosso trabalho, é um código em que para toda a combinação de parâmetros cosmológicos retorna-se um espectro de potência em função do número de ordem. Os dados por ele gerados servirão de auxilio na segunda parte de nosso trabalho na comparação com os dados experimentais, que serão gerados no Pikaia após a solução do problema de convergência encontrado recentemente.

O que nos possibilitará a comprovação da eficiência de nosso Algoritmo Genético no desenvolvimento do estudo do espectro de potência da RCFM para a obtenção dos parâmetros cosmológicos primários.

É importante, como em uma linguagem de programação pode ser desenvolvido um código, baseado em populações numéricas e em conceitos biológicos, como evolução e mutação, que possibilita o melhor conhecimento dos parâmetros cosmológicos primários que permitem entender a evolução do Universo jovem. Parâmetros que possibilitarão saber a idade, a composição química, o tamanho e a e a geometria do Universo nos dias de hoje.

7. CONCLUSÃO

Realizado todo o estudo sobre Algoritmos Genéticos, e feita a implementação da função de ajuste, composta por três gaussianas, embora não existam resultados concretos para concluir a confiabilidade do código, pode-se dizer que ele mostra eficiência e rapidez na obtenção dos dados. E que com a solução do problema de convergência por ele apresentado poderá ser feita a comparação a diversas simulações de um espectro de potência, correspondente a diferentes modelos cosmológicos, gerados pelo código CMBFAST.

Visando, sempre, é claro mostrar a eficiência da "Procura" com a utilização dos Algoritmos Genéticos no processo de determinação de parâmetros cosmológicos a partir do espectro de potência de flutuações de temperatura da RCFM, o que nos possibilita obter uma maior compreensão sobre a evolução do Universo jovem.

8. REFERÊNCIAS BIBLIOGRÁFICAS

8.1. Bevington, P. R., *Data Reduction and error analysis for the physical sciences*, McGraw- Hill, New YorK, NY, 1969

8.2. Boas, M. L., *Mathematical Methods in the Physical Sciences*, 2° edition, John Wiley & Sons, Chicago, Il, 1983

8.3. Charboneau, P., *Genetic Algorithms in Astronomy and Astrophysics.* Astrophysical Journal Suppl. Series, 101: 309-334, 1995.

8.4. Davidor, Y. *Genetic Algorithms and Robotics : A Heuristic Strategy for Optimization.* World Scientific, Teaneck, NJ, 1991.

8.5. Ellis, T.M.R.. *Fortran 77 Programming : With an Introduction to the Fortran 90 Standard (International Computer Science Series)*

8.6. Goldberg, David E. Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

8.7. Mandel, J., *The Statistical Analysis of Experimental Data,* Dover Publications, Inc., New York, 1964.

8.8. Milone, A. C.; et. al., *Introdução à Astronomia e Astrofisica,* Inpe- 7177-PUD/38, São José dos Campos, SP, 1999.

8.9. Mitchell, M., An Introduction to Genetic Algorithms (Complex Adaptive Systems Series), MIT Press, Boston, MA, 1998

8.10. Oliveira, K. S.; Saraiva, M. F. O., *Astronomia e Astrofisica*, Ed. da UFRGS,2000.

8.11. Press, W. H.; et. al., Numerical Recipes in Fortran (The art of scientific computing), 2° edition, Cambrige University Press, New York, NY, 1992.

8.12. Seljak, U.; Zaldarriaga, M. "A Line of Sight Approach to Cosmic Microwave Background Anisotropies". Preprint astro-ph/9603033, disponível no banco de preprints eletrônicos de Los Alamos (http:// br.arxiv.org)

8.13. Sodré Jr., L., *Introdução à Cosmologia,* Notas de aula do curso de cosmologia, IAG-USP, 1997.

8.14. Spiegel, *Probabilidade e Estatística,* Coleção Schaum, McGraw- Hill, Brasil, BRA, 1978

8.15. Wahde, M., *Determination of orbital parameters of interacting galaxies using a genetic algorithm (Description of the method and application to artificial data),* Preprint astro-ph/ 9710278, disponível no banco de preprints eletrônicos de Los Alamos (http:// br.arxiv.org)

8.16. Wright, E. L., *Scanning and Mapping Strategies for CMB Experiments,* Preprint astro-ph/ 9612006, disponível no banco de preprints eletrônicos de Los Alamos (http:// br.arxiv.org)