

**Universidade de São Paulo
IME - Instituto de Matemática e Estatística**

**MAC-499 – Trabalho de Formatura
Supervisionado**

Monografia

Projeto de Iniciação Científica:

Interface de Integração Excel / Gnuplot

para

Análise de Dados

Aluno:	Marcos Roberto Yukio Koga marcos@linux.ime.usp.br	N. USP: 3312021
Colaborador:	Luciano Vieira Araújo luciano@ime.usp.br	
Orientador:	João Eduardo Ferreira jef@ime.usp.br	

Sumário

INTRODUÇÃO.....	3
 A. PARTE TÉCNICA	
A.1. Introdução.....	4
A.2. Motivação.....	4
A.3. O Projeto.....	7
A.4. Tópicos Estudados.....	12
A.5. Atividades Realizadas.....	12
A.6. Resultados Obtidos.....	13
A.7. Conclusão.....	13
A.8. Referências Bibliográficas.....	14
 B. PARTE SUBJETIVA	
B.1. Desafios e Frustrações Encontrados.....	15
B.2. Disciplinas do BCC mais Relevantes para o Trabalho.....	16
B.3. Interação com o Orientador e Outros Profissionais.....	16
B.4. Aplicação Prática de Conceitos Estudados.....	16
B.5. Passos Futuros para Aprimoramento dos Conhecimentos.....	16

INTRODUÇÃO

Esta monografia relata a experiência que obtive realizando um projeto de iniciação científica sob orientação do professor João Eduardo Ferreira, no período que vai de abril a dezembro/2003.

Ela é composta de duas partes. A parte A descreve os aspectos técnicos do projeto, os objetivos, os tópicos estudados e as atividades realizadas. A parte B relaciona a experiência pessoal, as dificuldades encontradas e a importância da formação obtida no curso do BCC.

A. PARTE TÉCNICA

A.1. Introdução

Freqüentemente, quando se deseja entender algum fenômeno, coleta-se uma amostra relativamente grande de dados sobre o fenômeno e armazenam-se os dados num banco de dados para, posteriormente, realizar análise das informações.

Um dos métodos mais eficazes para analisar uma grande quantidade de dados é observar visualmente a variação dos dados por meio de gráficos. Para aplicação desse método, estão disponíveis muitas ferramentas capazes de gerar gráficos rapidamente a partir de uma base de dados.

O objetivo principal deste trabalho é desenvolver mais uma ferramenta para construção de gráficos, que possa ser utilizada por pesquisadores que necessitem analisar um grande volume de dados. Na realidade, não se trata de uma nova ferramenta independente, mas sim, da integração de duas ferramentas já existentes e muito utilizadas: o Microsoft Excel e o Gnuplot.

A.2. Motivação

O Laboratório de Bioinformática do IME tem desenvolvido diversos projetos para o Ministério da Saúde. Um deles objetiva fazer uma análise determinística de dados epidemiológicos e resistência genotípica do HIV-1.

O vírus HIV costuma sofrer muitas mutações ao se duplicar numa célula de um indivíduo contaminado. E, estas mutações podem ocorrer em várias partes diferentes de seu código genético.

Um levantamento realizado pelo Ministério da Saúde coletou informações sobre mais de 100.000 pacientes portadores do HIV. Uma análise bastante útil seria verificar se entre pacientes submetidos a um mesmo tratamento existe algum padrão para as mutações do HIV desenvolvidas nesses pacientes.

Um meio para tentar fazer essa análise, seria criar um gráfico em três dimensões, relacionando numa dimensão a posição do código genético do HIV onde houve a mutação, numa segunda dimensão, a droga utilizada no tratamento do paciente e, na terceira dimensão, para cada um dos pares de valores associados das duas primeiras dimensões, associar uma medida de agregação, como a quantidade de pacientes por mutação e por droga.

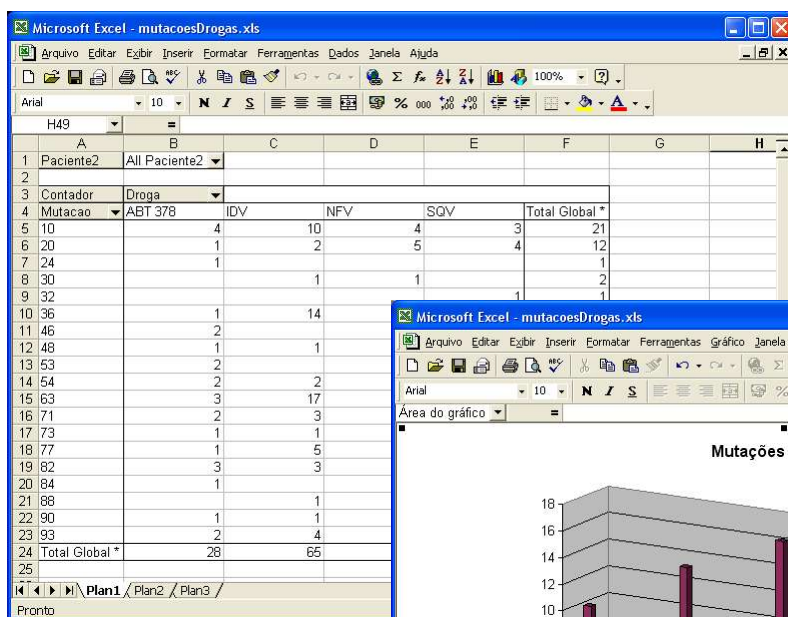
Pela solução desenvolvida atualmente pelo Laboratório de Bioinformática, os dados de pacientes coletados são armazenados no Sistema Gerenciador de Banco de Dados (SGBD) SQL Server, da Microsoft, escolhido pela melhor relação custo/benefício dentre outros sistemas existentes no mercado.

O Microsoft SQL Server permite criar estruturas multidimensionais para representar o inter-relacionamento entre dados em diferentes dimensões, como é o caso da estudo da correlação entre as três dimensões: drogas, mutações e número de pacientes.

Uma ferramenta muito conhecida no mercado, capaz de acessar os dados de estruturas multidimensionais do Microsoft SQL Server é o Microsoft Excel. Além de poder se comunicar facilmente com o SQL Server, o Excel também tem a vantagem de oferecer uma interface gráfica intuitiva e de fácil utilização para usuários com pouco conhecimento técnico em computação, como é o caso dos funcionários do Ministério da Saúde.

Por esses motivos e, pelo fato do Excel já ser adotado pelo Ministério da Saúde, ele constitui a principal ferramenta para a “visualização” dos dados.

A Figura 1 mostra um exemplo de uma planilha do Excel com informações de pacientes provenientes de uma base de dados em SQL Server e a Figura 2 mostra um gráfico construído a partir dessa planilha.

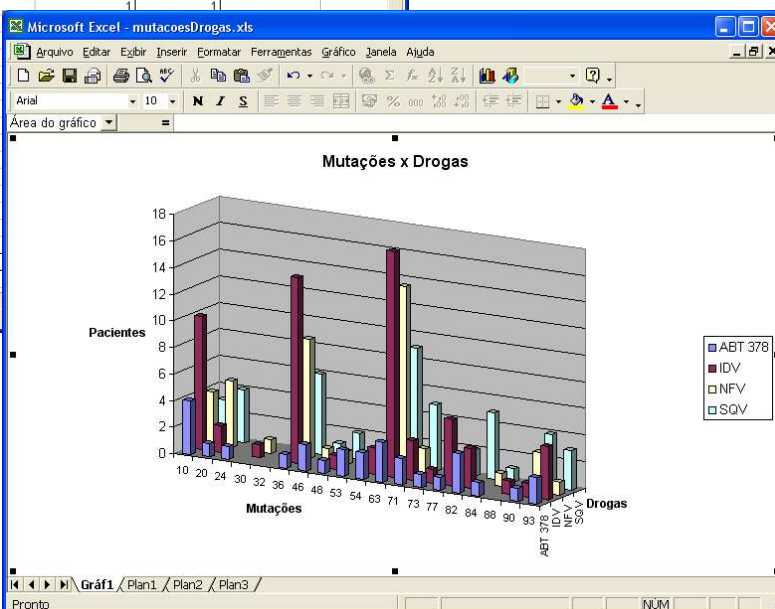


Microsoft Excel - mutacoesDrogas.xls

Paciente2	Druga	IDV	NFV	SQV	Total Global *
10	ABT 378	4	10	4	21
20		1	2	5	12
24		1			1
30			1	1	2
32				1	1
36		1	14		
46		2			
48		1	1		
53		2			
54		2	2		
63		3	17		
71		2	3		
73		1	1		
77		1	5		
82		3	3		
84		1			
88			1		
90		1	1		
93		2	4		
Total Global *		28	65		

Figura 1: Planilha com dados provenientes do SQL Server

Figura 2: Gráfico gerado pelo Excel a partir dos dados da planilha



Apesar de ser possível gerar muitos tipos diferentes de gráficos pelo Excel, o professor e orientador deste projeto de iniciação científica, João Eduardo Ferreira, constatou que o conjunto de tipos de gráfico oferecido ainda não supria todas as possibilidades de visualização dos dados; havia necessidade de uma variedade maior de gráficos 3D.

Para tentar atender suprir essa demanda foram estudadas várias outras ferramentas de construção de gráficos. Entretanto, nenhuma se mostrava plenamente satisfatória.

Dentre as ferramentas testadas, uma que obteve grande destaque foi o Gnuplot. O Gnuplot é um software gratuito que se destina à construção de gráficos variados, especialmente, os de aplicação científica. Como o Gnuplot foi desenvolvido para interpretar comandos entrados por um terminal, o seu uso exige a compreensão da sintaxe dos seus comandos, uma tarefa que não é trivial para qualquer usuário. Por outro lado, o Gnuplot apresenta alguns tipos de gráficos bem variados em relação ao Excel. As Figuras 3 e 4 mostram gráficos que podem ser gerados pelo Gnuplot.

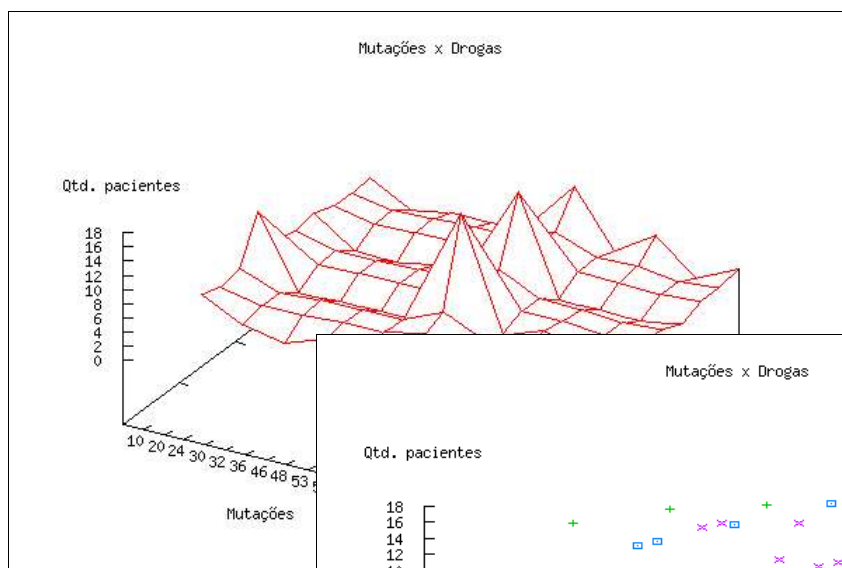
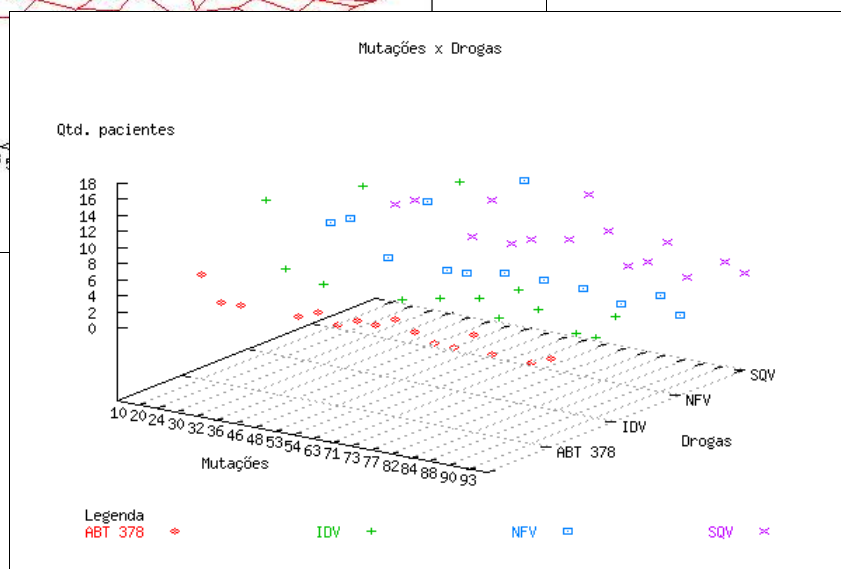


Figura 3: Gráfico de superfície

Figura 4: Gráfico de pontos (X,Y,Z)



Tendo, de um lado, um software com interface bem intuitiva e, de outro, um software não muito intuitivo, mas com recursos desejáveis no primeiro, torna-se natural tentar juntar o melhor dos dois lados. Nesse contexto, foi levantada a hipótese de uma integração do Excel com o Gnuplot e a partir de então, iniciaram-se os estudos para tentar concretizar essa integração.

A.3. O Projeto

Para integrar o Excel e o Gnuplot, a solução proposta foi a de utilizar a interface de programação em Visual Basic for Applications (VBA) disponibilizada pelo Excel.

A linguagem VBA é uma versão mais enxuta da linguagem Visual Basic (VB), com bem menos recursos. Ela é utilizada para automatização de tarefas por meio da criação de macros, nos aplicativos do Microsoft Office, como o Excel e o Word. Programando macros, tarefas que seriam repetitivas, podem ser realizadas com um simples clique num botão ou num comando de um menu.

Usando VBA foi possível desenvolver um protótipo para uma interface gráfica de integração do Excel com o Gnuplot. Através da interface gráfica, um usuário sem familiaridade com o Gnuplot pode gerar facilmente um gráfico 3D pelo Gnuplot sem precisar aprender comando algum do Gnuplot, bastando apenas clicar em alguns botões e preencher algumas caixas de texto para obter um gráfico da forma desejada.

Todo o código VBA da interface é incluído num arquivo comum do Excel, que depois é gravado no formato especial de suplemento (ou “add-in”). O Excel deve então ser configurado para carregar este suplemento ao inicializar. Assim, em toda execução do Excel, o suplemento é carregado. Após a carga do suplemento, um novo menu, “Gnuplot”, passa a ser exibido na barra de menus do Excel, dando acesso ao comando “Gerar gráfico”, que exibe a interface gráfica para geração de gráficos 3D por meio do Gnuplot. A figura 5 mostra em detalhe o novo menu.

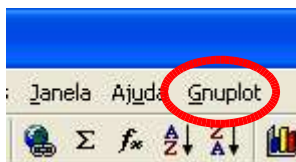


Figura 5: Menu Gnuplot criado ao inicializar o Excel com o suplemento desenvolvido

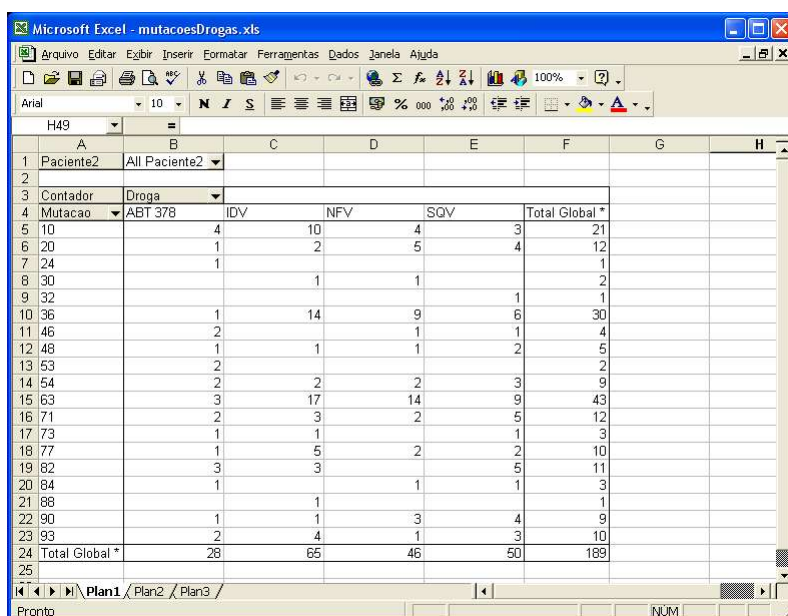
Resumidamente, o processo de geração de um gráfico 3D por meio do Gnuplot, a partir da interface em VBA no Excel consiste em:

1. Criação de uma tabela com os dados do gráfico

Criar uma tabela com os dados a serem utilizados para construir o gráfico. A tabela deve apresentar em sua primeira linha e em sua primeira coluna os rótulos de escala das dimensões da base do gráfico e nas coordenadas resultantes do cruzamento entre linhas e colunas, os valores para o eixo dos valores (perpendicular à base). O exemplo a seguir mostra uma tabela de 4 linhas por 3 colunas.

	d1	d2	d3
s1	2.0	5	6.89
s2		4.2	17.34
s3	11.23		
s4		2.34	5.49

A tabela anterior representa apenas um exemplo muito simples, com poucos dados, objetivando mostrar a estrutura da tabela. Num caso mais prático, certamente, os dados serão provenientes de um banco de dados como o SQL Server, o Oracle ou outros suportados pelo Excel, gerando a chamada tabela dinâmica, que permite ao usuário combinar dinamicamente os campos de tabelas do banco de dados para criar diferentes tabelas. A Figura 6 mostra um exemplo de tabela dinâmica criada com dados provenientes de uma estrutura multidimensional armazenada na estrutura de um cubo. Observa-se que na primeira coluna da tabela encontram-se as posições onde ocorrem mutação, na primeira linha da tabela encontram-se os tipos de drogas e nos cruzamentos entre linhas e colunas, o número de pacientes para cada combinação (as células vazias significam que não há ocorrência de uma determinada combinação de posição da mutação e tipo de droga).



Paciente2	All Paciente2					
Contador	Droga					
Mutacao	ABT 378	IDV	NFV	SGV	Total Global *	
10		4	10	4	3	21
20		1	2	5	4	12
24		1				1
30			1	1		2
32					1	1
36		1	14	9	6	30
46		2		1	1	4
48		1	1	1	2	5
53		2				2
54		2	2	2	3	9
63		3	17	14	9	43
71		2	3	2	5	12
73		1	1		1	3
77		1	5	2	2	10
82		3	3		5	11
84		1		1	1	3
88			1			1
90		1	1	3	4	9
93		2	4	1	3	10
Total Global *		28	65	46	50	189

Figura 6: Tabela dinâmica gerada com dados provenientes de uma base de dados externa ao Excel.

2. Iniciar a interface gráfica em VBA

Para iniciar a interface gráfica em VBA, basta executar o comando “Gerar gráfico” do menu Gnuplot. Será exibida a janela “Dados de origem do gráfico” (Figura 7).

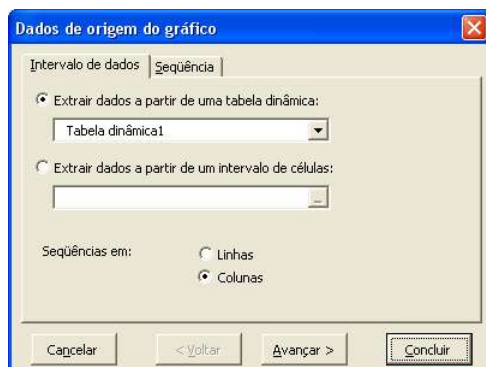


Figura 7: Definindo dados de origem do gráfico

3. Definir os dados de origem do gráfico

Na janela “Dados de origem do gráfico”, definir qual o intervalo de células da planilha contém os dados a serem utilizados no gráfico ou qual tabela dinâmica servirá como origem de dados para o gráfico.

4. Definir a formatação do gráfico

Na janela “Opções de gráfico”, definir a formatação do gráfico, ou seja, definir atributos como o título do gráfico, o título dos eixos, a exibição de legenda ou a rotação do gráfico (Figura 8).

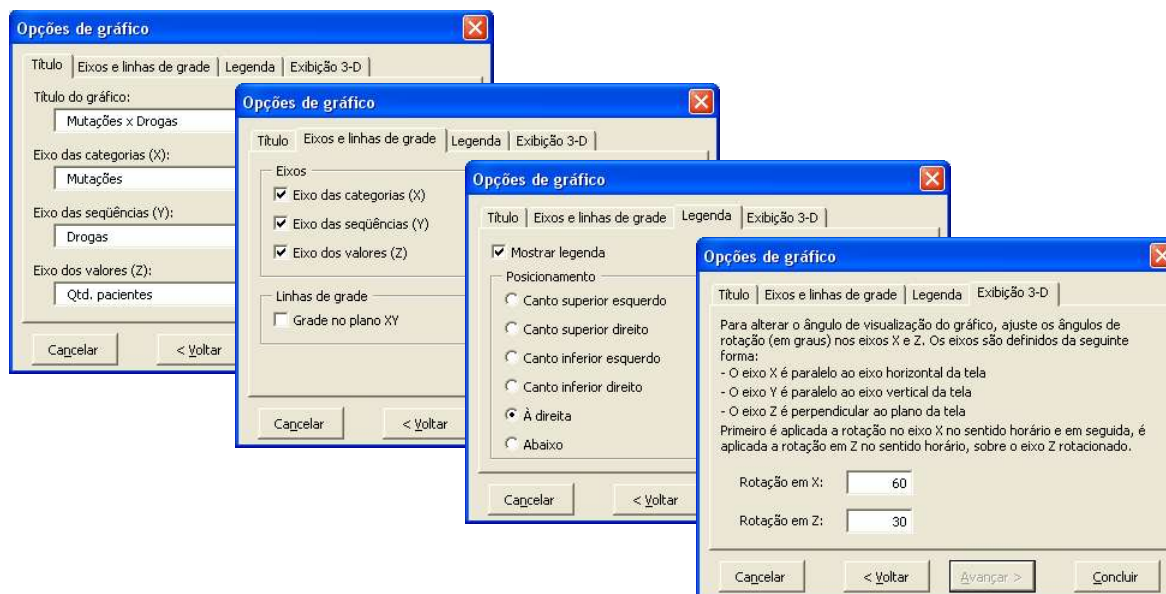


Figura 8: Definindo formatação do gráfico

5. Gravação de arquivos de entrada para o Gnuplot

O suplemento em VBA grava um arquivo com os comandos a serem executados pelo Gnuplot para obter um gráfico com as opções definidas e os dados definidos anteriormente. Os dados são gravados em um ou mais arquivos separados.

6. Execução do Gnuplot

O suplemento em VBA executa o Gnuplot usando como entrada o arquivo de comandos gravado anteriormente. O Gnuplot interpreta cada um dos comandos, lê os arquivos de dados e gera como saída uma imagem representando o gráfico. No momento, a imagem é gravada somente no formato GIF.

7. Exibição do gráfico gerado

O gráfico gerado pelo Gnuplot é carregado na janela “Visualização do gráfico” e exibido ao usuário (Figura 9).

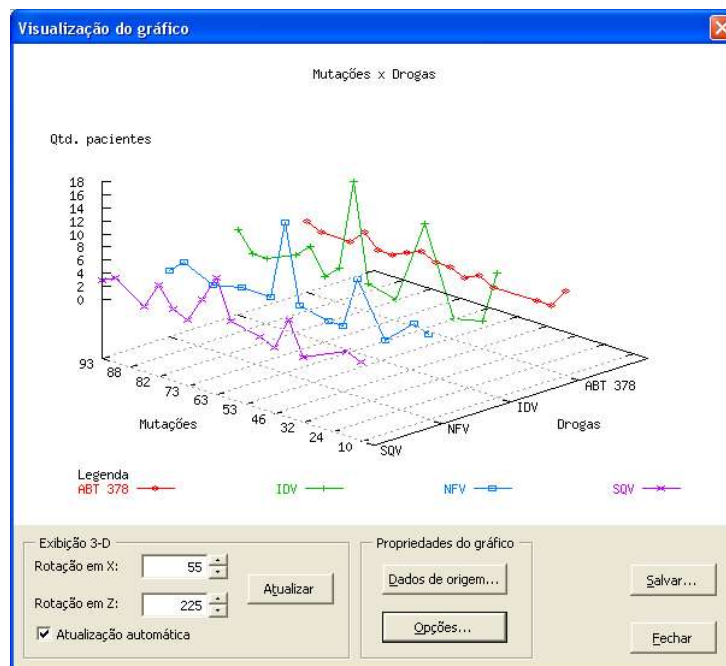


Figura 9: Visualização do gráfico

Nesta janela, o usuário pode ajustar o gráfico, rotacionando-o pelos eixos X e Z ou alterando opções de formatação. A atualização é feita gerando o gráfico novamente no Gnuplot e carregando-o em seguida.

O usuário também tem a opção de salvar o gráfico num arquivo para uso posterior, como por exemplo, exibição numa página web. No momento, a imagem é gravada somente no formato GIF.

A Figura 10 a seguir resume todo o processo de geração de gráfico usando o protótipo desenvolvido:

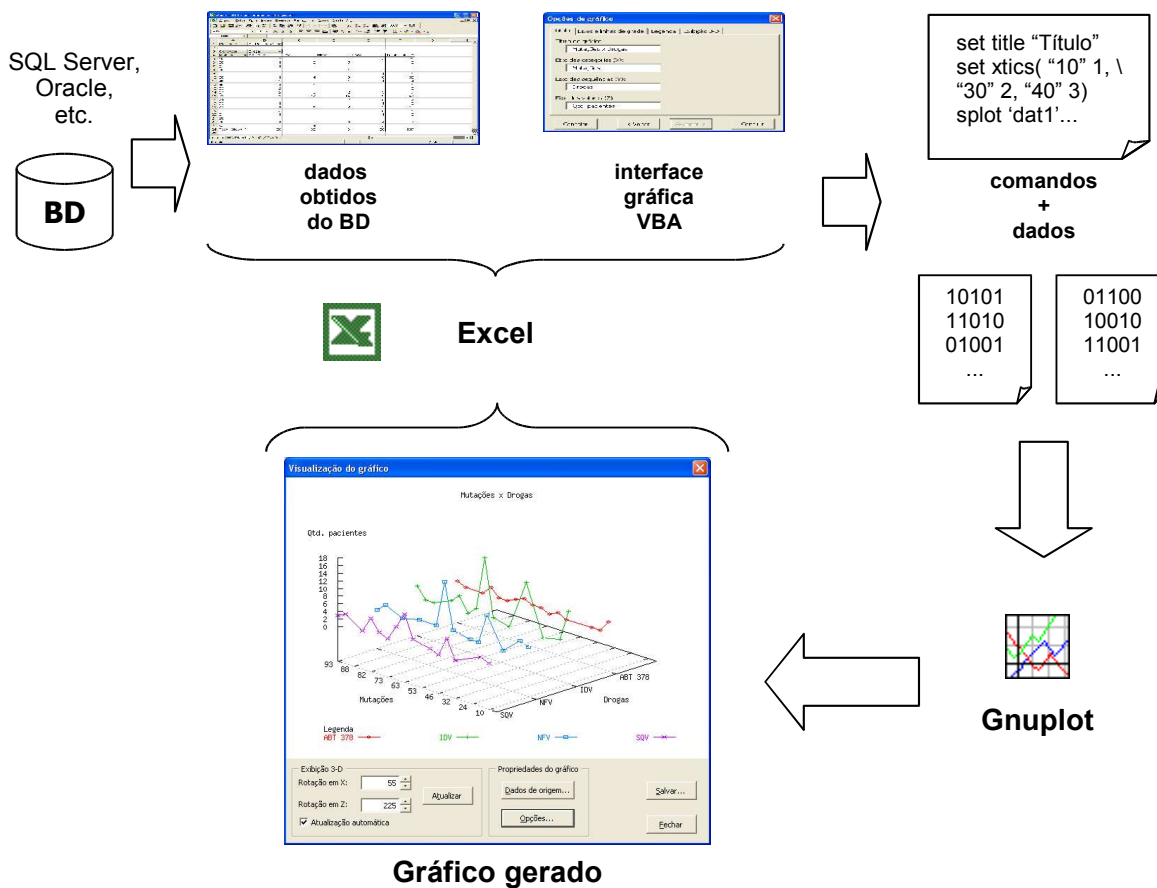


Figura 10: Processo de geração de gráfico integrando Excel e Gnuplot e Excel

A.4. Tópicos Estudados

Para o desenvolvimento deste projeto, foi necessário estudar o funcionamento do Gnuplot, principalmente, aprender a sintaxe dos comandos de:

- Definição de terminal gráfico e arquivo de saída;
- Definição de rótulos de marcas de escala dos eixos;
- Formatação de legenda, títulos e estilos de linha;
- Plotagem de gráfico tridimensional;
- Rotação de gráfico.

Além disso, antes de optar pelo formato GIF para geração de gráficos no Gnuplot, vários outros formatos de imagem (definidos pela escolha de diferentes terminais gráficos) tiveram que ser testados.

Pelo lado do Excel, foi necessário aprender a programar em Visual Basic for Applications, sobretudo, aprender a:

- Sintaxe básica da linguagem, os tipos de dados, operadores e estruturas de controle;
- Criar menus;
- Carregar macros ao inicializar o Excel;
- Tratar eventos dos controles de interface gráfica de VBA;
- Manipular arquivos (criar, copiar, apagar e localizar);
- Invocar caixas de diálogo internas do Excel, como a de seleção de arquivo para abrir;
- Trabalhar com os objetos da biblioteca do Excel, como o Range, que representa um intervalo de células;
- Executar programas externos de modo síncrono, como o Gnuplot.

A.5. Atividades Realizadas

Basicamente, o projeto foi evoluindo realizando várias iterações no seguinte ciclo de atividades:

- 1) Especificação de requisitos para o sistema.
- 2) Análise / estudo dos requisitos.
- 3) Implementação das funcionalidades exigidas pelos requisitos.
- 4) Teste das novas funcionalidades implementadas. Eventualmente, a detecção de falhas aqui leva a uma reimplementação, voltando ao passo 3.

Após a implementação e teste de uma versão com um conjunto reduzido de requisitos e posterior aprovação pelo orientador, novos requisitos eram solicitados. Depois de analisados, novas funcionalidades eram implementadas (e testadas) para atender aos novos requisitos. Dessa forma, o ciclo se repetia e gradualmente, novas funcionalidades eram agregadas ao projeto.

A.6. Resultados Obtidos

O projeto ainda se encontra em vias de conclusão. No momento, ainda não estão disponíveis pela interface todos os tipos de gráficos 3D e todas as opções de formatação oferecidos pelo Gnuplot. Ainda existem algumas correções a serem feitas e recursos a serem adicionados na interface gráfica.

Entretanto, de modo geral, os gráficos gerados no Gnuplot usando essa nova interface têm mostrado ter a sua utilidade, servindo como complemento aos tipos de gráficos 3D já oferecidos pelo Excel.

A interface tem a limitação de poder ser utilizada apenas no ambiente do Excel/ Windows, para o qual foi desenvolvida, devido ao forte acoplamento criado pelo uso de objetos especiais da biblioteca do Excel para acessar os dados de uma tabela.

Estudos foram realizados com o objetivo de tentar gerar uma versão da interface para uso com o OpenOffice, que é um pacote de ferramentas para escritório nos moldes do Microsoft Office, mas oferecido gratuitamente e com versões para várias plataformas. A linguagem para programação de macros oferecida pelo OpenOffice é o Basic, precursora do Visual Basic.

O resultado dos estudos é que não há compatibilidade direta entre o ambiente de programação Basic do OpenOffice e o de VBA do Excel e portanto, a interface gráfica e alguns dos módulos que se utilizam de objetos da biblioteca do Excel precisariam ser reescritos para o OpenOffice. Porém, um aspecto positivo é que uma parte da lógica utilizada no código gerado para o Excel poderia ser aproveitada, em especial, a parte que grava o arquivo de comandos para o Gnuplot e, além disso, a modelagem das várias opções de formatação de gráfico do Gnuplot também poderia ser aproveitada.

A.7. Conclusão

O intuito deste projeto não é substituir as ferramentas de geração de gráfico existentes, mas sim, oferecer uma nova opção aos pesquisadores, para que tenham melhores condições de fazer análise da imensa amostra de dados coletados em suas pesquisas.

Espera-se que quando este projeto for concluído, a interface de integração do Excel com o Gnuplot possa ser de grande valia no trabalho daqueles que estejam buscando novos meios de analisar os dados coletados em suas pesquisas.

A.8. Referências Bibliográficas

Livro:

- Roman, Steven - *Desenvolvendo macros no Excel*, Editora Ciência Moderna, 2000

Páginas na Internet:

- *Microsoft MSDN* – msdn.microsoft.com
- *Microsoft Support* – support.microsoft.com
- *Oreilly.com - Writing Excel macros* – www.oreilly.com/catalog/exlmacro/
- *KB - Alertz* – www.kbalertz.com
- *Gnuplot Central* – www.gnuplot.info
- *Introduction to Gnuplot* – www.cs.uni.edu/Help/gnuplot/
- *Gnuplot development* – sourceforge.net/projects/gnuplot/

Outras:

- Ajuda on-line do Office Visual Basic for Applications
- Ajuda on-line do Excel Visual Basic for Applications

B. PARTE SUBJETIVA

B.1. Desafios e Frustrações Encontrados

Um dos maiores desafios no desenvolvimento do projeto de iniciação científica foi tentar conciliar o tempo dedicado à iniciação científica com o dedicado às disciplinas do BCC cursadas e com o estágio que iniciei no começo de maio. Tive maior dificuldade no primeiro semestre, no qual estava cursando 22 créditos-aula, sendo 20 em disciplinas do BCC com EP's e além disso, havia o estágio, que tomava 20 horas por semana. Com todo esse acúmulo de atividades, não consegui dedicar muito tempo à iniciação científica até o final do primeiro semestre. No segundo semestre, com uma carga horária bem menor no BCC, de apenas 8 créditos-aula, pude dedicar muito mais tempo ao projeto.

Outro desafio considerável e que, de certa forma, também foi fonte de frustração foi o fato de precisar aprender a programar em Visual Basic for Applications (VBA) no Excel. Por estar acostumado a programar em C e, principalmente, em Java, tive um pouco de dificuldade para me adaptar a VBA. O VBA não tem toda a eficiência e o poder de abstração de C nem tão pouco uma biblioteca de suporte tão rica quanto a oferecida por Java. Por ser interpretada no ambiente do Excel, há algumas restrições com relação ao que se pode fazer com VBA. Muitas foram os obstáculos encontrados no VBA que exigiram esforço considerável para contorná-los; os principais foram:

- O reduzido conjunto de controles de interface gráfica oferecidos permite criar somente interfaces gráficas muito simples.
- Os controles de interface gráfica também ofereciam pouca flexibilidade na alteração de propriedades. O controle disponível para exibição de imagens, por exemplo, permite carregar apenas imagens nos formatos .EMF/ .WMF, .BMP, .ICO e .GIF. Assim, as únicas opções para formato da imagem gerada pelo Gnuplot seriam o .EMF e o .GIF.
- VBA só permitia a execução de outros programas de forma assíncrona. Assim, não era possível executar um outro programa cuja saída fosse dependente à uma macro de VBA. No caso do projeto, o Gnuplot, responsável pelo gráfico a ser exibido numa janela da interface gráfica em VBA, não podia ser executado utilizando apenas as funções básicas oferecidas pela biblioteca de VBA. A solução foi utilizar funções da biblioteca do Windows para poder executar o Gnuplot sincronamente e obter corretamente a saída do programa.
- O modelo de tratamento de eventos de controles de interface gráfica no VBA é bem diferente do de Java. Em VBA, os controles têm propriedades, que ao serem alteradas em qualquer ponto do código, geram eventos que são tratados pelas rotinas de tratamento correspondentes. Esse modelo não é muito intuitivo, já que normalmente, espera-se que um evento só seja gerado pelo controle quando há alteração de uma de suas propriedades pela interação com o usuário e, nunca quando essas propriedades são alteradas programaticamente no código.

B.2. Disciplinas do BCC mais Relevantes para o Trabalho

Para a realização do projeto, as disciplinas do BCC que mais relevantes foram:

- *Introdução a Programação* (MAC-110), *Princípios de desenvolvimento de algoritmos* (MAC-122) e *Estruturas de Dados* (MAC-323) – por permitirem contato com algoritmos e conceitos básicos de programação
- *Laboratório de Programação I* (MAC-211) e *Laboratório de Programação II* (MAC-242) – por estimularem a prática de programação em projetos relativamente longos e envolvendo outras pessoas.
- *Conceitos Fundamentais de Programação* (MAC-316) – por fornecer a base teórica para identificar os elementos básicos de qualquer linguagem de programação e assim, criar condições para que seja fácil compreender qualquer linguagem de programação.
- *Análise de Algoritmos* (MAC-338) – por ajudar a analisar melhor a complexidade de um algoritmo e, por consequência, ajudar a buscar algoritmos mais eficientes.

B.3. Interação com o Orientador e Outros Profissionais

Durante todo o desenvolvimento do projeto, fui auxiliado pelo mestre em Ciência da Computação pelo IME, Luciano Vieira Araújo, que há muito tempo tem trabalhado em diversos projetos no Laboratório de Bioinformática do IME.

Enquanto meu orientador, o professor João Eduardo Ferreira passava orientações gerais sobre quais requisitos deveriam ser implementados, o Luciano acompanhava dia-a-dia o andamento do projeto, passando orientações mais específicas sobre as funcionalidades desejáveis na interface de integração do Excel com o Gnuplot.

A convivência com ambos foi muito tranquila e, sem dúvida nenhuma, pude aprender bastante.

B.4. Aplicação Prática de Conceitos Estudados

Os conceitos aprendidos nas disciplinas básicas do BCC certamente foram de grande ajuda para o projeto, ajudando na compreensão da linguagem VBA e no desenvolvimento de abstrações para solucionar problemas e na geração de código mais eficiente.

B.5. Passos Futuros para Aprimoramento dos Conhecimentos

Pretendo aprimorar os conhecimentos adquiridos nesses quatro anos de BCC realizando mestrado no IME, em alguma das minhas áreas de maior interesse: sistemas distribuídos, programação orientada a objetos, sistemas de banco de dados ou bioinformática. Apesar dessas duas últimas estarem mais relacionadas com o trabalho de formatura que desenvolvi, ainda não optei por nenhuma das quatro áreas e então, existe a possibilidade de seguir para uma área bem distante da área na qual se insere o meu trabalho de formatura.