Uma metodologia para determinação do organismo de origem de seqüências de DNA com aplicação em projetos EST

João Paulo Piazza

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Uma metodologia para determinação do organismo de origem de seqüências de DNA com aplicação em projetos EST

João Paulo Piazza

Junho de 2004

Banca Examinadora:

- João Carlos Setubal (Orientador)
- André Carlos Ponce de Leon Ferreira de Carvalho
  SCE - ICMC - USP
- João Meidanis
  IC - UNICAMP
- Ricardo Dahab
  IC - UNICAMP

Uma metodologia para determinação do organismo de origem de seqüências de DNA com aplicação em projetos EST

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por João Paulo Piazza e aprovada pela Banca Examinadora.

Campinas, 05 de julho de 2004.

João Carlos Setubal (Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Resumo

Este trabalho apresenta uma nova metodologia para a determinação computacional do organismo de origem de seqüências de DNA, implementada na forma de um programa chamado QUEST. O QUEST é baseado em dois princípios: a extração de informações intrínsecas a cada seqüência, chamadas de características, e a extração de diferentes tipos de características e sua combinação para se chegar a melhores resultados. São utilizados 7 diferentes programas como extratores de características, alguns desenvolvidos por terceiros (GLIMMER e ESTSCAN) e outros desenvolvidos pelo autor. As características foram combinadas utilizando vários classificadores diferentes, variando desde uma soma simples até os baseados em vetores de suporte. O QUEST requer seqüências para treinamento. Em comparação com as abordagens baseadas em similaridade, as vantagens principais do QUEST estão no fornecimento de previsões para as taxas de erro e na capacidade de lidar com seqüências sem similaridades significativas em bancos de seqüências.

O QUEST foi aplicado ao problema de determinar automaticamente contaminantes em projetos EST. São apresentados resultados de experimentos simulados e de um projeto EST real (o projeto EST de *Schistosoma mansoni*). Nos experimentos simulados foram atingidas taxas de falsos positivos mais falsos negativos de aproximadamente 10%. No projeto de *S.mansoni* o QUEST sugere que a contaminação em seqüências supostamente legítimas poderia ser de pelo menos 6%. No teste com *S.mansoni*, o QUEST foi 10 vezes mais rápido que o tempo necessário para executar o BLASTX em todas as seqüências testadas. O QUEST tem outras aplicações, incluindo a determinação do organismo de origem na nova abordagem genômica chamada de genômica ambiental (também chamada de metagenômica).

Abstract

This work presents a new methodology for computational ascertainment of organismal origin of DNA sequences, which we call QUEST. QUEST is based on two principles: that of extracting intrinsic information from each sequence, which are called features, and of extracting different kinds of features and combining them to achieve a better result. We use as feature extractors 7 different programs, some third-party (GLIMMER and ESTSCAN) and others developed by the author. We combine features using many different standard classifiers, ranging from simple sum to support vector machines. QUEST requires training sequences. In comparison to similarity-based approaches, QUEST has the main advantages of providing predicted error rates and of being able to deal with sequences without a significant match in sequence databases.

We applied QUEST to the problem of automatically determining contaminants in EST projects. We present results from a simulated experiment and from a real EST project (the *Schistosoma mansoni* EST project). In the simulated experiment we achieved rates of false positives plus false negatives of around 10%. In the *S.mansoni* project QUEST suggests that contamination in supposedly bona fide sequences may be of at least 6%. In the *S.mansoni* test, QUEST was 10 times faster than the time it took to run BLASTX on all tested sequences. QUEST has a number of other applications, including the determination of organismal origin in the new approach to genomics called environmental genomics (also called metagenomics).

The silent stars turn above me
And slowly pass on their way to morning
And I must rise up and go now
And make my own long journey to dawn

And I must leave those who love me
And close my mind to the cries and warnings
I understand and I know now
Why I must make this journey to dawn

This is all as it should be
I have to do all I can
I must become all I could be
And stand up tall as a man
Then at last my longing eyes
Gaze on the blaze of the sunrise
I'll touch the sky with my fingers
And take my prize

Already I see before me
The first pale colours of early morning
I turn my face to the glow now
And set off on my journey to dawn

Milton Nascimento, Márcio Borges, Gene Lees - *Journey to Dawn*

Agradecimentos

Meus agradecimentos iniciais ao professor João Carlos Setubal, por sua constante disposição, paciência, sensatez, orientação e incentivo, em todos os momentos.

Agradecimentos especiais a João Paulo Kitajima, por ter me dado a primeira chance na área da biologia computacional, pelo incentivo ao ingresso no mestrado e na vida acadêmica, e pela compreensão e suporte indispensáveis durante toda a minha vivência na área.

Agradecimentos ao Instituto de Computação e à Unicamp, pela excelência em seus cursos de graduação e pós-graduação. Agradecimentos ao LBI por me oferecer o primeiro trabalho em biologia computacional, pela participação em tantos projetos interessantíssimos, e por me apresentar esse povo tão legal que trabalhou e ainda trabalha por lá. Agradeço também à Alellyx, por todo o apoio dado e pela oportunidade de continuar trabalhando na área com tantos outros projetos interessantes.

Meus agradecimentos ao professor Sérgio Verjovski-Almeida, pelo auxílio na obtenção das seqüências de *Schistosoma mansoni* utilizadas neste trabalho. Obrigado também ao Instituto Ludwig pelo fornecimento de bolsa pelo trabalho no projeto *Human Cancer Genome Project* e à Embrapa pela bolsa fornecida pelo trabalho no projeto do café, abrangendo meu período de permanência no LBI, de setembro de 2000 a dezembro de 2002.

Agradecimentos finais e imensuráveis a meus pais, minha namorada, meu irmão, meus sobrinhos, meus amigos e todos aqueles várias vezes abandonados mas jamais esquecidos nestes dois anos de muito trabalho.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Este trabalho se insere no contexto de extração de informação de seqüências de DNA. Essas informações podem ser utilizadas para responder a uma questão elementar, dada uma seqüência de DNA: a partir de qual organismo a seqüência foi originada ? Essa pergunta é importante para resolver diversos problemas da biologia computacional: detecção de contaminantes em projetos de seqüenciamento, distinção dos organismos de origem em seqüências obtidas através de metagenômica, dentre outros.

Em particular, o foco deste trabalho são as contaminações presentes em projetos de seqüenciamento denominados projetos EST. Este é um problema comum neste tipo de projeto, e que está presente também naquele que foi o principal motivador para este trabalho: o projeto EST de *Schistosoma mansoni* [29], parasita humano de grande importância. Neste projeto, as contaminações foram detectadas e eliminadas por meio de uma abordagem baseada em similaridade. Apesar da metodologia empregada ter conseguido identificar diversas possíveis contaminações, várias questões ficaram em aberto, como por exemplo as estimativas de erros e a maneira de se definir os melhores parâmetros para a classificação. Isso motivou o desenvolvimento e uso de uma abordagem diferente, baseada no reconhecimento de padrões em seqüências de DNA, como será visto ao longo deste trabalho.

A principal contribuição deste trabalho é a apresentação de uma metodologia alternativa a da similaridade, que combina diversas evidências para classificar seqüências com base em sua origem, além de fornecer taxas de erros estimadas sobre a classificação, características que não estão presentes nas abordagens convencionais baseadas em similaridade.

Os resultados parciais deste trabalho resultaram também em uma publicação e apresentação no 2º Workshop Brasileiro sobre Bioinformática [23] (WOB), além de futura publicação na Revista Tecnologia da Informação (RTInfo).

Este trabalho está organizado da seguinte maneira: no Capítulo 2 são apresentados os conceitos básicos utilizados. São descritos conceitos de biologia molecular, biologia

computacional e de reconhecimento de padrões.

O Capítulo 3 descreve duas maneiras de se determinar a origem de uma seqüência de DNA: a primeira baseada em similaridade, e a segunda baseada em reconhecimento de padrões, descrita em maiores detalhes ao longo deste trabalho. Uma comparação sobre os aspectos positivos e negativos de cada abordagem também é feita.

No Capítulo 4 é apresentada a metodologia e implementação desenvolvida neste trabalho para lidar com o problema da determinação da origem de seqüências de DNA, implementação denominada QUEST.

O Capítulo 5 apresenta os resultados do QUEST para um conjunto de projetos EST criados artificialmente, a fim de avaliar a ferramenta. São também apresentados e discutidos os resultados da aplicação do QUEST sobre o projeto EST de *Schistosoma mansoni*, projeto utilizado também para medições do tempo de execução do QUEST.

No Capítulo 6 são apresentadas as conclusões deste trabalho, sendo descritas as principais contribuições e melhorias previstas para a metodologia.

Por fim, o apêndice contém uma descrição da interface do QUEST e de seus módulos principais.

Capítulo 2

Notação e conceitos

2.1 Notação básica

Neste trabalho, um conjunto é sempre representado por uma ou mais letras maiúsculas, e seus elementos não ordenados representados pelas mesmas letras minúsculas seguidas ou não de um índice, como por exemplo o conjunto $A = \{a_1, a_2, ..., a_n\}$, com $n = |A|$.

Da mesma maneira, um vetor é denotado por uma ou mais letras maiúsculas, sendo os seus elementos ordenados representados pelas mesmas letras minúsculas acompanhadas de um índice. O vetor $V = [v_1, v_2, ..., v_m]$, com $m = |V|$ exemplifica esta notação.

2.2 Conceitos básicos de biologia molecular

Todos os seres vivos possuem uma característica em comum: eles possuem a capacidade da reprodução. Com exceção dos vírus, todo ser vivo é formado por uma ou mais células. Cada célula é envolta por uma membrana, que separa o seu conteúdo (denominado citoplasma) do ambiente que a cerca, e que permite a obtenção de substâncias necessárias, descarte de metabolitos indesejados e interação com o ambiente [19, 22].

Dentro do citoplasma, encontram-se uma série de estruturas responsáveis pelo funcionamento e multiplicação da célula. Essas estruturas têm funções bastante diversas. A presença ou ausência de algumas delas é usada como critério para a diferenciação de dois grandes grupos de organismos: aqueles que têm estruturas denominadas *núcleo* e *organelas* separadas por membranas são denominados organismos *eucariotos*. Os organismos que não possuem essa separação são denominados de organismos *procariotos*, e são em geral considerados menos complexos que os eucariotos. Nos organismos eucariotos, o material genético (fundamental para a reprodução da célula) encontra-se encerrado no núcleo, enquanto nos procariotos esse material encontra-se disperso no citoplasma. As

organelas também contêm material genético. Como exemplos de organelas pode-se citar as *mitocôndrias*, responsáveis pela produção de energia, e os *cloroplastos*, onde ocorre a fotossíntese.

Tanto organismos eucariotos quanto procariotos perpetuam-se através da reprodução. Através dela o material genético do organismo é copiado, originando-se um novo organismo da mesma espécie, evento este também denominado de *transferência vertical* de material genético. Há também situações em que o material genético de um organismo é transferido para um organismo de outra espécie, caracterizando o evento da *transferência horizontal* ou *lateral* de material genético.

O material genético de uma célula está organizado em *unidades replicáveis*, também chamadas de *replicons*. Como exemplos de replicons pode-se citar os cromossomos, os plasmídeos e os cosmídeos, estes últimos exclusivos de procariotos. O conjunto de replicons de um organismo corresponde ao seu *genoma*. Os replicons podem ser lineares ou circulares, e são compostos pelo denominado *ácido desoxiribonucleico*, ou simplesmente *DNA*[1].

O DNA é uma molécula composta pelo encadeamento de moléculas menores denominadas *nucleotídeos*. Os nucleotídeos têm em sua composição um fosfato, uma ribose e uma base nitrogenada que os diferencia. No DNA podem existir quatro tipos de nucleotídeos, denominados de acordo com a base nitrogenada que possuem: *adenina*, *citosina*, *guanina* e *timina*, geralmente representadas por suas iniciais `A`, `C`, `G` e `T`, respectivamente. Uma seqüência de nucleotídeos representando um fragmento de DNA é denotada neste trabalho simplesmente por *seqüência*.

O DNA nos replicons é encontrado sempre aos pares, o que confere maior estabilidade ao conjunto. As moléculas de DNA em cada par são também denominadas de *fitas*, unidas por interações entre as bases nitrogenadas de seus nucleotídeos. Uma adenina sempre se liga a uma timina, e uma citosina sempre se liga a uma guanina, formando neste sentido fitas ditas *complementares*. A primeira fita é comumente denominada *fita senso* ou *fita* +, e a segunda *fita antisenso* ou ainda *fita* −. Cada fita tem duas extremidades livres, denominadas *5'* e *3'*. A extremidade 5' de uma fita está sempre ligada à extremidade 3' da outra fita, e por isso as fitas são ditas *antiparalelas*. Uma representação plana da ligação entre duas fitas é dada abaixo

```
fita senso      5' C A G T C A G C A G 3'
                   | | | | | | | | | |
fita antisenso  3' G T C A G T C G T C 5'
```

Uma unidade de medida comumente usada para moléculas de DNA em fita simples é seu número de nucleotídeos (abreviado para *nt*), ou ainda pares de bases (*bp*[2]), quando o

[1] em inglês: *desoxyribonucleic acid*

[2] abreviação do inglês *base pairs*

DNA está na forma de fita dupla (como nos replicons). O genoma dos organismos varia entre aqueles tão pequenos quanto o do HIV tipo I (9.181bp), a tão grandes quanto o da ameba *Amoeba dubia*, com 6.7×10^{11}bp.

O DNA dos replicons possui regiões especiais contendo diversas receitas codificadas para a produção de proteínas e RNAs[3], regiões estas denominadas *genes* [25]. Isso é feito através da *tradução* do DNA dessas regiões em proteína, em um processo descrito mais adiante. Um gene pode estar em qualquer uma das fitas de DNA de um replicon.

Apenas parte de cada replicon é composta por genes, sendo o restante composto por regiões de função ainda desconhecida, denominadas *regiões intergênicas*. A proporção entre genes e regiões intergênicas nos organismos varia bastante. Nos organismos procariotos a proporção é de aproximadamente 90% e 10%, e em eucariotos de 5% e 95%, respectivamente para genes e regiões intergênicas [22].

Os genomas de organismos de diferentes espécies podem conter genes com seqüências muito semelhantes, presentes em algum ancestral comum a ambos, e que foram mantidos de certa forma *conservados* por pressão evolutiva. Esses genes são chamados de *genes ortólogos*.

2.2.1 Dogma central da biologia molecular

Simplificadamente, para um gene ser transformado em uma proteína, seu DNA correspondente deve ser *transcrito* em RNA, e esse RNA traduzido em uma proteína. Este é o chamado *dogma central da biologia molecular*. A Figura 2.1 ilustra este processo, descrito com maiores detalhes a seguir.

Contexto para a produção da proteína

Quando a proteína codificada por um certo gene A é necessária à célula, seu aparato de tradução "percebe" a mudança de contexto ocorrida e inicia-se a *expressão* do gene A. A expressão dependente de contexto é importante como forma de resposta a estímulos externos e para a expressão regulada dos genes.

Cópia em mRNA

O mesmo DNA de um replicon deve servir para a expressão de um grande número de genes, possivelmente inúmeras vezes. Portanto, a expressão de um gene deve preservar a integridade do DNA. Através da *transcrição*, uma cópia de parte do gene é feita em

[3]ácido ribonucleico, de composição semelhante a do DNA, mas com uma *ribose* em lugar da desoxiribose, e a base nitrogenada *uracila* em lugar da timina

RNA. Especificamente, a transcrição dá origem a um RNA denominado mensageiro, ou simplesmente *mRNA*.

Processamento do mRNA

No caso dos organismos eucariotos, o mRNA gerado pela transcrição contém regiões denominadas *exons* e *introns*. Destas regiões, apenas os *exons* serão traduzidos em proteínas; os introns, portanto, deverão ser removidos do mRNA. Isso é feito através de um processo chamado de *splicing*, resultando em um *mRNA maduro*. Ocorre também o chamado *splicing alternativo*, onde um ou mais exons acabam não estando presentes no mRNA maduro.

Nos eucariotos o mRNA maduro recebe em sua extremidade 3' uma seqüência de 20 a 200 adeninas para conferir-lhe maior estabilidade enquanto o mRNA se dirige do núcleo para o citoplasma. Esta seqüência adicional recebe o nome de cauda poli-A.

Tradução do mRNA

O mRNA processado deve ser finalmente traduzido em uma proteína. Em organismos eucariotos, as extremidades do mRNA maduro contêm regiões não traduzidas[4] (UTR). A porção do mRNA que de fato é traduzida em proteína é denominada de *seqüência codificadora*[5] (CDS). O CDS é processado em grupos de três nucleotídeos denominados de *codons*. Para o processamento do CDS e seus codons diversas outras moléculas são necessárias:

RNA transportador (tRNA) Dentro da célula existem diversos RNAs chamados de *RNAs transportadores*, ou simplesmente *tRNAs*. Estes tRNAs são transcritos a partir dos replicons, mas não são traduzidos. Eles contêm em uma extremidade um grupamento de três nucleotídeos, chamados de *anti-codons*, e na outra extremidade um entre 20 aminoácidos possíveis. Em um organismo, cada anti-codon terá sempre o mesmo aminoácido associado, e essa relação determina o chamado *código genético*. Por exemplo, o codon `AUG` está sempre associado ao aminoácido *metionina* (*Met*). Diversos codons podem estar associados a um mesmo aminoácido, existindo ainda codons especiais que sinalizam o término da tradução (*codons de parada*).

RNA ribossomal (rRNA) Assim como os tRNAs, os *rRNAs* são transcritos mas não traduzidos. Conjuntamente com determinadas proteínas, os rRNAs formam os *ribossomos*, que irão promover a interação entre os codons do CDS e os anti-codons dos tRNAs para efetuar a tradução do CDS.

[4] em inglês: *untranslated regions*
[5] em inglês: *coding sequence*

Mediado por ribossomos, cada codon do CDS é pareado com o anti-codon presente em um tRNA, liberando o aminoácido antes ligado ao tRNA, até alcançar um codon de parada. Cada aminoácido recém liberado liga-se aos outros aminoácidos já liberados, formando enfim a proteína, conforme ilustrado na Figura 2.1.

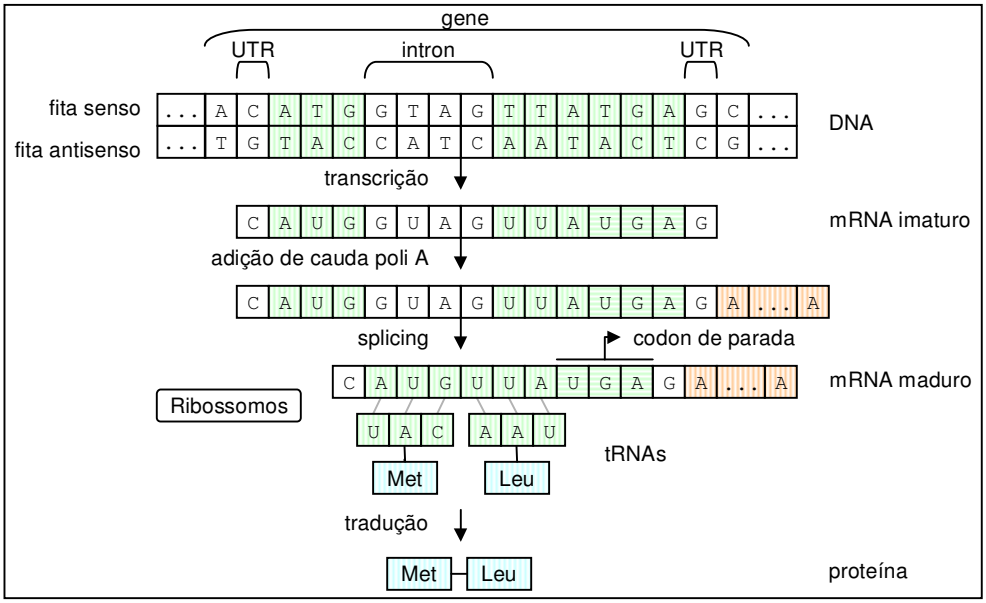

Figura 2.1: Diagrama da expressão de um gene em um organismo eucarioto. O produto final neste exemplo é uma proteína composta pelos aminoácidos *metionina* (*Met*) e *leucina* (*Leu*).

2.3 Conceitos básicos de projetos de seqüenciamento

Uma vez que o material genético dos organismos carrega praticamente toda a informação necessária para a sua constituição, ele é uma importante fonte de estudo. Por isso, foram desenvolvidas técnicas para efetuar sua leitura, através de um processo denominado *seqüenciamento*. Através do seqüenciamento é possível a leitura dos nucleotídeos que compõem uma molécula de DNA de um certo *organismo alvo*.

No entanto, a tecnologia atual de seqüenciamento consegue efetuar a leitura de cerca de 1.000 nucleotídeos em seqüência [7, 22]. Dessa forma, para a leitura de todos os nucleotídeos de um genoma, o DNA de seus replicons deve ser *fragmentado* em pedaços menores. Cada fragmento lido é representado por uma série de letras `A`, `C`, `G` e `T`, correspondendo aos nucleotídeos existentes na molécula de DNA de origem. Associada a cada letra há um valor de qualidade, variando entre 0 e 99; quanto maior o valor desta qualidade, maior a possibilidade da letra corresponder ao nucleotídeo presente no DNA.

Outra restrição da tecnologia atual de seqüenciamento é que ela requer a existência de inúmeras cópias do mesmo DNA alvo para funcionar, o que é conseguido através da chamada *amplificação*. A técnica de amplificação pode resultar na inserção de DNAs auxiliares nas extremidades do DNA alvo, como por exemplo dos chamados *primers* e *vetores*. Como nem sempre é possível eliminar essas seqüências auxiliares antes do processo de seqüenciamento, elas devem ser identificadas e eliminadas após o mesmo[6].

Uma vez seqüenciados os fragmentos, deve-se então *montá-los* novamente, utilizando-se as sobreposições existentes entre eles, de forma a que voltem a representar a molécula de DNA de onde foram originados. No entanto, a montagem não é um processo simples; sua descrição e dificuldades estão descritas mais adiante na Seção 2.4.2.

Dependendo da técnica de seqüenciamento utilizada, as seqüências lidas podem estar em qualquer orientação (senso ou anti-senso). Além disso, no caso de seqüências provenientes de regiões codificadoras, o *quadro de leitura* não será conhecido. O quadro de leitura de uma seqüência indica a partir de qual nucleotídeo inicia-se o primeiro codon presente para efetuar a tradução em proteína. Há três possibilidades na fita senso (denominadas quadros de leitura +1, +2 e +3) e outras três na fita anti-senso (quadros de leitura -1, -2 e -3), conforme ilustrado na Figura 2.2. Os codons formados em cada quadro estão destacados.

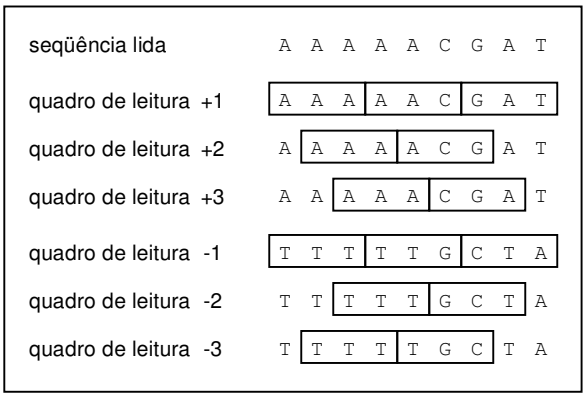

Figura 2.2: Exemplo dos quadros de leitura possíveis para uma mesma seqüência de DNA. Cada grupo de três nucleotídeos em destaque representa um codon no quadro de leitura correspondente.

[6]este processo é chamado usualmente de *screening*

2.3.1 Principais tipos de projetos de seqüenciamento

Existem dois tipos principais de projetos de seqüenciamento, denominados respectivamente de *projetos genoma* e *projetos EST*[7] [1]. Em um projeto genoma, o objetivo é obter as seqüências completas dos DNAs de todos os replicons de uma determinada espécie. Isso é feito em geral para organismos procariotos, uma vez que o tamanho de seus genomas é pequeno quando comparado com o genoma dos eucariotos. No entanto, projetos genoma para organismos eucariotos de maior interesse são também feitos, como por exemplo para o ser humano e para a planta modelo *Arabidopsis thaliana*. O produto final de um projeto genoma são diferentes seqüências representando os replicons do organismo. É comum também que seja feita a *anotação* da localização e função dos genes presentes no genoma.

Já em um projeto EST, o objetivo é a obtenção de uma parte do *transcriptoma* de um organismo, ou seja, uma amostra dos mRNAs maduros produzidos por seu genoma. Apesar de não se obter ao final do projeto o genoma do organismo alvo, os projetos EST apresentam vantagens sobre os projetos genoma:

- Menor custo, uma vez que pode ser feita uma amostragem proporcional aos recursos disponíveis para o projeto, uma opção não presente nos projetos genoma.
- Amostragem apenas das regiões transcritas do genoma, que em geral são de maior interesse.

Conforme ilustrado na Figura 2.3, uma seqüência EST é obtida partindo-se de um mRNA maduro, transformado através de *transcrição reversa* em DNA (neste caso chamado de *cDNA*[8]), e por fim amplificado e seqüenciado.

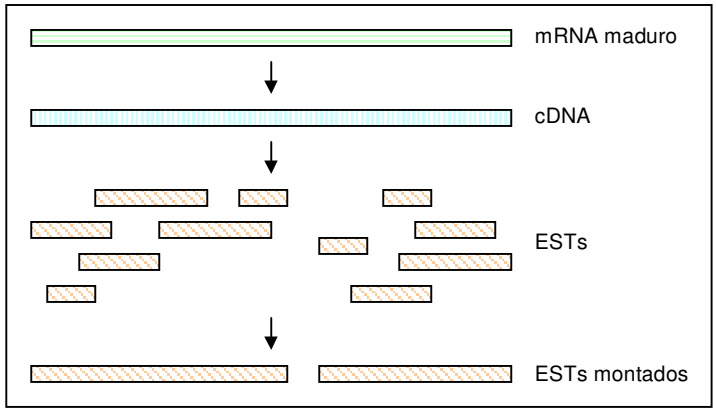

Figura 2.3: Diagrama mostrando a obtenção de ESTs a partir de um cDNA.

[7] do inglês: *Expressed Sequence Tag*
[8] do inglês: *complementary DNA*

Assim como em um projeto genoma, a montagem das seqüências lidas em um projeto EST também é necessária. Se duas ESTs representam amostras de diferentes posições de um mesmo cDNA, e se elas têm sobreposição, a montagem irá transformá-las em apenas uma seqüência, diminuindo o número de seqüências para análise, e eliminando a redundância (mais de duas seqüências representando o mesmo trecho do mesmo DNA de origem). Uma ilustração de montagem é dada na Figura 2.3.

2.3.2 Problemas

Tanto projetos genoma quanto projetos EST apresentam dificuldades, que podem comprometer o andamento e resultados dos mesmos. O objetivo deste trabalho é lidar com o problema da *contaminação* por seqüências de organismos que não o organismo alvo especificamente em projetos EST. Em projetos genoma as seqüências contaminantes acabam não entrando nas montagens finais, visto que não têm similaridade com as demais seqüências. As origens e problemas decorrentes das contaminações são discutidos a seguir.

Contaminações

Um dos problemas que podem acontecer em projetos de seqüenciamento é a ocorrência de seqüências contaminantes dentre as seqüências do projeto. Por contaminante entende-se quaisquer seqüências obtidas mas que não são de interesse do projeto.

Muitas vezes, seqüências provenientes do próprio organismo alvo são consideradas contaminações. Por exemplo, seqüências de mitocôndrias, cloroplastos, rRNAs ou quaisquer outras seqüências obtidas mas que não seriam traduzidas (exceto as UTRs) não são em geral de interesse em projetos EST. Contaminações do próprio organismo alvo não são no entanto tratadas neste trabalho.

Por outro lado, são bastante comuns as contaminações por organismos diferentes do organismo alvo, denominados *organismos contaminantes*. Este tipo de contaminação pode surgir por diversos motivos, como por exemplo:

(1) organismos que coexistem com o organismo alvo podem ser difíceis de separar. Essa separação é ainda mais complicada se o organismo alvo viver em *simbiose* com algum outro, pois nesse tipo de relação há ainda uma maior proximidade entre ambos.

(2) manipulação incorreta de equipamentos e reagentes, fazendo com que o material genético de outros organismos esteja presente indevidamente no momento da extração do material do organismo alvo. Material genético indesejado pode ser introduzido por exemplo por recipientes mal esterilizados.

(3) no caso de um laboratório trabalhar com mais de um organismo no mesmo período, pode ocorrer de placas de seqüenciamento de um organismo serem erroneamente rotuladas como de outro organismo. Esse tipo de erro já foi verificado mesmo em repositórios públicos de seqüências, como o genbank [6] (detalhes em *http://www.ncbi.nlm.nih.gov/dbEST/synopsis_detailsR.html*).

Embora haja técnicas laboratoriais para diminuir a ocorrência das contaminações descritas anteriormente, elas raramente podem ser removidas por completo. Se mantidas, as seqüências contaminantes serão confundidas com seqüências do organismo alvo, acarretando erros em diversas estatísticas e análises do projeto. Por exemplo, o número de genes do organismo alvo pode ser superestimado, ou ainda o seqüenciamento pode ser interrompido antes de se atingir o número de seqüências correto previsto pelo projeto.

Dada a importância do problema e a impossibilidade de eliminá-lo usando-se apenas técnicas laboratoriais, é necessário lidar com as contaminações através de técnicas de biologia computacional, descritas mais adiante.

2.4 Conceitos básicos de biologia computacional

Por biologia computacional entende-se o uso de técnicas e ferramentas da ciência da computação aplicadas aos problemas da biologia [25], notadamente para as áreas de biologia molecular e genética. A biologia computacional é uma área bastante ampla; por isso, nesta seção há apenas a descrição dos conceitos mais relevantes para o trabalho.

2.4.1 Similaridade entre seqüências

A comparação de seqüências é um problema fundamental da biologia computacional. Quando se compara duas seqüências, busca-se na verdade um *alinhamento* entre elas. Por exemplo, para a comparação de duas seqüências $s =$ `GACGGATTA` e $r =$ `GATCGGAATA`, poderia ser construído o seguinte alinhamento [25]:

seqüência s	G	A	-	C	G	G	A	T	T	A
seqüência r	G	A	T	C	G	G	A	A	T	A
casamento ?	S	S	N	S	S	S	S	N	S	S

No entanto, este é apenas um dentre vários alinhamentos possíveis. Assim, é necessário algum critério para a escolha de um alinhamento. Como exemplo de critério, pode-se criar um sistema de pontuação baseado no alinhamento: caso as letras casem (letra "S" no exemplo), tem-se um valor positivo; caso contrário, tem-se um valor negativo. Assim, a *similaridade* é definida pelo valor da pontuação do alinhamento que a maximiza. Pode

haver ainda variações no problema; por exemplo, pode-se precisar alinhar apenas trechos de seqüências (prefixo, sufixo ou regiões internas). O alinhamento de três ou mais seqüências é denominado *alinhamento múltiplo*.

Além do valor da similaridade dado por um alinhamento com certo esquema de pontuação, o *percentual de identidade* também é uma medida importante, e é simplesmente o percentual de casamentos feitos com relação ao total de casamentos possíveis no alinhamento correspondente. No exemplo dado, esse valor é de 80%.

2.4.2 Montagem de seqüências

A montagem de seqüências é um processo fundamental em projetos de seqüenciamento, como forma de agrupar as seqüências provenientes de um mesmo DNA e conhecer sua composição original. A montagem envolve basicamente três etapas principais [25]:

(1) agrupamento das seqüências em grupos relacionados, por exemplo através da similaridade entre elas. Cada grupo deve idealmente ter sido originado a partir de um mesmo DNA.

(2) construção de um alinhamento múltiplo para cada grupo.

(3) determinação do consenso para cada grupo.

Um exemplo simplificado das etapas **(2)** e **(3)** é dado abaixo (retirado de [22]):

seqüências relacionadas	alinhamento
ACGGAGCA	ACGGAGCA----------
GAGCACTTG	---GAGCACTTG--------
CTTGAGTC	--------CTTGAGTC----
GGAGCACT	--GGAGCACT----------
TGAGTCAAAC	----------TGAGTCAAAC
GCACTTG	-----GCACTTG--------
consenso	ACGGAGCACTTGAGTCAAAC

A montagem é um problema computacional bastante complexo, devido a uma série de complicadores que ocorrem com freqüência, sejam eles de origem biológica (B) ou biotecnológica (BT). Uma lista de alguns desses complicadores é dada abaixo:

Cauda poli-A/T (B) a cauda poli-A (ou poli-T se tiver sido seqüenciada a outra fita) pode estar presente em várias seqüências, o que pode fazer com que elas sejam agrupadas indevidamente, por compartilharem estas regiões similares.

Seqüências repetitivas (B) seqüências repetitivas são aquelas que ocorrem mais de uma vez em posições diferentes de um DNA. Se a amostra seqüenciada estiver contida inteiramente em uma repetição, o processo de montagem poderá colocar a amostra em um grupo de seqüências incorreto.

Splicing alternativo (B) seqüências provenientes do mesmo mRNA mas de diferentes formas de splicing (como por exemplo splicing alternativos) podem compartilhar semelhanças que poderão atrapalhar o processo de montagem.

Parálogos (B) genes parálogos são aqueles originados pela duplicação de um gene. Os mRNAs correspondentes podem conter semelhanças que atrapalham a montagem.

Erros nas seqüências (BT) nem sempre o seqüenciamento retorna a composição exata do DNA original, podendo haver troca, inserção ou deleção de nucleotídeos.

Orientação da seqüência (BT) conforme já descrito, o seqüenciamento pode resultar em seqüências tanto da fita senso quanto da fita anti-senso, e portanto as duas possibilidades devem ser levadas em conta.

Falta de cobertura (BT) se forem amostradas seqüências nas pontas de um DNA, e não houver seqüências representando o seu interior, serão formados na verdade dois consensos para um mesmo DNA, por não haver sobreposição entre eles. A região interna do mRNA representado na Figura 2.3 exemplifica uma região sem cobertura.

Quimeras (BT) pode ocorrer de dois ou mais fragmentos de DNA se unirem indevidamente antes do seqüenciamento, sendo então seqüenciados conjuntamente e representando a seqüência de uma molécula de DNA inexistente.

Seqüências de DNA auxiliares (BT) seqüências de primers e vetores quando não removidas podem atrapalhar a montagem.

Embora haja formalizações para lidar com o problema da montagem, são usadas na prática implementações heurísticas, como as encontradas nas ferramentas PHRAP [13] e CAP3 [15].

2.5 Conceitos básicos de reconhecimento de padrões

Como será discutido adiante, o foco deste trabalho está no *reconhecimento de padrões* em seqüências dos organismos alvo e contaminantes, com o objetivo de diferenciá-las. Nesta seção os conceitos básicos de reconhecimento de padrões utilizados são apresentados.

2.5.1 Visão geral

Um procedimento de reconhecimento de padrões opera basicamente sobre *objetos*. Um objeto pode possuir diversas *características*[9], que são como representações simplificadas para ele. Por exemplo, uma seqüência de DNA é um objeto que possui como características o seu tamanho, suas contagens de nucleotídeos etc.

Com base nas características de um objeto, pode-se categorizá-lo em diferentes classes, usando duas maneiras distintas: a primeira utiliza-se de um *sistema de aprendizado supervisionado*, e a segunda um *sistema de aprendizado não supervisionado*. Um sistema de aprendizado supervisionado categoriza um objeto como uma de n classes pré-definidas, usando para isso o aprendizado obtido com as características de objetos de classes já conhecidas (ditos *rotulados*). Já em um sistema de aprendizado não-supervisionado, nenhum objeto rotulado é fornecido *a priori*, e a classificação é feita com base em agrupamentos de objetos formados "naturalmente" através de características semelhantes [11].

Este trabalho implementa um sistema de aprendizado supervisionado, cujo objetivo é classificar seqüências de origem incerta como uma de duas classes pré-definidas, separando aquelas provenientes do organismo alvo (e portanto *legítimas*) daquelas originadas de algum outro organismo contaminante. As seções a seguir ilustram este processo.

2.5.2 Escolha das características

O passo inicial para se trabalhar em um sistema de aprendizado supervisionado é a escolha das características a serem obtidas a partir dos objetos; tais características devem ser escolhidas idealmente de forma a que objetos de uma mesma classe tenham valores próximos, e objetos de classes diferentes tenham valores distantes. Como exemplos de características bastante conhecidas em seqüências de DNA pode-se citar o chamado *conteúdo* `GC`, que é o seu número de bases `G` e `C` dividido pelo seu total de bases[10], e de maneira análoga o *percentual* `GC`, que é o valor do conteúdo `GC` multiplicado por 100.

Dependendo do problema, podem ser necessárias várias características para que se alcance taxas de erros adequadas na classificação.

[9] *em inglês:* features

[10] o genoma de cada espécie tem em geral um conteúdo `GC` relativamente constante mesmo em suas subcadeias [21]

2.5.3 Obtenção das características

Uma vez escolhidas as características de trabalho, elas devem ser extraídas dos objetos. Isso é feito através de um procedimento chamado de *extração de características*. Um determinado objeto é dado ao extratores escolhidos, que retornam os valores das características correspondentes. Por exemplo, um objeto s é associado através de dois extratores de características a um vetor X contendo as características x_1 e x_2

$$X = [x_1, x_2]$$

Cada característica empregada define uma dimensão no chamado *espaço de características*, como ilustrado na Figura 2.4. Cabe a um *classificador* distinguir as diferentes classes neste espaço.

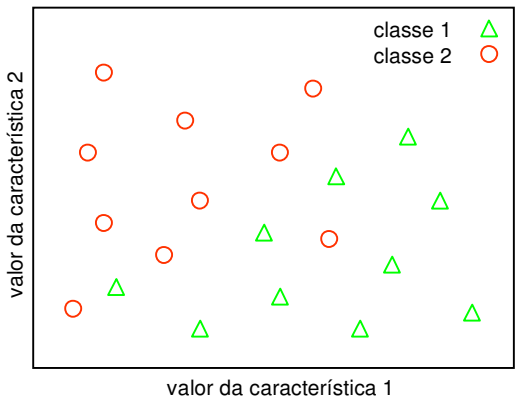

Figura 2.4: Exemplo de espaço de características de dimensão dois, contendo características de objetos de duas classes.

2.5.4 Classificadores

Um classificador em um sistema de aprendizado supervisionado deve analisar as características de objetos rotulados e usá-las para classificar objetos não rotulados fornecidos.

Mais especificamente, um classificador recebe como entrada um conjunto de características obtidas a partir de objetos rotulados. Através delas, ele deve decidir quais são seus parâmetros internos que fornecem uma melhor separação entre os objetos de cada classe. Esses parâmetros servirão de base para a classificação dos objetos não rotulados fornecidos, e devem ser escolhidos de forma a classificarem bem também este conjunto. Classificadores com essa característica possuem a chamada boa capacidade de *generalização*, ilustrada na Figura 2.5. A capacidade de generalização neste trabalho denota simplesmente a diferença entre os erros observados na classificação dos conjuntos não rotulados e rotulados.

Um classificador pode dar basicamente três tipos de resultados para cada objeto analisado: **(a)** simplesmente retornar a classe que ele considera mais provável; **(b)** retornar todas as classes, ordenadas da mais provável para a menos provável; **(c)** retornar todas as classes, juntamente com um valor de confiabilidade para cada uma. A forma **(c)** é a mais informativa, pois dela podem ser derivadas as demais. Esses resultados estão exemplificados na Tabela 2.1.

Tabela 2.1: Diferentes formas de um classificador retornar sua decisão sobre um conjunto de características de um objeto. Em todos os casos, a classe escolhida é a classe c_2. A forma (c) apresenta entre parênteses o valor de confiabilidade na predição dado pelo classificador.

forma	classe escolhida		
	1ª	2ª	3ª
(a)	c_2	-	-
(b)	c_2	c_1	c_3
(c)	c_2 (90)	c_1 (8)	c_3 (2)

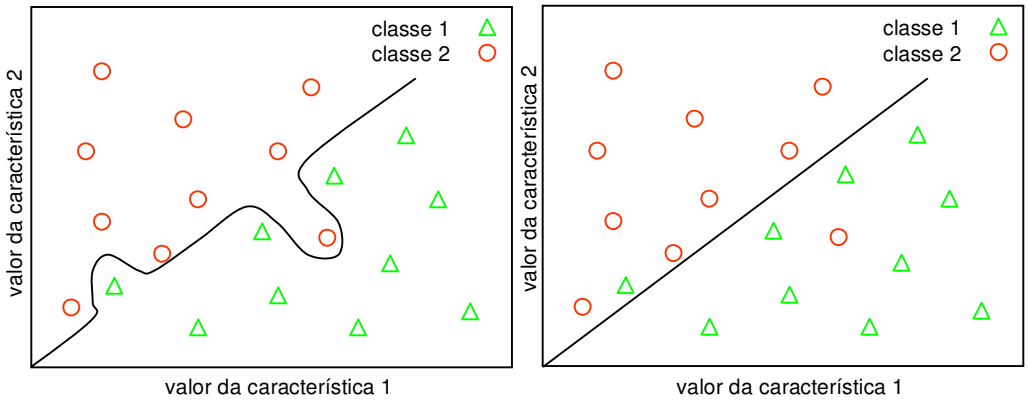

Figura 2.5: Distinção entre duas classes através de dois classificadores diferentes, representados pela linha separando as classes 1 e 2. Intuitivamente, o classificador da esquerda é bastante específico para as características apresentadas, podendo não ser tão bom para separar novos objetos quanto o classificador da direita.

Na análise do erro associado à classificação pode-se descobrir que ou o classificador ou as características utilizadas não são suficientes para que se obtenha o erro desejado. O uso de um novo classificador ou ainda a adição de novas características (através do uso de novos extratores) pode ajudar a diminuir esse erro. Características adicionais podem ajudar na diminuição dos erros pois podem conter informações complementares àquelas das características já empregadas.

2.5.5 Estimativa de erros

A decisão pelo uso de uma ou mais características ou entre um ou outro classificador é normalmente direcionada pelo *erro* associado a cada uma dessas escolhas. A maneira mais simples de se medir esse erro é calcular a quantidade de classes incorretamente atribuídas decorrentes da classificação do conjunto de objetos rotulados. Se este erro estiver acima do limite aceitável para o problema, pode ser necessário escolher outras ou mais características, outro classificador, ou no pior dos casos pode-se concluir que o problema é muito difícil para ser tratado desta maneira.

2.5.6 Seleção de subconjunto de características

Sob determinadas circunstâncias que podem depender dos objetos analisados, a adição de novas características pode piorar os erros obtidos. Neste caso, para se obter uma performance ótima utilizando as características disponíveis, pode ser necessária a escolha de um subconjunto dessas características de forma a minimizar o erro produzido na classificação. Este procedimento é chamado de *seleção de subconjunto de características*. A busca do subconjunto ótimo deve percorrer idealmente todas as combinações de características (no exemplo, $2^{|X|} - 1$), o que pode ser utilizado na prática apenas quando há um número pequeno de características. Para conjuntos maiores, diversas outras estratégias podem ser empregadas [24].

2.5.7 Combinação de classificadores

Podem existir classificadores que funcionam melhor quando trabalham com certas características, e outros que trabalham melhor com outras. Ainda assim, esses classificadores podem ser usados em conjunto a fim de melhorar os erros da classificação, através de alguma função f que combine suas predições individuais.

Por exemplo, dois classificadores p_1 e p_2 poderiam classificar bem respectivamente as características obtidas pelos extratores e_1 e e_2, e assim receberiam como entrada não o vetor X inteiro, mas apenas as características x_1 e x_2 apropriadas. Dessa maneira p_1 e p_2 retornariam suas predições individuais:

$$\begin{aligned} p_1([x_1]) &\rightarrow \textit{classe } c_2 \\ p_2([x_2]) &\rightarrow \textit{classe } c_1 \end{aligned}$$

Neste caso, os dois classificadores tiveram opiniões diferentes analisando os subconjuntos do vetor X. Uma função de combinação f recebe como entrada essas classificações, retornando uma classe para o vetor X em questão:

$$f(p_1([x_1]), p_2([x_2])) \rightarrow \textit{classe } c_2$$

Os combinadores são na verdade classificadores que recebem predições de outros classificadores ao invés de um vetor de características. Assim, os combinadores também podem retornar a classe escolhida de três formas diferentes (ver Tabela 2.1).

2.5.8 Classificação

Após a definição das melhores características, do classificador e combinador (se necessário) utilizados, e após a definição de parâmetros e estimativas de erros através da execução do sistema com objetos rotulados, é a vez dos objetos de origem desconhecida serem analisados. Cada objeto é associado a uma classe, podendo opcionalmente (dependendo da saída do classificador ou combinador) ter o restante das classes associadas em ordem decrescente de confiabilidade.

Capítulo 3

Métodos para determinação da origem de seqüências de DNA

3.1 Determinação da origem de uma seqüência: similaridade versus informação intrínseca

Dada uma seqüência de DNA qualquer, a pergunta mais básica que se pode fazer é: a partir de qual organismo esta seqüência foi obtida ? Esta pergunta é importante em diversos contextos: na detecção de seqüências contaminantes em projetos de seqüenciamento, na separação de seqüências obtidas a partir de organismos vivendo em simbiose, na distinção entre seqüências dos diferentes organismos amostrados via metagenômica [28] dentre outros.

Existem basicamente duas maneiras de se responder à pergunta acima: a primeira é procurando outra seqüência muito similar em um banco de seqüências de diversos organismos já conhecidos. A segunda é, conhecendo-se um padrão que caracteriza as seqüências de um determinado organismo X, pode-se buscar esse padrão na seqüência de origem desconhecida; se encontrado, então postula-se que o organismo da seqüência de origem desconhecida é X; se não encontrado, postula-se que o organismo não é X.

A primeira abordagem é implementada em geral através de algum sistema baseado em *similaridade*, e a segunda abordagem através da extração e análise de características intrínsecas às seqüências. As duas abordagens apresentam vantagens e desvantagens que serão discutidas adiante.

Embora essas abordagens sejam genéricas para a determinação da origem de seqüências de DNA, ao longo deste capítulo e no restante do trabalho utiliza-se o problema da contaminação em projetos de seqüenciamento para testes e discussão.

3.1.1 Abordagem baseada em similaridade

A procura de contaminações através de similaridade é bastante direta. É dado um conjunto de *seqüências de origem desconhecida* denominado *SOD*, onde cada seqüência pode ou não ser originária de um organismo alvo. É também necessário um banco contendo diversas seqüências de um grande número de organismos, denominado *conjunto de seqüências de origem conhecida* ou simplesmente *SOC*. Cada seqüência sod_i é comparada com cada seqüência soc_j através de alguma ferramenta de busca baseada em similaridade, como o BLAST (*basic local alignment search tool* [3]).

Para cada sod_i essa busca retorna dois resultados: ou sod_i tem ou não tem uma seqüência suficientemente similar no banco *SOC*. Sobre as seqüências sod_i sem similares (também chamadas de *sem similaridade*[1]), essa abordagem não irá concluir nada sobre seu organismo de origem, e elas permanecerão sem classificação. Já as seqüências sod_i com similaridade são analisadas com maior cuidado. A similaridade pode ter ocorrido por um dos motivos abaixo:

(a) a seqüência soc_j similar é originária **(a.1)** ou ortóloga **(a.2)** a alguma seqüência do organismo alvo.

(b) a seqüência soc_j similar é originária **(b.1)** ou ortóloga **(b.2)** à seqüência de um organismo contaminante.

A dificuldade está em se diferenciar os casos **(a)** e **(b)**, visto que os resultados dos alinhamentos para ambos os casos podem ser indistinguíveis mesmo quando analisados por profissionais experientes. O que se faz na prática é analisar o organismo de origem de soc_j: se ele for *suficientemente distante* em termos filogenéticos do organismo alvo, e se o alinhamento entre soc_j e sod_i obedecer a certos parâmetros p, então sod_i é classificada como contaminante, ou como proveniente do organismo alvo caso contrário.

Apesar da dificuldade na aplicação dos critérios usados acima, a separação de contaminantes através da similaridade é bastante utilizada na prática, pela sua facilidade de implementação. Para isso, os parâmetros p são definidos de forma empírica, analisando-se o resultado da ferramenta de busca. Por exemplo, no projeto EST de *Schistosoma mansoni* [29], quaisquer seqüências do projeto apresentando similaridade e percentual de identidade de pelo menos 98% em uma região alinhada de ao menos 75nt (que são os parâmetros p), com qualquer seqüência de *SOC* de gênero diferente de *Schistosoma*, eram consideradas contaminações. A Tabela 3.1 apresenta exemplos encontrados no projeto EST de *Schistosoma mansoni* que ilustram a dificuldade de se diferenciar os casos **(a)** e **(b)** mencionados anteriormente.

[1] seqüências sem similaridade suficiente são denominadas em inglês de "no matches"

Tabela 3.1: Exemplos reais de seqüências do projeto EST de *Schistosoma mansoni* ilustrando os casos **(a)** e **(b)** possíveis quando há similaridade entre uma seqüência de origem desconhecida *sod* do projeto com uma de origem conhecida *soc*. A coluna *sod* representa seqüências do projeto, *soc* a seqüência mais similar encontrada em bancos públicos, e o *%id* é o percentual de identidade na região alinhada.

sod	*soc* mais similar	%id	observação
C600796.1	cistatina de *Schistosoma mansoni*	100	caso **(a.1)**: *sod* é legítima, pois *soc* tem *S.mansoni* como origem
C603209.1	proteína de ligação de RNA de *Schistosoma japonicum*	98	caso **(a.2)**: *sod* é legítima, sendo *soc* ortóloga a ela, com organismo de origem de mesmo gênero (*Schistosoma*)
C716619.1	possível proteína de ligação de GTP de *Pseudomonas fluorescens* (bactéria)	100	caso **(b.1)**: *sod* é contaminação por *P.fluorescens*, correspondendo à mesma *soc*
C718951.1	proteína regulatória de *Staphylococcus aureus* (bactéria)	58	caso **(b.2)**: *sod* é contaminação por organismo desconhecido, sendo *soc* ortóloga a ela

Sobre essa abordagem, diversas outras questões podem ser levantadas: os parâmetros p escolhidos são mesmo os melhores ? Existe uma maneira automática para definí-los ? Qual o percentual esperado de seqüências do organismo alvo classificadas como contaminação e vice-versa ?

Ainda no projeto *Schistosoma mansoni*, as seqüências sem similaridade foram mantidas e consideradas legítimas, embora não houvesse nenhuma evidência para serem ou não classificadas como tal. Esse fato mostra um outro problema na abordagem de busca por similaridade: os resultados são bastante dependentes do conjunto *SOC* utilizado. Conforme descrito, se a seqüência do contaminante não estiver neste conjunto, ele será identificado apenas se houver nele outra seqüência ortóloga similar de um outro organismo, que ainda assim poderia ser ainda confundida com um ortólogo legítimo.

3.1.2 Abordagem baseada em informações intrínsecas

A detecção de contaminações através da análise de características intrínsecas é menos direta que a abordagem da similaridade. Usando-se um sistema de aprendizado supervisionado, devem ser definidos inicialmente quais organismos poderiam ter originado as seqüências de origem incerta. Para cada organismo deve ser fornecido um conjunto de

seqüências rotuladas. Usando-se um ou mais extratores de características, serão obtidos os vetores de características para cada seqüência desses conjuntos. Os vetores de características são então passados a um classificador, que irá determinar como separar as seqüências em dois conjuntos: aquelas originadas do organismo alvo e aquelas originadas de um dos organismos contaminantes utilizados. A detecção de contaminantes não previstos irá acontecer caso as seqüências destes contaminantes tenham vetores de características que se assemelhem mais aos vetores de um contaminante do que do organismo alvo.

Definidos os parâmetros de categorização sobre o conjunto de seqüências rotuladas, o classificador irá utilizá-los agora para classificar as seqüências de origem incerta, fornecendo idealmente um valor de confiabilidade sobre a predição feita. O erro associado a esta classificação é estimado como o mesmo obtido na classificação do conjunto de seqüências rotuladas. Essa estimativa de erros não existe na metodologia baseada em similaridade como a descrita na seção anterior, representando assim um importante ponto a favor da metodologia baseada em características intrínsecas. A Tabela 3.2 resume alguns dos principais pontos a favor e contra as duas metodologias apresentadas.

3.2 Trabalhos anteriores baseados em características intrínsecas

A busca por padrões em seqüências de DNA é bastante empregada para diversos fins, como por exemplo na descrição da estrutura de genes, na sua classificação funcional e em várias outras aplicações. Apesar disto, o seu uso para a detecção de contaminações tem sido pouco explorado. Os únicos trabalhos disponíveis são baseados na análise de hexâmeros (subcadeias de 6 nucleotídeos), que embora tenham apresentado resultados razoáveis, ainda poderiam ser mais genéricos e fazer uso de outras características complementares àquela dada pelos hexâmeros. Estes trabalhos estão descritos a seguir.

3.2.1 Método para controle de qualidade baseado em hexâmeros

O programa HBQCM (método para controle de qualidade baseado em hexâmeros[2]) foi desenvolvido em 1993 por White et al. [30], com a finalidade de identificar seqüências contaminantes em projetos EST. O interesse principal era o de se empregar outra metodologia conjuntamente com a da similaridade para a busca de contaminações, visto que naquela época a quantidade de seqüências depositadas em bancos públicos (e portanto disponíveis para buscas por similaridade) era algumas ordens de grandeza menor que a

[2]em inglês: *hexamer based quality control method*

Tabela 3.2: Resumo das vantagens e desvantagens das abordagens para detecção de contaminações baseadas em similaridade e em características intrínsecas. As características positivas são aquelas marcadas com ↑, e aquelas marcadas com "*" referem-se somente ao estado atual das implementações das abordagens citadas.

característica	similaridade	carac. intrínsecas
valor de confiabilidade na predição de cada seqüência	não	sim ↑
estimativas gerais de erros sobre as predições*	não	sim ↑
detecção de novos contaminantes	às vezes	às vezes
classificação de todas as seqüências apresentadas	não	sim ↑
dependência de treinamento	não ↑	sim
combinação de múltiplas evidências para tomada de decisão	não	sim ↑
necessidade de grandes bancos de seqüências	sim	não ↑
dificuldade na implementação*	não ↑	sim
sofre influência por diferenças na proporção entre seqüências legítimas e contaminantes	não ↑	sim
requer seqüências do organismo alvo	não ↑	sim

disponível hoje em dia (aproximadamente 140 mil seqüências em 1993 contra 31 milhões em 2003, dados do genbank [6]).

O HBQCM requer três conjuntos: um conjunto X com as seqüências a serem classificadas, um conjunto A com as contagens de hexâmeros feitas em seqüências de um organismo filogeneticamente próximo ao organismo alvo, e um conjunto B com contagens de hexâmeros feitas em seqüências de um organismo distante filogeneticamente do organismo alvo. A distribuição de hexâmeros de cada seqüência x é comparada com aquela dos conjuntos A e B. Dessa comparação resulta um valor, que pode ser considerado como uma característica; quanto menor o valor desta característica, menor a probabilidade da seqüência testada ser originária do organismo alvo. Uma descrição mais detalhada do método é dada na Seção 4.2.5.

Através desse programa foi possível identificar algumas prováveis contaminações por

levedura e bactérias em seqüências supostamente de origem humana. Os mesmos resultados foram obtidos também através da similaridade, o que confirma de certa forma os resultados da metodologia usada pelo programa.

Embora o programa tenha apresentado resultados promissores, várias características importantes não estão presentes nele, como por exemplo o uso de vários contaminantes simultaneamente e uso de vários métodos para extração de características.

3.2.2 Separação de ESTs de simbiontes

Em 2001, Hraber et al. [14] utilizou o programa HBQCM para uma finalidade um pouco diferente da inicialmente proposta: diferenciar seqüências obtidas de organismos vivendo em simbiose, especificamente diferenciando seqüências de plantas das de microorganismos, como fungos e bactérias.

O problema de se diferenciar as ESTs de simbiontes é muito importante. No trabalho citado, foram observados transcritos de fungos ou bactérias em amostras vegetais, em percentuais que variaram entre 0 e 75%. O uso de amostras vegetais desafiadas (ou seja, colocadas em contato) contra algum outro organismo é bastante comum, pois é importante estudar a existência de transcritos expressos especificamente em resposta à presença desse organismo, como é o caso do trabalho de Hraber.

As seqüências utilizadas passaram por um controle de qualidade antes de serem analisadas. Foram usadas apenas aquelas obedecendo aos critérios abaixo:

- Possuir tamanho mínimo de 300nt.
- Não conter nenhuma seqüência de rRNAs, tRNAs, mitocôndrias ou cloroplastos.
- Subcadeias de mais de 13 `A`s ou `T`s deveriam ser ignoradas nas análises.

Além do controle de qualidade das seqüências testadas, o trabalho utilizou também de forma mais intensiva estimativas de erros para as classificações feitas. Para isso, subcadeias de seqüências foram obtidas aleatoriamente dos conjuntos que originaram A e B (ver Seção 3.2.1) formando o conjunto X'. Seguiu-se a execução do HBQCM, tendo como entrada os conjuntos A, B e X'. Dado que se sabia a origem das seqüências em X', estimou-se o erro previsto para classificar outras seqüências de origem desconhecida. O erro é dado por dois parâmetros α e β, respectivamente a taxa de seqüências de microorganismos classificados como plantas, e a taxa de seqüências de planta classificadas como microorganismos. Essa maneira de apresentação dos erros obtidos é bastante semelhante à utilizada neste trabalho.

Em geral, a aplicação do programa HBQCM permitiu uma separação razoável entre os transcritos dos organismos em questão. Apenas em um caso específico, ao se utilizar

seqüências de ESTs obtidas de uma planta não desafiada, várias seqüências de planta se mostraram mais semelhantes a seqüências de fungo, quando na verdade nenhuma seqüência a não ser de planta seria esperada. Isso pode ter sido causado por limitações na separação provida pelos hexâmeros, sinalizando a necessidade do uso de outras características para auxiliar a classificação.

Embora o trabalho tenha apresentado bons resultados, algumas características desejáveis não foram introduzidas, tais como o uso de vários contaminantes simultaneamente e o uso de diferentes métodos para a extração de características.

3.3 Escolha da metodologia

Conforme apresentado, as metodologias baseadas em similaridade e na busca de padrões apresentam vantagens e desvantagens. Neste trabalho optou-se pelo trabalho com a busca de padrões por três razões principais:

- Ela provê estimativas de erros para as classificações feitas.
- As classificações são baseadas em evidências dadas por diferentes métodos.
- Todos os problemas já discutidos da metodologia baseada em similaridade.

No entanto, a situação ideal seria a combinação das duas metodologias para tentar retirar o melhor de ambas. Embora isso não tenha sido feito neste trabalho, esta combinação feita de forma adequada traria melhorias significativas na classificação, merecendo estudos adicionais. A combinação de metodologias baseadas em similaridade e na busca de padrões já foi usada com resultados satisfatórios no auxílio à procura de genes em organismos eucariotos [2], resultados que não foram atingidos pelas metodologias usadas isoladamente.

Capítulo 4

Descrição do programa QUEST

Para lidar com o problema das contaminações através de um sistema de aprendizado supervisionado, foi desenvolvida uma ferramenta denominada QUEST, abreviação de *Query EST* (*consulta EST*). Uma visão geral desta ferramenta é dada na Figura 4.1. Seus detalhes de implementação estão na Seção 5.3, e a metodologia empregada será vista ao longo deste capítulo.

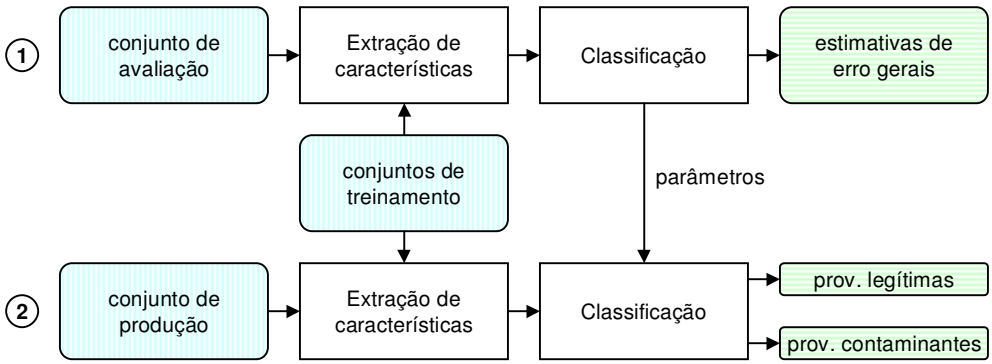

Figura 4.1: Diagrama resumido do funcionamento do QUEST, dividido em duas etapas. Na primeira etapa, um conjunto de seqüências rotuladas de avaliação tem suas características extraídas com o auxílio de conjuntos de treinamento do organismo alvo e dos contaminantes. As características são analisadas por um classificador, que utilizará as informações obtidas para a definição de parâmetros. Na segunda etapa, esses parâmetros serão usados na categorização do conjunto de produção, originando um conjunto com seqüências provavelmente legítimas e outro com prováveis contaminantes. A estimativa de erros vem da primeira etapa.

4.1 Configurações iniciais

O programa QUEST trabalha com *projetos*. Por projeto entende-se a tentativa de separar as ESTs de um organismo alvo de quaisquer contaminantes de outros organismos possivelmente presentes. Podem existir projetos para teste, onde todas as seqüências empregadas têm origem conhecida, e ainda projetos reais, onde um conjunto de seqüências de origem incerta adicional é fornecido.

O primeiro passo para a criação de um projeto é a definição do organismo alvo e dos contaminantes a serem utilizados. O usuário deve fornecer seqüências para cada um desses organismos, que serão usadas pelo programa na definição de seus três conjuntos de trabalho:

Treinamento Deve haver um conjunto de treinamento para cada organismo utilizado (alvo ou contaminante), contendo seqüências rotuladas. Eles são usados para treinamento dos extratores de características dos conjuntos de avaliação e produção descritos a seguir. Conforme será visto na Seção 5.1.3, o estudo de diferentes tamanhos de conjuntos de treinamento sugere uma quantidade mínima de seqüências para a obtenção de taxas de erro menores.

Avaliação Este conjunto contém seqüências rotuladas dos organismos do projeto, e é usado para a obtenção de parâmetros para os classificadores e para a estimativa dos erros na classificação.

Produção Dependendo do tipo do projeto, este conjunto contém seqüências rotuladas (projetos de teste) ou não rotuladas (projetos reais). Suas seqüências são classificadas com os parâmetros obtidos pelo conjunto de avaliação.

Os conjuntos descritos acima devem ser idealmente disjuntos, isto é, não devem ter nenhuma seqüência em comum para que não haja interferência nas estimativas de erros. Esses conjuntos, o organismo alvo, os potenciais contaminantes e quaisquer outras informações relativas ao projeto devem estar presentes no *arquivo de configurações de projeto*, descrito no apêndice A.1.

As seqüências *rotuladas* dos conjuntos utilizados devem obedecer a determinados critérios:

(1) possuir alta qualidade, isto é, confiança de que suas seqüências não apresentam erros ou apresentam poucos erros. Esses erros podem fazer com que elas não representem

fielmente os nucleotídeos presentes no DNA do organismo de origem. Isso é importante porque os extratores de características dependem basicamente do conteúdo das seqüências para um funcionamento adequado.

(2) elas devem representar apenas a região codificadora (CDS) parcial ou completa de algum gene do seu organismo de origem, sem regiões não traduzidas (UTR). Isso porque as regiões CDS têm composição significativamente diferentes daquela encontrada em regiões UTR [20]. Caso ambas as regiões fossem usadas, os modelos construídos seriam afetados, com conseqüente prejuízo para o desempenho da ferramenta.

(3) ter tamanho entre 200 e 2.000 nucleotídeos, para maior uniformidade nas análises, e diminuição do efeito que o tamanho pode ter sobre os extratores de características.

Similarmente, é esperado que as seqüências *não rotuladas* também obedeçam a certos critérios:

(1) possuir alta qualidade, pelo mesmo motivo citado anteriormente. Isso pode ser obtido pela remoção das extremidades com baixa qualidade, ou ainda pela montagem das seqüências empregadas através de alguma ferramenta de montagem.

(2) não conter quaisquer seqüências de RNAs ribossomais (rRNAs). Os rRNAs são bastante conservados entre as diferentes espécies, podendo atrapalhar os extratores de características. Essas seqüências podem ser encontradas e removidas através de alguma ferramenta de busca por similaridade, como o BLAST.

(3) não conter quaisquer seqüências de organelas, como mitocôndrias e cloroplastos, pois sua composição intrínseca pode variar bastante com relação a das seqüências provenientes do núcleo [14], que são em geral o interesse nos projetos EST.

(4) ter sua cauda poli-A/T, vetor e primers removidos previamente, uma vez que são trechos de seqüências não existentes no genoma do organismo alvo, e por isso podem causar a geração de modelos inadequados.

(5) ter tamanho entre 200 e 2.000 nucleotídeos, pelo motivo citado anteriormente.

4.2 Programas extratores de características

Para a extração das características necessárias à classificação das seqüências, são empregados diversos programas, referenciados neste trabalho simplesmente como *extratores*. Esses programas não foram necessariamente desenvolvidos com o propósito de detectar

contaminações. No entanto, todos devem receber como entrada um conjunto de seqüências de treinamento e um conjunto de seqüências a ser avaliado, e retornar para cada uma de suas seqüências o valor de uma característica. Quanto maior o valor de uma característica para uma seqüência, maior deve ser a probabilidade desta seqüência ter sido originada do mesmo organismo utilizado no treinamento do extrator. Os extratores utilizados foram escolhidos pensando-se na obtenção de características com certo grau de complementaridade, que combinadas poderiam atingir erros menores na classificação das seqüências.

Foram utilizados tanto programas desenvolvidos por terceiros como implementações próprias. Os programas desenvolvidos por terceiros foram o ESTScan (ES) e o Glimmer (GL). Aqueles implementados foram denominados de Distribuição binomial do conteúdo GC (BN), Distribuição multinomial com di (DN) e trinucleotídeos (TN), Distribuição de hexâmeros (HN) e Assinatura de dinucleotídeos (DS). Detalhes dos programas são dados a seguir.

4.2.1 ESTScan (ES)

O ESTScan [16] foi desenvolvido com o objetivo de traduzir seqüências de ESTs, identificando a região codificadora e efetuando eventuais correções no quadro de leitura das mesmas. Essas correções são necessárias, pois as ESTs contêm freqüentemente erros de seqüenciamento, que podem remover ou inserir bases em suas seqüências, fazendo com que o quadro de leitura seja modificado e se tenha uma tradução incorreta.

O programa deve ser treinado com seqüências do mesmo ou de um organismo próximo a aquele que deu origem às seqüências a serem trabalhadas. Dado este treinamento, o programa constrói um modelo de Markov para estados ocultos[1], contendo informações sobre qual é o padrão das seqüências transcritas do organismo. Cada seqüência a ser checada é comparada com o modelo construído, sendo retornada a seqüência com o quadro de leitura corrigido (se necessário), e uma pontuação que reflete o grau de adequação da seqüência ao modelo usado. Mesmo se houver necessidade da correção da seqüência (situação diferente da esperada neste trabalho), o ESTScan consegue em geral determinar o quadro de leitura correto.

Neste trabalho, o ESTScan foi treinado com cada conjunto de treinamento de cada organismo do projeto, e então executado sobre as seqüências de avaliação e produção. A pontuação retornada pelo ESTScan foi usada como o valor da característica retornado pelo programa para cada seqüência. Eventualmente, o ESTScan retorna pontuações com grandes valores negativos (pe, -2.147.483.648), indicando que a seqüência difere bastante do modelo empregado. Nesses casos, foi atribuído o valor zero para a característica

[1] abreviação em inglês: *HMM*

retornada.

4.2.2 Glimmer (GL)

O Glimmer [9] é um programa desenvolvido para a busca de genes em genomas de organismos procariotos. Para isso, ele procura inicialmente os genes longos (o padrão de tamanho são 500nt) presentes no genoma, visto que estes genes longos têm grande possibilidade de serem genes verdadeiros. Com esses genes longos é construído um modelo de Markov para estados ocultos com o intuito de obter o padrão de composição dos genes do organismo.

Obtido este modelo, o genoma do organismo é percorrido novamente, e o padrão de composição de cada gene encontrado é comparado com ele. Essa comparação resulta em uma pontuação, proporcional à probabilidade do gene ser mesmo um gene real. Os genes que obtiveram uma pontuação acima de um limite fornecido são então reportados.

Para uso neste trabalho, o Glimmer foi treinado separadamente com cada um dos conjuntos de treinamento dos organismos do projeto. As seqüências verificadas precisaram ser modificadas para que tivessem alguns "sinais" reconhecidos pelo Glimmer como marcadores de inicio e fim de genes.

Para cada seqüência r a ser analisada foram inseridos em seu início os prefixos $p1$ e $p2$ abaixo

$$\begin{aligned} p1 &= \texttt{ATGNATGNATG} \\ p2 &= \texttt{TCANTCANTCA} \end{aligned}$$

Isso foi feito para gerar codons de início (`ATG`) e codons de parada (`TGA`, que complementado e revertido dá `TCA`) em todos os quadros de leitura nas fitas senso e antisenso respectivamente, uma vez que não se conhece o quadro correto de r. Da mesma maneira, foram inseridos no final de r os sufixos

$$\begin{aligned} s1 &= \texttt{TGANTGANTGA} \\ s2 &= \texttt{CATNCATNCAT} \end{aligned}$$

a fim de gerar os mesmos codons de parada e início. A seqüência analisada é a concatenação de $p1$, $p2$, r, $s1$ e $s2$, de forma a simular um gene de um organismo procarioto. Em testes sem os prefixos e sufixos mencionados, o Glimmer não conseguiu terminar a análise, apresentando um erro. O parâmetro `-l` mostrou-se importante: ele evita que o Glimmer considere a seqüência como circular, fato que é comum nos organismos procariotos comumente analisados por ele.

A característica retornada pelo Glimmer é escolhida como a pontuação da melhor predição de gene encontrada nos 6 quadros de leitura analisados pelo programa. No entanto, em certos casos o Glimmer não obteve nenhuma predição, sendo então retornada a característica de valor zero.

4.2.3 Distribuição binomial do conteúdo GC (BN)

O conteúdo GC de uma seqüência é definido pela razão de seus Gs e Cs pelo número total de nucleotídeos. Essa razão é relativamente conservada mesmo em subcadeias do genoma de um mesmo organismo, mas pode variar substancialmente entre os diversos organismos existentes, como discutido em [21]. Este fato pode ser usado para distinguir a origem de seqüências, quando os organismos potenciais possuem conteúdos GC bastante diferentes.

No entanto, é comum o fato dos organismos possuirem conteúdos GC bastante próximos. Neste caso, devido a esta proximidade, a distinção da origem das seqüências pode ser feita, mas com taxas de erro inaceitáveis na prática. O conteúdo GC pode, portanto, ser utilizado apenas em determinadas circunstâncias.

Neste trabalho, os extratores devem receber necessariamente como entrada um conjunto de treinamento e um conjunto a ser avaliado. Assim, um extrator para analisar o conteúdo GC de seqüências deve ter como modelo o conteúdo GC do organismo de treinamento, e retornar um valor dado em função deste modelo e de cada seqüência avaliada. Isso é feito tendo como base a fórmula da distribuição binomial, como visto a seguir.

Para descrever o modelo, pode-se partir de um exemplo mais simples, como uma série consecutiva de n arremessos de um dado. Sabendo-se que a probabilidade p de se tirar um 6 é de 1/6 em qualquer um dos arremessos, pergunta-se: qual a probabilidade $P(k)$ de se tirar exatamente k números 6 em n arremessos, independente da ordem em que eles apareçam ? Essa probabilidade é fornecida pela fórmula da distribuição binomial [12], dada por

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \tag{4.1}$$

Analogamente para seqüências de DNA, deseja-se saber qual a probabilidade de existir um determinado número k de Gs e Cs em uma sequência de tamanho n, sabendo-se que a chance de existir um G ou C é de p (freqüência de Gs e Cs obtida através das seqüências do organismo de treino). Embora a Equação 4.1 suponha independência entre os k eventos, as ocorrências de Gs e Cs não são eventos independentes; ainda assim, a equação fornece uma aproximação razoável para o problema.

Devido aos fatoriais e às exponenciações necessárias, é freqüente a ocorrência de problemas computacionais de *underflow* e *overflow* quando se tenta calcular o valor de $P(k)$ diretamente, e por esse motivo é utilizado $\log P(k)$. Quanto maior o valor de $\log P(k)$, menor a probabilidade da seqüência checada ter sido originada do organismo de treino. Para atender às exigências descritas na Seção 4.2, o valor da característica utilizado de fato foi $-\log(P(k))$.

Para possibilitar a comparação entre os valores de características retornados por seqüências de diferentes tamanhos, foi necessário normalizar previamente os seus tamanhos,

através da normalização da contagem de seus `G`s e `C`s. O valor do tamanho padrão foi escolhido arbitrariamente como 500. Por exemplo, uma seqüência de tamanho 250 com 40 `G`s e 35 `C`s teria a contagem de `G`s mais `C`s modificada de 75 para $75 \times 500/250 = 150$.

4.2.4 Distribuição multinomial com di e trinucleotídeos (DN e TN)

Dada uma seqüência de entrada s, $|s| = n$, ela conterá exatamente $n - k + 1$ subcadeias de tamanho k, onde $1 \leq k \leq n$. Para um alfabeto de 4 letras como é o caso do DNA, o número de diferentes cadeias de tamanho k é dado por 4^k.

Nesta abordagem o objetivo é contar os números de ocorrências de subcadeias de tamanho k na seqüência de entrada e compará-los com as freqüências esperadas. As subcadeias são contadas na fita senso e anti-senso, uma vez que não se sabe *a priori* qual das duas é a correta. As freqüências esperadas vêm do que se observa nas seqüências pertencentes ao organismo de treinamento, onde as subcadeias também são contadas nas duas fitas. Dessa comparação resulta um número que reflete qual a probabilidade da seqüência checada ter sido originada ou não do organismo de treinamento.

A fórmula da distribuição multinomial provê esse tipo de resultado, visto que as freqüências das diversas subcadeias de tamanho k em s *não* são eventos independentes. Os resultados para dinucleotídeos e trinucleotídeos são obtidos quando $k = 2$ e $k = 3$, respectivamente. Na discussão abaixo considera-se subcadeias w de um certo tamanho k dado.

Para obter os valores esperados, são necessárias seqüências de treino. Para obter os valores observados precisa-se de uma seqüência s de entrada. Define-se então:

T = o conjunto de freqüências nas seqüências de treino
t_i = um elemento de T

Deve-se observar que $|T| = 4^k$ e que $0 \leq t_i \leq 1$. Analogamente são definidos:

C = o conjunto de números de ocorrências das diferentes subcadeias em s
c_i = um elemento de C

Observa-se que $|C| = 4^k$ e que $0 \leq c_i \leq 2(n-k+1)$. Definidos T e C, pode-se aplicar a fórmula da distribuição multinomial [12]:

$$P(C) \quad = \quad \frac{(2(n-k+1))!}{\prod_i c_i!} \prod_i (t_i)^{c_i} \tag{4.2}$$

onde $1 \leq i \leq 4^k$. A interpretação de $P(C)$ é a seguinte: é a probabilidade do conjunto C de ocorrências para uma dada seqüência s dadas as freqüências em T. Na prática, calcula-se o valor de $\log P(C)$, devido às exponenciações e fatoriais empregados. De forma similar ao método BN, utiliza-se $-\log P(C)$ como valor da característica retornado para o método.

Como o tamanho das seqüências s analisadas varia, o seu número de subcadeias w foi fixado arbitrariamente em 500, e as contagens no conjunto C normalizadas de acordo. Para um exemplo onde $k = 2$, se a contagem do dinucleotídeo `AA` fosse 50 em uma seqüência de tamanho 251 (portanto contendo 250 dinucleotídeos), o valor 50 seria normalizado para $(50 \times 500/250) = 100$, e de maneira similar para o restante das contagens.

4.2.5 Distribuição de hexâmeros (HN)

Conforme descrito na Seção 3.2.1, o programa HBQCM foi desenvolvido com o objetivo de identificar contaminações em projetos de seqüenciamento de DNA, através da análise da composição de hexâmeros encontradas em diferentes organismos. Neste programa, o usuário deve fornecer três conjuntos de seqüências:

1. conjunto composto por seqüências de treinamento do organismo alvo ou de outro filogeneticamente próximo; seus hexâmeros formam o conjunto A.

2. conjunto composto por seqüências de treinamento de um organismo de controle, filogeneticamente distante do organismo alvo; seus hexâmeros formam o conjunto B.

3. conjunto composto pelas seqüências a serem checadas. O conjunto de hexâmeros de cada seqüência individualmente é denominado de X.

Sendo W o conjunto de todos os possíveis hexâmeros (e portanto $|W| = 4^6$), e w_i um elemento de W, a_i, b_i e x_i denotam as quantidades de hexâmeros w_i em A, B e X, respectivamente. Cada seqüência a ser checada é comparada com os conjuntos A e B através de uma função de teste definida por

$$Test(A, B, X) \quad = \quad D(A, X) - D(B, X).$$

Caso o valor de $Test(A, B, X)$ seja positivo, a composição de hexâmeros do conjunto X será mais semelhante àquela de A que a de B, e portanto haverá maior probabilidade de X ter sido originado pelo organismo alvo do que pelo organismo de controle. A função $D(A, X)$ (e similarmente $D(B, X)$) é dada por

$$D(A,X) \;=\; -2\times \log \lambda(A,X),$$

onde a função $\lambda(A,X)$ é a razão da probabilidade de se encontrar um certo hexâmero em X pela probabilidade de encontrá-lo em A. O valor de $-2 \times \log \lambda(A,X)$ é dado por

$$-2\times \log \lambda(A,X) \;=\; 2 \times [\log L(A,A) - \log L(AX,A) + \log L(X,X) - \log L(AX,X)].$$

Para cada subcadeia w_i, a soma de a_i com x_i é atribuída a $(ax)_i$, sendo o conjunto AX formado pelos elementos $(ax)_i$. A função $\log L$ é uma razão de verossimilhança, dada por

$$\log L(P,Q) \;=\; \sum_i \left[q_i \log \left(\frac{p_i}{\sum_j p_j} \right) \right], \tag{4.3}$$

com $1 \leq i,j \leq 4^6$ e onde P e Q são quaisquer conjuntos possuindo contagens de hexâmeros. A Equação 4.3 é sensível à quantidade de hexâmeros presentes em P e Q. Por isso, na implementação original do HBQCM as seqüências com mais de 400 nucleotídeos foram reduzidas aleatoriamente para uma seqüência de tamanho 300. Testes feitos pelos autores do programa indicaram que a orientação (senso ou antisenso) das seqüências utilizadas não tem influência sobre os resultados finais obtidos.

A classificação final de cada seqüência é dada por $Test(A,B,X)$: se o seu valor for positivo, a seqüência é classificada como proveniente do organismo alvo, e se o valor for negativo, ela é classificada como contaminação.

Para este trabalho, o HBQCM foi reimplementado, e o valor da característica retornado é simplesmente aquele dado por $Test(A,B,X)$. Para maior uniformidade nas análises, o programa foi modificado de forma a que as contagens de hexâmeros fossem normalizadas, somando 500 para cada um dos conjuntos A, B e X. Por exemplo, se o hexâmero `GATACA` ocorresse 70 vezes no conjunto A, e esse conjunto tivesse um total de 1.000 hexâmeros, o valor de 70 seria normalizado para $(70 \times 500/1.000) = 35$, e de maneira análoga para o restante das contagens de todos os conjuntos.

4.2.6 Assinatura de dinucleotídeos (DS)

A assinatura de dinucleotídeos é uma metodologia desenvolvida por Karlin [17] que tem por objetivo detectar regiões com composição anômala de dinucleotídeos em um genoma, quando comparadas com um modelo de assinatura genômica pré-calculado. Regiões com

esse tipo de composição anômala poderiam indicar ilhas de patogenicidade[2] e eventos de transferência lateral gênica.

A metodologia inicialmente analisa todos os dinucleotídeos de uma seqüência genômica concatenada com seu complemento reverso pela fórmula

$$\rho^*_{XY} \quad = \quad f^*_{XY}/f^*_X f^*_Y, \tag{4.4}$$

onde f^*_X é a freqüência do nucleotídeo X, e f^*_{XY} a freqüência do dinucleotídeo XY. Assim, a cada dinucleotídeo XY está associado um valor de ρ^*_{XY}, que avalia a freqüência observada de certo dinucleotídeo com aquilo que seria esperado pela associação aleatória de seus nucleotídeos componentes. O conjunto dos diferentes ρ^*_{XY} forma o modelo G para o organismo. Este modelo G, segundo o autor, é relativamente bem conservado e característico do genoma de cada organismo, mesmo para seqüências (com $\geq$ 50kb) em diferentes regiões de seus cromossomos.

Para procurar trechos do genoma estudado com composição diferente daquela dada pelo modelo G, monta-se primeiramente este modelo pela Equação 4.4. Percorre-se então todo o genoma utilizando janelas de tamanho n (geralmente 10kb $\leq n \leq$ 50kb), e para cada seqüência correspondente a cada janela é construído um modelo F, usando também a Eq. 4.4. Cada modelo F é comparado com o modelo G do genoma através da equação

$$\delta^*(F,G) \quad = \quad \frac{1}{16}\sum |\rho^*_{XY}(F) - \rho^*_{XY}(G)|. \tag{4.5}$$

Quanto maior o valor de $\delta^*(F,G)$, maior é a diferença entre os modelos F e G, e portanto maior a chance da subcadeia que gerou F (ou sua maior parte) ter sido adquirida de um organismo de outra espécie com modelo G' distinto.

Para uso neste trabalho, a metodologia proposta por Karlin foi implementada, tendo como entrada as seqüências do organismo de treinamento (concatenadas com seu reverso complementado) contabilizadas conjuntamente para formar o modelo G. Cada seqüência a ser checada teve o seu modelo F calculado individualmente, e comparado com o modelo G pela Eq. 4.5. O valor retornado para cada seqüência de modelo F é o mesmo de $\delta^*(F,G)$.

Deve-se observar que o tamanho médio das seqüências analisadas pela ferramenta QUEST varia entre 200 e 2.000nt, sendo portanto de 5 a 250 vezes menores que as esperadas pelo método original; esse fato influencia bastante a qualidade das características retornadas por essa adaptação da metodologia.

[2]regiões com maior densidade de genes ligados a patogenicidade

4.2.7 Nota sobre os extratores DN, TN, HN e DS

Os extratores DN, TN, HN e DS trabalham com subcadeias de nucleotídeos através de três metodologias diferentes. Sobre eles três questões poderia ser levantadas: é viável a utilização de subcadeias de tamanhos diferentes ? Por que foram usadas subcadeias de tamanho x para um método y, e não subcadeias de outros tamanhos ? Por que todas as possibilidades de tamanho não foram exploradas ?

Na metodologia que usa a fórmula da distribuição multinomial (DN e TN), poderiam ser usadas subcadeias de quaisquer tamanhos limitadas pelo tamanho das seqüências analisadas. Subcadeias de tamanho 2 (DN) foram usadas por Karlin em sua metodologia de assinatura de nucleotídeos, e subcadeias de tamanho 3 (TN) correspondem ao tamanho de um codon; assim, é esperado que subcadeias de tamanho 2 e 3 contenham informações características de seus organismos de origem.

A metodologia baseada em hexâmeros (HN) utiliza uma razão logarítmica de verossimilhança para gerar suas características, com um embasamento estatístico bastante adequado. Embora seja teoricamente possível a utilização de subcadeias de tamanho menor ou igual ao tamanho das seqüências analisadas, testes realizados pelo autor do artigo sobre hexâmeros indicam o tamanho ideal de 6 para as subcadeias empregadas no método [30]. Subcadeias maiores poderiam ser utilizadas, já que quanto maior o seu tamanho, mais específicas elas se tornam para determinado organismo. No entanto, o número de subcadeias possíveis de um certo tamanho aumenta exponencialmente conforme aumenta o tamanho das subcadeias, o que impõe uma limitação para a metodologia, tendo sido considerado o tamanho 6 como valor intermediário ideal.

Para o método da assinatura de dinucleotídeos (DS), poderiam ser usadas também subcadeias de tamanhos variados. No entanto, segundo o autor Karlin [17], experimentos bioquímicos conduzidos entre as décadas de 60 e 70 sugerem que o uso da metodologia com dinucleotídeos provê informações suficientemente características sobre o organismo estudado, sendo a melhor opção inicial.

Conforme descrito, a utilização de subcadeias de tamanhos variados é possível para os métodos descritos acima. No entanto, este trabalho tem a preocupação de manter o número de extratores de características utilizados pequeno, por duas razões: (1) o objetivo principal deste trabalho é mostrar a viabilidade da metodologia, e não explorá-la a fundo; (2) o uso de mais características provenientes dos mesmos extratores (apenas com variação de parâmetros) pode não melhorar significativamente o desempenho dos classificadores, mas irá aumentar consideravelmente o tempo gasto pela etapa de seleção de subconjunto de características, descrita mais adiante.

4.3 Treinamento e execução dos programas extratores de características

O treinamento e execução dos extratores requer um conjunto E de extratores de características, um conjunto O de organismos utilizados, e um conjunto S com seqüências a terem suas características extraídas. O conjunto S pode ser tanto o conjunto de avaliação quanto o de produção, dependendo da etapa de execução do QUEST.

Cada extrator $e \in E$ é executado, tendo como entrada o conjunto de treinamento do organismo $o \in O$, e um conjunto S. Assim, para uma mesma seqüência $s \in S$, um extrator retorna uma característica para cada organismo o usado no treinamento. Por exemplo, a característica retornada pelo extrator DN com treinamento do organismo o pode ser lida como "característica que reflete a distribuição de dinucleotídeos segundo a fórmula da distribuição multinomial com modelo do organismo o".

Conforme os extratores são executados, as características de cada seqüência s vão sendo conhecidas, formando sua matriz C de características correspondente, como representado na Tabela 4.1. Nessa matriz C, o valor da característica c_{oe} denota o valor da característica obtida ao se executar o extrator e, tendo como entrada o conjunto de treinamento do organismo o e s como seqüência analisada. A matriz C equivale ao vetor de características usado no reconhecimento de padrões descrito no capítulo de conceitos.

Tabela 4.1: Matriz C de características obtidas para uma seqüência s. Os elementos c_{oe} indicam a característica obtida pelo extrator e, treinado com o organismo o tendo como seqüência de entrada s. A primeira linha e coluna foram colocadas apenas para ilustração, e não fazem parte da matriz C.

	e_1	e_2	$\cdots$	e_m
o_1	c_{11}	c_{12}	$\cdots$	c_{1m}
o_2	c_{21}	c_{22}	$\cdots$	c_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	c_{n1}	c_{n2}	$\cdots$	c_{nm}

O valor da característica retornada por cada extrator pode variar enormemente. Por exemplo, em um dado projeto, o ESTSCAN pode retornar valores variando entre 0 e 3.000, enquanto o GLIMMER retorna por exemplo valores entre 0 e 3. Essa diferença entre os valores mínimos e máximos das características pode dificultar o trabalho dos classificadores, pois uma característica com um valor muito alto pode ter mais influência que uma outra com valor mais baixo, sem que isso represente na verdade uma maior significância da primeira característica [26]. Assim, os valores das características devem

ser normalizados de alguma maneira.

As características são normalizadas de forma linear, de modo a que fiquem dentro da faixa $[x_1, x_2]$, onde x_1 significa a menor e x_2 a maior confiança possível. Foi usado $x_1 = 0$ e $x_2 = 100$, embora qualquer outra faixa (com $x_1 < x_2$) pudesse ter sido empregada. Para essa normalização, são analisadas conjuntamente todas as características geradas por um mesmo extrator, independente do organismo usado no treinamento, para todas as seqüências dos conjuntos de avaliação e produção. São obtidos os valores mínimo min e máximo max, sendo o valor de cada característica redefinido para

$$valor' \quad = \quad (valor - min) \times 100/(max - min).$$

Após a normalização, a matriz C associada a cada seqüência será preenchida com o valor das características normalizadas, sendo então passada para os classificadores.

4.4 Classificação

Após a execução dos extratores de características, cada seqüência s a ser classificada estará associada a uma matriz C de características normalizadas. Dada essa matriz C, um classificador deverá escolher o organismo (alvo ou contaminante) que provavelmente gerou uma matriz como C. Para isso, dois tipos de classificadores foram testados: os *não-paramétricos* e os *paramétricos*. Os classificadores não-paramétricos utilizam somente os valores da matriz C, enquanto os classificadores paramétricos requerem valores adicionais, conforme descrito na Seção 4.4.3. As duas abordagens de classificação foram implementadas e seus detalhes estão descritos nas seções 4.4.2 e 4.4.3. Além de trabalhar com as características de vários extratores simultaneamente, os dois tipos de classificadores funcionam também para as características de um extrator individualmente, o que permite comparar a performance conseguida com um extrator com a performance com um conjunto de extratores.

4.4.1 Valores de confiabilidade

Os classificadores paramétricos e não paramétricos devem retornar também um *valor de confiabilidade* na predição feita para cada matriz C. Esses valores servem para a comparação entre as predições de seqüências de mesma classe (legítimas ou contaminantes) dadas para o mesmo projeto. Os valores de confiabilidade mínimos e máximos não são fixos e podem variar de classificador para classificador e também para cada projeto processado. Não há rigor estatístico na definição desses valores; eles servem apenas como referências para a comparação entre duas predições de mesma classe.

Por exemplo, se em um projeto A o QUEST atribuir a uma seqüência r um valor de confiabilidade 3 para predição como contaminante, e a outra seqüência s um valor de confiabilidade 2 também para contaminante, isso deveria significar que a predição dada para r é mais confiável que aquela dada a s. No entanto, se em um outro projeto B o QUEST atribuir um valor de 4 para uma outra seqüência t considerada contaminação, isso não implica em que essa predição seja mais confiável que aquelas dadas para as seqüências r e s do projeto A, uma vez que os valores de confiabilidade servem apenas para a comparação entre as predições de um mesmo projeto.

4.4.2 Classificadores não-paramétricos

Cada linha da matriz C contém as características obtidas para cada um dos organismos o. Assim, denomina-se L_o o conjunto formado pelas características obtidas pelo organismo o. Os elementos de cada conjunto L_o são combinados por uma certa *função consolidadora* $g(L_o)$. Quanto maior for o valor de $g(L_o)$, maior será a probabilidade da seqüência correspondente ser originária do organismo o. Foram empregadas quatro diferentes funções g de consolidação, discutidas em [27] e descritas a seguir.

Soma Dado um conjunto de características L_o, a função g definida pela *soma* é simplesmente a soma dos elementos de L_o.

Produto O *produto* é semelhante ao procedimento da soma, mas com a troca da operação de soma pela multiplicação. Esta estratégia é particularmente sensível a características com valores pequenos. Por exemplo, se um extrator retornar zero para algum organismo, o valor retornado pela estratégia para este organismo será zero, independentemente dos valores das outras características.

Pluralidade Nesta estratégia os valores de cada coluna da matriz associada C são previamente substituídos pelo seu *ranking* na coluna, conforme exemplificado na Tabela 4.2. Em seguida os conjuntos L_o são montados e é retornado para cada um o número de elementos "1"s presentes. Assim, o organismo que obteve o maior valor de característica para o maior número de extratores receberá o maior valor.

Borda Na função denominada *Borda*[3] a matriz C é modificada assim como na estratégia da *pluralidade*. É retornado o valor de $n \times m$ menos a soma dos valores de L_o, onde n e m são respectivamente o número de linhas e colunas da matriz C.

[3]método possivelmente inventado pelo engenheiro francês Jean-Charles Borda, no século XVIII

Tabela 4.2: Exemplo de matriz C com suas características trocadas pelos seus valores ordenados dentro de cada coluna. Por exemplo, o maior valor de uma coluna i terá sempre o número um na coluna i da nova matriz, o segundo maior terá o numero dois e assim por diante.

	e_1	e_2	e_3
o_1	5	3	2
o_2	3	4	5
o_3	4	2	3
o_4	0	5	6

$\Rightarrow$

e_1	e_2	e_3
1	3	4
3	2	2
2	4	3
4	1	1

Escolhida uma função g, haverá para uma seqüência s um valor consolidado para cada organismo o do projeto. Esses valores devem ser usados para decidir com qual dos organismos esta seqüência s mais se parece. Neste trabalho, o organismo escolhido é aquele que tiver o conjunto L_o de maior $g(L_o)$. Se este organismo for contaminante, a seqüência é classificada como contaminante, ou como legítima caso contrário. O valor de confiabilidade associado a cada seqüência é dado pela diferença entre o primeiro e segundo maiores $g(L_o)$. A Tabela 4.3 apresenta um exemplo de classificações dadas por cada uma das estratégias apresentadas.

Tabela 4.3: Exemplo de aplicação das diferentes estratégias de combinação mencionadas. Os valores em destaque indicam qual é o organismo escolhido para cada estratégia.

	e_1	e_2	e_3
o_1	5	3	2
o_2	3	4	5
o_3	4	2	3
o_4	0	5	6

$\Rightarrow$

soma	produto	pluralidade	Borda
10	30	1	4
12	**60**	0	5
9	24	0	3
11	0	**2**	**6**

4.4.3 Classificadores paramétricos

Cada coluna da matriz C contém as características obtidas para um extrator e treinado com cada um dos organismos o empregados. Assim, denomina-se K^e o vetor formado pelas características obtidas pelo extrator e. Por definição, o primeiro elemento k_1^e irá conter sempre a característica obtida pelo extrator e treinado com seqüências do organismo alvo, sendo o restante dos elementos de K^e obtidos com treinamentos de organismos contaminantes.

Tendo como parâmetro um vetor K^e, uma função consolidadora g irá retornar um número v, combinando os elementos de K^e. Quanto maior o valor de v, maior será a probabilidade da seqüência s correspondente ser originária do organismo alvo. Para obtenção de v, foram definidas empiricamente duas funções consolidadoras g, descritas na Tabela 4.4. Foram testadas outras funções g substituindo o termo $max(k_i^e)$ pela média aritmética ou geométrica dos elementos k_i^e, mas os resultados obtidos foram na maioria dos casos piores que aqueles dados pelas funções dmc e rmc descritas.

Tabela 4.4: Descrição das funções consolidadoras g empregadas, tendo como entrada o vetor K^e. O elemento k_1^e contém sempre a característica obtida com treinamento do organismo alvo. O índice i obedece a $2 \leq i \leq |E|$.

descrição	abreviação	definição
diferença da maior característica	dmc	$k_1^e - max(k_i^e)$
razão da maior característica	rmc	$k_1^e / max(k_i^e)$

Escolhida e aplicada uma função g aos vetores K^e, cada um deles terá sido reduzido a um número. Estes números são agrupados em um vetor M, havendo um elemento m_i para cada extrator e_i. Neste ponto, a matriz C original terá sido reduzida ao vetor M.

É necessária então a construção de um outro vetor Q, denominado *vetor divisor*. Ele será usado com todo vetor M gerado. Cada elemento q_i indica o valor mínimo para que m_i caracterize sua seqüência geradora como legítima. Abaixo deste valor q_i, a mesma será classificada como contaminante. A definição e uso desse vetor Q é responsável por essa abordagem ser denominada de *paramétrica*. Seus elementos devem ser escolhidos de forma a classificar bem (isto é, com a menor taxa de erro possível) os vetores M. Um exemplo de como um valor q_i é escolhido é dado na Tabela 4.5.

Os vetores M e Q são usados na construção do vetor R de classificações da seguinte maneira: se $m_i \geq q_i$, então r_i recebe "legítima", ou "contaminação" caso contrário. O esquema abaixo apresenta um exemplo da construção de um vetor R. Os elementos $r_i = L$ indicam predição do i-ésimo extrator como legítima, e aqueles $r_i = C$ indicam predição como contaminação.

$$\begin{array}{rcl} Q & = & [\ 10,\ 20,\ 75,\ 35\ ] \\ M & = & [\ 80,\ 46,\ 15,\ 43\ ] \\ \hline R & = & [\ L,\ L,\ C,\ L\ ] \end{array}$$

Assim, ao final do processo, cada seqüência estará associada a um vetor R contendo as predições individuais dadas pelos diferentes extratores utilizados. O problema agora é como combinar as classificações individuais dos extratores para obter a classificação final

Tabela 4.5: Exemplo de escolha de um valor q_i, para $i = 1$. Os elementos m_i provenientes do conjunto de avaliação (tabela superior) são ordenados, e q_i é escolhido varrendo esses elementos e usando o valor de m_i que melhor separa os elementos legítimos (L) dos contaminantes (C), de acordo com a função de minimização de erro descrita na Seção 4.5. O valor de q_i escolhido neste exemplo está indicado.

seqüência	classe	pos. i no vetor M				
		1	2	3	4	5
1	leg	**3**	7	1	4	10
2	leg	**11**	5	4	5	8
3	leg	**12**	19	4	9	11
4	leg	**17**	10	2	8	14
5	leg	**23**	14	13	2	15
6	cont	**2**	6	2	3	7
7	cont	**3**	4	5	6	3
8	cont	**8**	8	9	7	8
9	cont	**10**	12	3	4	2
10	cont	**20**	18	15	11	12

elementos m_1 ordenados	2	3	3	8	10	11	12	17	20	23
classes	C	C	L	C	C	L	L	L	C	L
						$\uparrow q_1$				

para cada seqüência. Para isso, o vetor de classificações R pode ser passado para duas diferentes estratégias de combinação de classificadores implementadas: uma baseada em *votação*, e outra chamada de *vetores binários*, ambas descritas a seguir. Essas estratégias devem retornar para cada seqüência representada pelo seu vetor R uma classificação (legítima ou contaminante).

Votação A idéia da votação é bastante simples: descobrir qual é o número n mínimo de extratores com predição "legítima" que cada seqüência deve ter em seu vetor R para ser classificada como tal. O valor escolhido para n é obtido fazendo-o variar entre 1 e o número de extratores, classificando a cada passo as seqüências como legítimas se seu número de predições legítimas forem maiores ou iguais a n, ou como contaminantes caso contrário. A definição do valor de n deve minimizar o erro implicado na sua escolha. Uma vez definido n sobre as seqüências de avaliação, o mesmo valor é aplicado para classificar as seqüências de produção. Um exemplo é dado na Tabela 4.6.

Além da predição como contaminante ou legítima, cada seqüência deve receber também

Tabela 4.6: Exemplo de definição do melhor valor de n em seqüências do conjunto de avaliação. A coluna *classe* indica a classe real a que a seqüência pertence. O vetor R indica a predição individual dos quatro extratores empregados, onde C e L significam contaminação e legítima, respectivamente. O valor de n é o número de elementos L em R. Neste exemplo, o valor de n escolhido seria 3, por resultar em maior número de predições corretas (4/4), e conseqüentemente menor erro.

seqüência	classe	vetor R	predição se exigido $n \geq$			
			1	2	**3**	4
1	leg	$[C, L, L, L]$	leg	leg	leg	cont
2	cont	$[C, C, L, L]$	leg	leg	cont	cont
3	leg	$[L, L, C, L]$	leg	leg	leg	cont
4	cont	$[L, C, C, C]$	leg	cont	cont	cont
predições corretas			2/4	3/4	**4/4**	2/4

um valor de confiabilidade. Para esta estratégia, o valor de confiabilidade é o número de elementos L presentes no vetor R se a seqüência for classificada como legítima, ou o número de elementos C caso contrário.

Nesta estratégia, todas as predições dos extratores têm o mesmo peso, o que torna sua implementação bastante simplificada. Também pelo uso de um peso comum a todos os extratores, é esperado que a estratégia seja prejudicada por extratores com altas taxas de erros na classificação individual. Isso pode ser contornado através da seleção apenas dos extratores mais relevantes à classificação (ver Seção 4.4.5), ou ainda através da escolha de diferentes pesos para os extratores, de acordo com o desempenho do classificador [18].

Vetores binários Esta estratégia baseia-se na observação e uso de todos os vetores R gerados pelas seqüências de avaliação. Os elementos desses vetores R podem conter apenas dois valores (L ou C, respectivamente "legítima" ou "contaminante"), e por isso os vetores R são também chamados de *vetores binários*. Como exemplificado na Tabela 4.7, alguns vetores R podem estar predominantemente associados a seqüências legítimas, e outros às contaminantes, de acordo com o conjunto de avaliação. Neste exemplo, é intuitivo classificar como contaminação qualquer seqüência do conjunto de produção que esteja associada ao vetor $R = [C, C]$. De uma maneira geral, se 60% das seqüências contaminantes presentes no conjunto de avaliação estiverem associadas ao vetor $[C, C]$, espera-se também que 60% dos contaminantes do conjunto de produção tenham a mesma associação, o mesmo valendo para o restante dos vetores.

No entanto, como se observa na Tabela 4.7, existem vetores associados tanto com seqüências legítimas quanto contaminantes, como por exemplo o vetor $R = [L, C]$. Nes-

Tabela 4.7: Exemplo de vetores R obtidos a partir do conjunto de avaliação. Os vetores R nas tabelas superiores foram contabilizados para formar a tabela inferior. Na tabela inferior, a fração 3/5 associada ao vetor $R = [C, C]$ indica que ele ocorre em três do total de cinco contaminantes presentes.

seqüência	classe	vetor R
1	cont	$[C, C]$
2	cont	$[C, C]$
3	cont	$[C, C]$
4	leg	$[C, L]$
5	cont	$[L, C]$

seqüência	classe	vetor R
6	leg	$[L, C]$
7	leg	$[L, C]$
8	cont	$[L, L]$
9	leg	$[L, L]$
10	leg	$[L, L]$

vetor R	contaminantes	legítimas	maior associação com
$[C, C]$	3/5 (60%)	0/5 (0%)	contaminantes
$[C, L]$	0/5 (0%)	1/5 (20%)	legítimas
$[L, C]$	1/5 (20%)	2/5 (40%)	legítimas
$[L, L]$	1/5 (20%)	2/5 (40%)	legítimas

sas situações escolhe-se a classe de maior percentual; assim, na classificação do conjunto de produção, as seqüências que apresentarem o vetor $[L, C]$ seriam classificadas como legítimas (já que $40\% > 20\%$). Os 20% de seqüências contaminantes classificadas incorretamente tornariam-se parte do erro deste classificador, e o valor de confiabilidade dessas seqüências seria definido como 20 (40%-20%). No caso do empate entre os percentuais, opta-se por classificar as seqüências como legítimas, uma vez que é esperado haver em geral um maior número de seqüências do organismo alvo no conjunto de produção.

As estimativas de erros gerais para este classificador são dadas simplesmente pela soma dos percentuais de classificações incorretas de cada vetor R para legítimas e contaminantes. Para a Tabela 4.7 os erros seriam $(20\% + 20\%) = 40\%$ de contaminações classificadas como legítimas e 0% de seqüências legítimas classificadas como contaminações. Essa forma de se definir os erros é tratada com maiores detalhes na Seção 4.5.

Um caso mais complicado envolve o surgimento no conjunto de produção de vetores não amostrados no conjunto de avaliação, por insuficiência de amostras neste último. Uma vez que não há conhecimento *a priori* de como usar tais vetores, as seqüências que os apresentem permanecem com classificação incerta. Além desse problema, o número de vetores R diferentes aumenta exponencialmente com o número de extratores empregados, fazendo com que o número de vetores amostrados no conjunto de produção mas não amostrados no de avaliação tenda a aumentar bastante.

Apesar dos problemas citados, a estratégia dos vetores binários apresenta vantagens

sobre a da votação. Por exemplo, dois vetores $R' = [L, L, C, C]$ e $R'' = [C, C, L, L]$ podem ter significados bastante diferentes. Estas configurações são identificadas pela estratégia dos vetores binários, mas não pela estratégia da votação.

4.4.4 Outros classificadores

Cada classificador descrito neste trabalho foi escolhido principalmente por sua simplicidade e facilidade de implementação. No entanto, existe um grande número de classificadores diferentes que poderiam ser empregados.

Como forma de averiguar o poder de classificação e generalização dos classificadores empregados, foram feitos testes com os chamados *classificadores baseados em vetores de suporte* [4], denominados aqui simplesmente de *SVM*. Em linhas gerais, uma SVM é um classificador binário (distingue membros de duas classes) que recebe como treinamento vetores de características rotulados de cada uma das classes. Com base nisso ele classifica um outro conjunto de vetores não-rotulados. A descrição a seguir sobre SVMs é sucinta; maiores detalhes podem ser obtidos em [8, 11].

Conforme explicado na Seção 2.5, um vetor de características de tamanho n define um espaço de características também de ordem n. Usando uma função não-linear, uma SVM mapeia os vetores de características recebidos em um espaço de características de ordem superior, em geral muito superior ao definido pelos vetores de características originais [11]. Usando uma função de mapeamento e uma dimensão apropriada, pode-se sempre definir um hiperplano que separa os vetores das duas classes de maneira perfeita, embora a dimensão empregada seja limitada por questões computacionais. Um hiperplano é construído de forma a fornecer boa separação entre os vetores transformados de cada classe. Para a definição deste hiperplano, são utilizados apenas os vetores de treino transformados mais relevantes de ambas as classes (chamados *vetores de suporte*), de forma a que a distância entre esses vetores e o hiperplano seja máxima. A Figura 4.2 ilustra a definição de um hiperplano separando as características transformadas de duas classes (adaptado de [11]).

Assim, usando-se uma função linear simples, os vetores não-rotulados acima do hiperplano serão classificados como pertencentes à classe 1, ou como classe 2 caso contrário. Diferentemente dos classificadores descritos neste trabalho, que procuram somente minimizar o erro no conjunto de avaliação, uma SVM preocupa-se também em minimizá-lo nos conjuntos de produção, o que resulta em um melhor poder de generalização, algumas vezes em detrimento ao erro no conjunto de avaliação.

Para testar o desempenho das SVMs no contexto deste trabalho, foi utilizada uma implementação específica denominada GIST, disponível em *http://svm.sdsc.edu*. Foram

[4] em inglês: *support vector machines* (SVM)

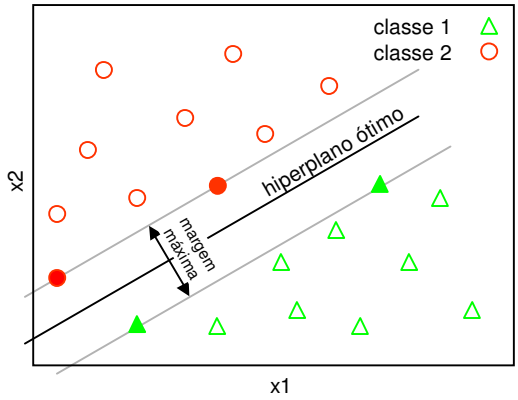

Figura 4.2: Exemplo de hiperplano gerado através de vetores de treino. O hiperplano é definido de forma a que tenha distância máxima dos vetores de suporte, representados aqui por formas cheias.

utilizados os parâmetros padrão da ferramenta, e foram passados para ela exatamente os vetores M definidos para os classificadores paramétricos na Seção 4.4.3. O valor de confiabilidade usado foi o mesmo do campo chamado `discriminant` retornado pelo próprio GIST. Os resultados da utilização da ferramenta estão descritos na Seção 5.1.

4.4.5 Seleção de subconjunto de características

Os sistemas baseados em reconhecimento de padrões são em geral beneficiados pelo uso conjunto de uma determinada quantidade de características complementares para o seu trabalho de classificação.

No entanto, muitas características podem apresentar comportamentos bastante distintos conforme se muda o objeto a partir do qual elas são extraídas. Por exemplo, certas características que ocasionam pequenas taxas de erro quando usadas para distinguir seqüências entre um organismo A e B podem ser inúteis para separar A de C.

Nestes casos, um procedimento bastante utilizado para melhorar o desempenho dos classificadores é o uso de um subconjunto das características obtidas. O objetivo é, dado um conjunto de características obtidas para o conjunto de seqüências de avaliação, descobrir qual o subconjunto de características gera a menor taxa de erro para um certo classificador. Isso é fundamental neste trabalho, pois o comportamento dos extratores pode variar enormemente de acordo com os organismos usados em cada projeto, como exemplificado na Figura 4.3.

Os subconjuntos de características são diferentes combinações entre as características disponíveis. Essas diferentes combinações podem fazer com que os classificadores funcionem de maneiras bastante distintas. Dessa maneira, o melhor subconjunto de carac-

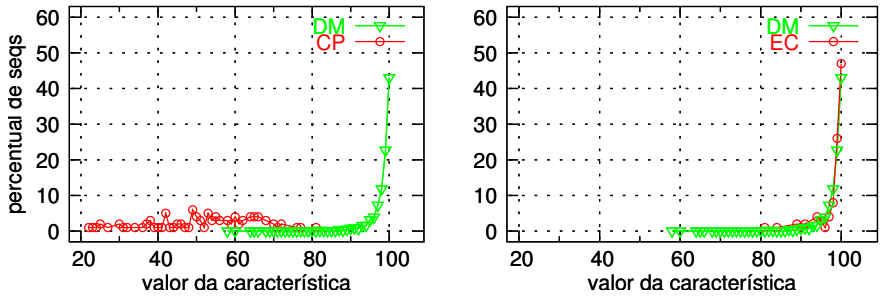

Figura 4.3: Valores das características do extrator BN para as seqüências dos organismos *Drosophila melanogaster* (DM), *Clostridium perfringens* (CP) e *Escherichia coli* (EC), com treinamento de DM. As características retornadas pelo extrator BN proporcionam melhor separação entre DM e CP que entre DM e EC.

terísticas deve ser obtido através do teste de todas as combinações possíveis entre as características para cada classificador. Porém, no caso de haver muitas características, esta busca se mostra computacionalmente muito cara, uma vez que o número de combinações a serem testadas é $(2^{|E|} - 1)$, onde $|E|$ é o número de extratores utilizados. Sistemas que empregam um grande número de características (da ordem de centenas), como problemas de reconhecimento de caracteres e voz, utilizam freqüentemente heurísticas como as descritas por Roli [24] para lidar com essa questão.

Neste trabalho é feita uma busca exaustiva dentro das combinações possíveis de extratores utilizados para cada classificador a ser testado, conforme exemplificado na tabela 4.8. Isso é viável porque o número de extratores utilizados (7) faz com que o espaço de busca tenha tamanho 127 $(2^7 - 1)$, o que torna a seleção de extratores rápida mesmo em computadores com pequeno poder de processamento.

4.5 Estimativa de erros

Este trabalho tem por objetivo classificar as seqüências dadas entre duas classes: a classe das legítimas e a classe das contaminações. No entanto, o processo de classificação nem sempre é perfeito, o que faz com que algumas seqüências legítimas acabem classificadas como contaminantes e vice-versa. Desta maneira, ao se efetuar a classificação, as seqüências estarão em uma de quatro situações distintas:

- Classificadas como legítimas, sendo de fato legítimas: tais eventos são chamados de positivos verdadeiros (*true positives*, ou simplesmente *TP*).
- Classificadas como legítimas, sendo na verdade contaminantes: tais eventos são chamados de falsos positivos (*FP*).

Tabela 4.8: Exemplo de como a seleção do subconjunto de características é feita quando há um total de três extratores. São enumeradas todas as combinações com ao menos um extrator, e é utilizada a combinação que resulta no menor erro sobre o conjunto de avaliação para um dado classificador. O conjunto selecionado neste exemplo está em destaque.

conjunto de extratores	erro associado (%)
ES	10,00
GL	25,50
BN	51,75
ES GL	12,50
ES BN	**9,00**
GL BN	21,50
ES GL BN	12,75

- Classificadas como contaminantes, sendo contaminantes: esses eventos são chamados de negativos verdadeiros (*true negatives*, ou simplesmente *TN*).
- Classificadas como contaminantes, sendo legítimas: esses eventos são chamados de falsos negativos (*FN*).

O número total de seqüências classificadas é necessariamente igual à soma de *TP*, *FP*, *TN* e *FN*. A partir deles são feitas as definições da Tabela 4.9.

A *precisão* da classificação mede o total de seqüências preditas corretamente (tanto legítimas quanto contaminações) com relação ao total de seqüências classificadas, e dá uma medida de quão bom é o classificador usado para separar as classes. No entanto, esta medida só é útil quando as seqüências de ambas as classes estão em proporções próximas à 50%, o que não é o caso na maioria das situações abordadas neste trabalho.

As taxas de falsos positivos e negativos medem separadamente a capacidade do classificador para distinguir cada uma das classes, sendo mais apropriadas para as situações encontradas neste trabalho. Essas taxas devem ser mantidas tão pequenas quanto possível. No entanto, quando se consegue diminuir uma dessas taxas, a outra freqüentemente aumenta, restando aos classificadores encontrar um compromisso entre elas que atenda às necessidades do usuário.

Em um projeto EST que não passe por nenhum tipo de verificação contra contaminações, 100% dos positivos verdadeiros estarão presentes, mas o percentual de falsos positivos (contaminações) é desconhecido. Um dos objetivos deste trabalho é manter o percentual de positivos verdadeiros o mais alto possível, ao mesmo tempo em que se

Tabela 4.9: Definições usadas para a descrição dos erros neste trabalho.

item	representação	definição
taxa de falsos positivos	tFP	$\frac{FP}{(TN+FP)}$
taxa de falsos negativos	tFN	$\frac{FN}{(TP+FN)}$
percentual de falsos positivos	pFP	$100\times tFP$
percentual de falsos negativos	pFN	$100\times tFN$
soma dos percentuais de falsos positivos e negativos	$pFPFN$	$pFP+pFN$
precisão	Ac	$\frac{100\times(TP+TN)}{(TP+FN+TN+FP)}$
percentual de positivos verdadeiros	pTP	$100-pFN$
percentual de negativos verdadeiros	pTN	$100-pFP$

mantém o percentual de falsos positivos em níveis aceitáveis, fornecendo essas estimativas de erro ao usuário. Os valores de pFN e pFP que gostaríamos de alcançar com esta metodologia são de $pFN \leq 1\%$ e $pFP \leq 2\%$.

Projetos EST em geral produzem um número de seqüências do organismo alvo muito maior do que o de organismos contaminantes. Nestas situações, para pFN e pFP iguais, o número absoluto de seqüências legítimas classificadas erroneamente é maior que o de contaminantes também com classificação incorreta, conforme ilustrado na Figura 4.4. Assim, em projetos EST, é preferível em geral manter pFN menor que pFP.

No entanto, existem também situações em que é preferível manter pFP tão pequeno quanto possível, como por exemplo quando se quer obter um conjunto bastante confiável de seqüências do organismo alvo. Dessa maneira, como meio termo entre as situações descritas anteriormente, a medida de erro que o sistema busca minimizar é o percentual de falsos positivos mais o percentual de falsos negativos ($pFPFN$). Essa é apenas uma dentre diversas formas de se medir a performance de classificadores, apresentando vantagens e desvantagens discutidas em [5]. Ela é usada neste trabalho tanto para definir o vetor divisor Q dos classificadores paramétricos (Seção 4.4.3) quanto para a escolha do subconjunto de características de cada classificador (Seção 4.4.5). A minimização de erros outros que não $pFPFN$ foram previstas mas ainda não completamente implementadas.

Neste trabalho a estimativa de erros é feita através do conjunto de avaliação. Os classificadores empregados para categorizar as seqüências desse conjunto irão procurar minimizar $pFPFN$. Esse valor é usado também como estimativa para a classificação feita sobre o conjunto de produção. É bastante desejável que este valor seja o menor possível,

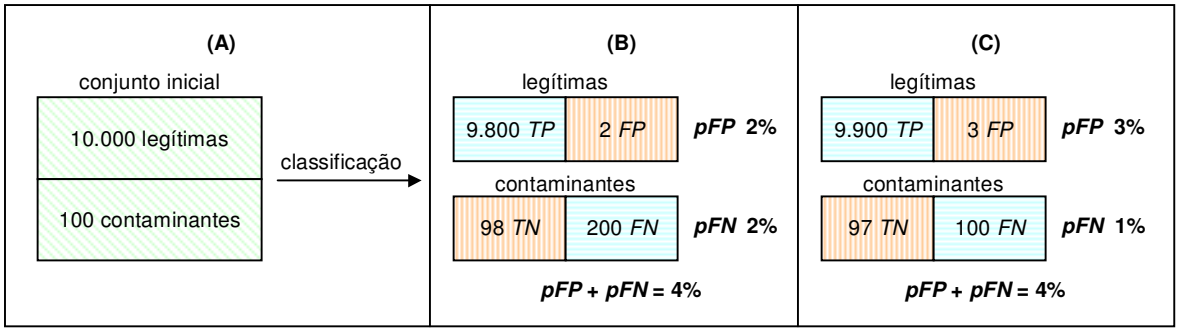

Figura 4.4: Exemplo da situação comum em projetos EST onde é preferível ter-se pFN menor que pFP. Neste exemplo, existem 100 vezes mais seqüências legítimas que contaminantes. Em (A) estão representadas as seqüências amostradas por um projeto EST fictício; a separação entre legítimas e contaminantes não é conhecida *a priori*. Na separação apresentada em (B), consegue-se um erro (pFN + pFP) total de 4%, mas 200 seqüências legítimas são classificadas erroneamente (falsos negativos). Na separação (C), mantém-se o mesmo erro total de 4%, com um falso positivo a mais e 100 falsos negativos a menos.

indicando que os classificadores conseguiram boa distinção entre as seqüências do conjunto de avaliação. No entanto, é também muito importante que o valor estimado de $pFPFN$ esteja muito próximo daquele real do conjunto de produção, indicando que o sistema "aprendeu" como distinguir novas seqüências apresentadas a ele e caracterizando portanto um bom poder de generalização. Por exemplo, pouco adianta um classificador categorizar com 100% de precisão o conjunto de avaliação, mas com apenas 60% as seqüências do conjunto de produção. Embora se espere que os classificadores utilizados apresentem boa capacidade de generalização, a única maneira de se averiguar isso é através de conjuntos de produção rotulados, como será mostrado na Seção 5.1.

4.5.1 Nota sobre *pFP*, *pFN* e *pFPFN*

É importante ressaltar que o QUEST não fornece diretamente estimativas sobre o percentual de contaminações no arquivo de prováveis seqüências legítimas, tampouco sobre o percentual de seqüências contaminantes presentes no conjunto de seqüências original. Assim, se o QUEST estimar pFP de 3%, isso significa que no conjunto de prováveis legítimas gerado estarão teoricamente 3% das contaminações presentes no conjunto de seqüências original, e não que 3% das seqüências no conjunto de prováveis legítimas são possíveis contaminações.

Como forma de estimar o percentual de contaminações no conjunto de seqüências originais e de prováveis legítimas, pode-se utilizar os próprios valores dados pelo QUEST,

bastando para isso observar algumas relações entre eles, como se segue:

$$\begin{cases} TP + FP & = & A \\ TN + FN & = & B \\ 100 \times FN/(FN + TP) & = & pFN \\ 100 \times FP/(FP + TN) & = & pFP, \end{cases}$$

onde TP, FP, TN e FN são as variáveis procuradas, A é o número de seqüências presentes no conjunto de prováveis legítimas gerado, B o número de seqüências presentes no conjunto de prováveis contaminações, e pFN e pFP são as estimativas de erros dadas pelo QUEST. A partir da resolução desse sistema, o percentual estimado de possíveis contaminações no conjunto de prováveis legítimas é dado por $FP{\times}100/A$, e o percentual estimado de contaminações presentes no conjunto de seqüências original é dado por $(TN{+}FP){\times}100/(A + B)$.

Uma outra observação sobre $pFPFN$ é que rigorosamente falando este valor não é uma porcentagem, pois pFP e pFN referem-se a razões distintas. No entanto, no que segue os valores de $pFPFN$ são apresentados acompanhados do símbolo "%".

4.6 Resumo

Para utilização do QUEST, o usuário deve fornecer um arquivo de descrição do projeto, contendo um conjunto de organismos O (com um organismo alvo e um ou mais contaminantes), os conjuntos de seqüências de treinamento, de avaliação e de produção, e um classificador dentre: *soma*, *produto*, *pluralidade*, *Borda*, *votação* ou *vetores binários*. No caso dos classificadores *votação* e *vetores binários* deve ser fornecida também a função de consolidação (dmc ou rmc). O conjunto E dos sete extratores é fixo. Cada projeto processado pelo programa QUEST gera uma saída correspondente. Um exemplo de saída está no apêndice A.2.

Um diagrama geral da ferramenta QUEST é dado na Figura 4.5. O pseudo-código da ferramenta é apresentado na Figura 4.6, onde a indentação das linhas define os blocos, e os comentários são precedidos por barras duplas (//).

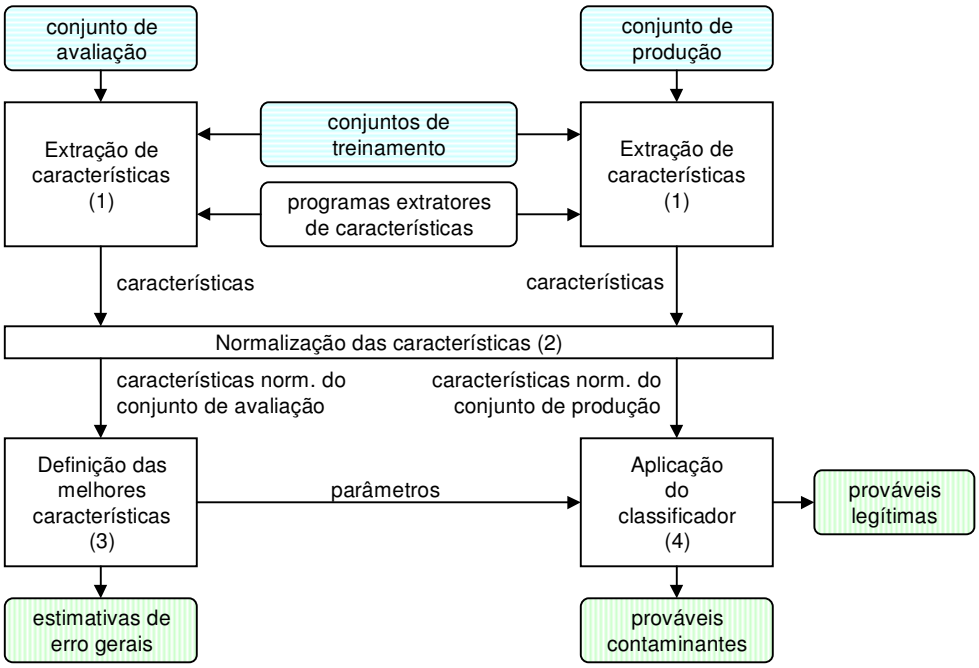

Figura 4.5: Diagrama geral do funcionamento do QUEST. Os números entre parênteses indicam a ordem em que as etapas são executadas.

Figura 4.6: Pseudo-código do programa QUEST

```
Entradas: O      // conjunto de organismos
          E      // conjunto de extratores
          C      // classificador a ser usado
          SR     // seqüências de treinamento (rotuladas)
          SA     // seqüências de avaliação (rotuladas)
          SP     // seqüências de produção (rotuladas ou não)
programa QUEST

    // Extração dos conjuntos de características
    CA ← {}          // caracs. relativas às seqüências de SA
    CP ← {}          // caracs. relativas às seqüências de SP
    para cada seqüência sa_i
        ca_i ← constrói matriz de características(sa_i,SR,O,E)
    para cada seqüência sp_i
        cp_i ← constrói matriz de características(sp_i,SR,O,E)

    // Normalização das características
    (CA',CP') ← normaliza características(CA,CP)

    // Seleção do subconjunto das melhores características
    MERRO ← ∞     // menor erro conseguido
    MCC ← {}      // melhor combinação de características
    PARAMS ← {}   // parâmetros do classificador
    para cada combinação CC de características
        (PARAMS,ERRO) ← classificador(C,CA')
        se ERRO < MERRO
            MCC ← CC
            MERRO ← ERRO

    // Classifica SP
    (RESUMO,RESULTADO) ← classificador(C,CP',PARAMS)

    // Exibe resultados
    imprime MCC, MERRO, RESUMO
    para cada seqüência sp_i
        imprime sp_i, resultado_i
```

Capítulo 5

Testes, resultados e análise

Neste trabalho foram realizados dois testes principais utilizando o QUEST. No primeiro, foi feita uma verificação da metodologia, usando somente seqüências cuja origem era conhecida. No segundo teste, a metodologia foi testada utilizando-se seqüências provenientes do projeto EST de *Schistosoma mansoni*. Os resultados são apresentados neste capítulo.

5.1 Testes com projetos EST fictícios

O primeiro teste foi feito como forma de validar a metodologia. Pelo fato dela ser baseada em um sistema de reconhecimento de padrões, diversas questões relacionadas precisam ser abordadas:

- A classificação de seqüências de DNA baseada somente em suas características intrínsecas é possível ? Ela produz resultados satisfatórios ?
- Quão úteis para a classificação são as características obtidas pelos extratores ? O uso conjunto de características traz benefícios à classificação ?
- Qual o desempenho dos classificadores utilizados sob condições controladas ? Qual o classificador mais preciso ? Qual o de maior poder de generalização ?
- Qual o tamanho do conjunto de treinamento necessário para que se consiga um desempenho razoável ?

Essas e outras questões precisam ser respondidas para que se possa trabalhar com mais confiança em projetos reais. Para isso, diversos projetos EST foram simulados, sendo escolhido como organismo alvo a mosca *Drosophila melanogaster* (referenciada aqui simplesmente como DM), e como contaminantes os organismos da Tabela 5.1. Os contaminantes foram escolhidos pela maior ou menor distância filogenética de *D.melanogaster*

e também por seu conteúdo GC, e não por serem organismos contaminantes prováveis em um projeto EST de *D.melanogaster*.

Tabela 5.1: Organismos contaminantes usados com o organismo alvo *D.melanogaster* na criação dos projetos fictícios. Todos estes organismos já tiveram seus genomas completamente seqüenciados. A coluna *abreviação* indica como os organismos são referenciados no trabalho.

organismo	abreviação	motivo da escolha
Clostridium perfrigens	CP	Bactéria, filogeneticamente distante de DM, com percentual GC bastante diverso (30% contra 54%)
Escherichia coli	EC	Bactéria, também filogeneticamente distante de DM, com percentual GC próximo (52% contra 54%)
Saccharomyces cerevisiae	SC	Levedura, mais próxima a DM que qualquer bactéria, com percentual GC de 40%
Caenorhabditis elegans	CE	Nematelminto, mais próximo a DM que os outros organismos acima, com percentual GC de 45%

Todo projeto simulado continha DM como organismo alvo, e uma das combinações possíveis de contaminantes, totalizando $(2^4 - 1) = 15$ projetos. Como os projetos têm sempre o mesmo organismo alvo, eles são identificados pelos contaminantes empregados. Por exemplo, o projeto (CP, CE) refere-se ao projeto contendo DM como organismo alvo e CP e CE como contaminantes. Para cada um dos 15 projetos foi criado um arquivo de configuração conforme descrito na Seção 4.1. As seqüências dos organismos foram obtidas do genbank [6], sendo escolhidas apenas aquelas representando regiões codificadoras de proteínas (CDS).

5.1.1 Análise dos programas extratores de características

A análise das características obtidas pelos extratores utilizados é muito importante, pois o desempenho dos classificadores depende diretamente delas. Para o estudo dos sete extratores empregados, a ferramenta QUEST foi utilizada no projeto (CE), por ele ser o caso de separação teoricamente mais difícil (devido à proximidade filogenética entre os organismos do projeto) usando apenas um contaminante. Para DM, foram usadas 200 seqüências para o conjunto de treinamento e outras 200 para o de avaliação, e 500 e 100 seqüências de treinamento e avaliação respectivamente para CE.

A capacidade de cada característica de distinguir os organismos pode ser estimada graficamente conforme ilustrado na Figura 5.1, onde o eixo x representa o valor da característica dada pelo extrator com treinamento de DM e o eixo y o valor com treinamento de CE. Quanto maior a separação entre as características de DM e de CE mais fácil será o trabalho de cada classificador, e conseqüentemente melhor será seu desempenho.

Os gráficos dados pelas implementações de BN e DS mostram que, se eles retornam um valor x de característica para treinamento com DM, retornam também sempre um outro valor y fixo para o treinamento com CE, sinalizando uma maior simplicidade de seus modelos internos.

Apesar de visualmente não ser possível dizer qual dos extratores pode fornecer a melhor separação entre as classes, o problema da separação parece factível de ser resolvido, embora exija um tratamento sofisticado. Uma análise dos classificadores utilizando as características de cada extrator separadamente e em conjunto é feita a seguir.

5.1.2 Análise de desempenho usando combinação de características

A fim de verificar o benefício introduzido pelo uso de diversas características nos classificadores, eles foram testados usando as características isoladamente e em conjunto. Para isso, foram definidos os conjuntos de seqüências rotuladas descritos na Tabela 5.2, e testados todos os classificadores paramétricos e não-paramétricos com suas funções de consolidação.

Tabela 5.2: Informações sobre os conjuntos rotulados utilizados para avaliar a performance dos classificadores com uma e várias características. Os conjuntos de cada organismo não contêm seqüências em comum.

organismo	treinamento	avaliação	produção	total	tam.médio (nt)	%GC
DM	200	200	5.000	5.400	1.012	54,1
CP	500	100	100	700	875	29,6
EC	500	100	100	700	860	52,0
SC	500	100	100	700	1.042	41,1
CE	500	100	100	700	859	45,7
total	2.200	600	5.400	7.500	-	-

A quantidade de seqüências presentes em cada conjunto da Tabela 5.2 tenta simular a situação encontrada em projetos EST reais. Para o organismo alvo, o número de seqüências existentes anteriores ao projeto é em geral pequeno, e deve ser distribuído

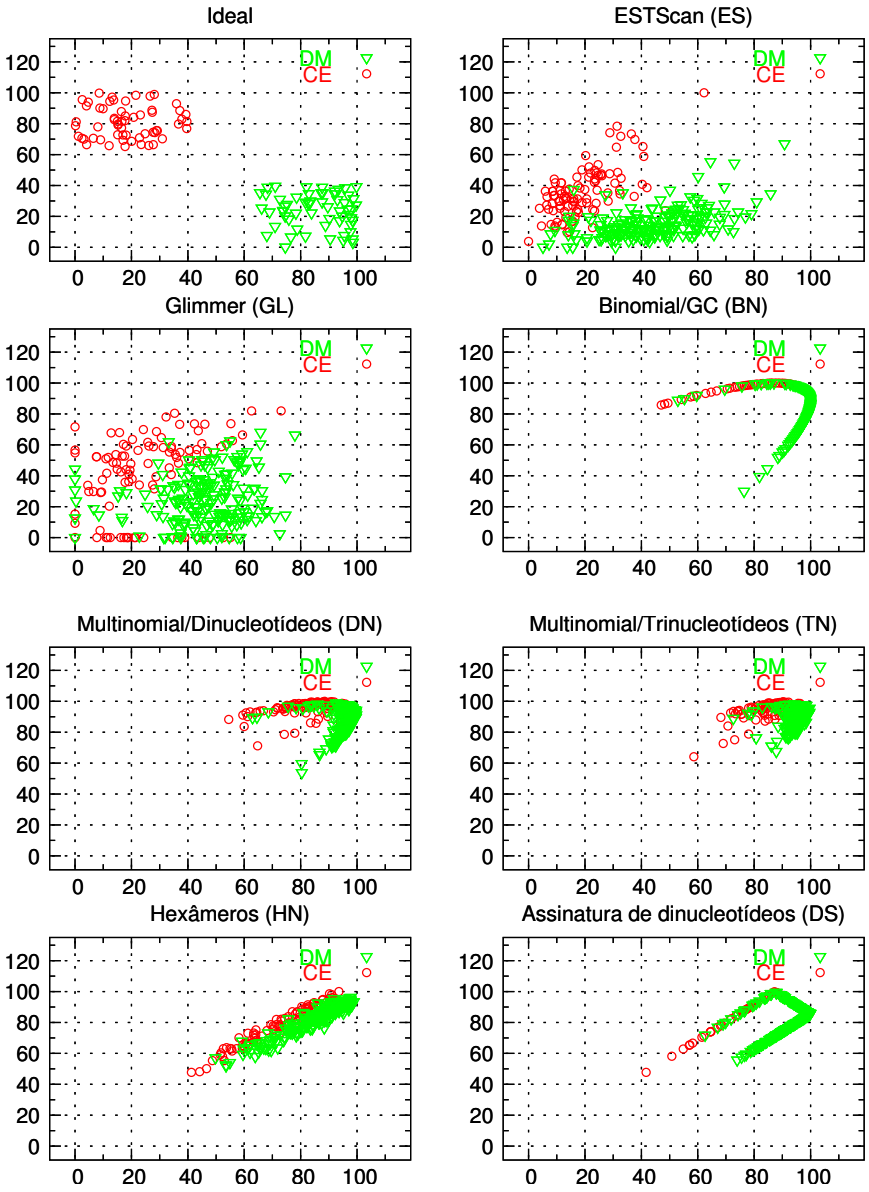

Figura 5.1: Características dadas por cada extrator com diferentes treinamentos para as seqüências do conjunto de avaliação do projeto (CE). O eixo x representa o valor da característica retornada pelo extrator com treinamento de DM e o eixo y o valor da característica com treinamento de CE. O gráfico intitulado "Ideal" foi colocado como referência de optimalidade para a análise dos gráficos restantes. O valor das características foi normalizado conforme descrito na Seção 4.3.

homogeneamente para formar os conjuntos de treinamento e avaliação. Já para os organismos contaminantes, é comum existirem diversas seqüências disponíveis, permitindo a construção de conjuntos de treinamento e avaliação maiores. As seqüências de produção em projetos reais são em sua maioria originárias do organismo alvo, o que explica o tamanho do conjunto de produção desse organismo em comparação aos mesmos conjuntos dos contaminantes.

Resultados gerais para os projetos A Tabela 5.3 apresenta os dois melhores classificadores para cada um dos 15 projetos criados. Os erros (*pFPFN*) para os melhores classificadores variaram entre 0,42% para o projeto (CP) e 10,22% para o projeto (SC, CE); a pior capacidade de generalização foi de 3,48% para o projeto (EC, SC). O classificador baseado em SVMs foi o melhor em 11 dos 15 projetos testados, seguido dos classificadores da votação (dmc) com 2, e da soma e vetores binários (dmc) com um projeto cada. Todos os melhores classificadores utilizaram no mínimo as características de dois extratores, sendo constante a presença do ESTScan.

Aparentemente o número de organismos contaminantes empregados tem alguma influência sobre os valores de *pFPFN*, apesar dessa influência não ser decisiva. Por exemplo, o projeto (SC, CE) foi o que apresentou o maior *pFPFN*, curiosamente maior que SC e CE combinados com os outros contaminantes CP e EC empregados. Isso ocorre porque CP e EC são relativamente fáceis de separar (sozinhos e combinados são os três projetos com menores *pFPFN*), e não atrapalham a separação dos contaminantes SC e CE. No entanto, a presença de CP e EC consegue trazer melhorias aparentes nos valores de *pFP* e *pFN*, pelos seguintes motivos:

- Como CP e EC são mais fáceis de se separar, a proporção entre os contaminantes corretamente classificados e o total de contaminantes aumenta, sem que haja necessariamente um aumento na quantidade de seqüências de SC e CE classificadas corretamente. A diminuição no valor de *pFP* pode ser portanto atribuída a esse fato.

- A adição de EC ao projeto (SC, CE) fez com que houvesse um pequeno acréscimo na proporção de seqüências de DM corretamente classificadas (de 97,5% para 98%), implicando em uma redução no valor de *pFN*. Essa melhora não é significativa (ver exemplo de piora para os projetos (CP) e (CP, EC)) e é causada por modificações nos classificadores introduzidas pelas novas características.

Assim, pela amostra de projetos feita, o erro é mais influenciado pela escolha dos organismos contaminantes do que pela sua quantidade.

Um fato curioso observado na Tabela 5.3 é que o valor de *pFN* é menor que *pFP* em 14 dos 15 projetos analisados. Isso parece ser um viés introduzido pelos classificadores ao

usarem as características do ESTSCAN , uma vez que os projetos (CP, EC) e (EC, SC) utilizaram um maior número de extratores e obtiveram valores mais próximos entre *pFN* e *pFP*, talvez por diminuirem a influência do ESTSCAN nos classificadores. Embora valores de *pFN* menores que *pFP* sejam desejáveis em projetos EST, não há um motivo claro para que isso tenha acontecido, e portanto garantias de que isso irá ocorrer na maior parte dos casos.

Os resultados em geral indicam que apesar da separação ter sido possível, os percentuais de falsos positivos e negativos ainda são relativamente altos, sendo necessárias melhorias adicionais para sua redução. Neste sentido, este trabalho representa um primeiro passo, demonstrando a viabilidade da metodologia proposta.

O fato dos melhores classificadores terem usado ao menos dois extratores, a presença do ESTSCAN em todos eles e os bons resultados atingidos pelo classificador baseado em SVMs indicam que o uso de novos extratores de características e classificadores mais poderosos poderiam trazer benefícios, devendo ser estudados mais profundamente.

Análise de um projeto específico Para que o desempenho dos classificadores utilizando uma ou mais características fosse analisado, foi escolhido o projeto que contém o maior número de organismos contaminantes: (CP, EC, SC, CE), considerado por isso o mais próximo a um projeto real. Os resultados para cada classificador estão na Tabela 5.4.

Para que os classificadores funcionassem usando as características de apenas um extrator por vez, foi utilizada apenas a coluna correspondente na matriz C de características. Para estes casos, o classificador paramétrico usado foi o da votação (*(P)* na Tabela 5.4), e o não paramétrico foi o da soma (*(NP)* na mesma tabela).

Os resultados devem ser analisados idealmente pelo valor de *pFPFN* do conjunto de produção, visto que ele irá conter as seqüências de origem desconhecida em um projeto real. Neste sentido, o melhor extrator para o projeto (CP, EC, SC, CE) é o ESTSCAN usando o classificador não paramétrico, atingindo *pFPFN* igual a 10,54%, além de apresentar boa capacidade de generalização (diferença de 0,54% entre *pFPFN* do conjunto de produção e avaliação).

Combinando as características de diversos extratores alguns classificadores conseguiram um desempenho melhor que o apresentado pelo ESTSCAN individualmente, como foi o caso dos classificadores do produto, soma e votação (função de consolidação dmc). Esses classificadores apresentaram para o conjunto de produção valores de *pFPFN* de 9,65%, 10,36% e 9,27% respectivamente, também com boa capacidade generalização, variando entre 1,02% e 1,86%. O caso do classificador da votação exemplifica a complementaridade existente entre as características de diferentes extratores: os extratores usados (ES, GL e HN) tinham *pFPFN* de 9,75, 18,00 e 10,00% respectivamente, que foi reduzido para 9,27% pela sua combinação.

Tabela 5.3: Dois melhores classificadores para cada projeto criado. Estão também indicados para cada um os extratores escolhidos pela seleção de subconjunto de características. As colunas *pFP*, *pFN* e *pFPFN* referem-se ao valores do conjunto de produção, e a coluna *dif* contém a diferença entre *pFPFN* dos conjuntos de produção e avaliação.

projeto	classificador	extratores	*pFP*	*pFN*	*pFPFN*	dif
(CP)	vet.bin,dmc	ES HN	0,42	0,00	0,42	0,42
	votação,dmc	ES HN	0,42	0,00	0,42	0,42
(CP, EC)	votação,dmc	ES GL TN BN HN	1,44	1,50	2,94	2,94
	SVM,dmc	todos	1,96	1,00	2,96	-0,04
(EC)	SVM,dmc	todos	2,26	1,00	3,26	0,26
	soma	ES HN	3,62	0,00	3,62	-0,38
(CP, CE)	SVM,dmc	todos	4,92	1,50	6,42	-0,08
	produto	ES HN	4,46	2,50	6,96	0,96
(SC)	SVM,dmc	todos	5,62	1,00	6,62	-2,38
	soma	ES GL	6,86	2,00	8,86	1,86
(CP, EC, CE)	SVM,dmc	todos	5,30	1,33	6,63	0,80
	produto	ES TN HN	5,84	1,00	6,84	1,01
(CP, SC)	SVM,dmc	todos	4,44	2,50	6,94	0,94
	votação,dmc	ES GL TN HN	3,46	4,00	7,46	3,96
(CE)	soma	ES HN	5,20	2,00	7,20	-1,80
	votação,dmc	ES	6,68	1,00	7,68	0,68
(CP, EC, SC)	SVM,dmc	todos	5,08	2,33	7,41	1,91
	votação,rmc	ES	3,40	4,67	8,07	1,24
(EC, CE)	SVM,dmc	todos	5,80	2,00	7,80	-0,20
	soma	ES	5,76	2,50	8,26	0,76
(EC, SC)	votação,dmc	ES GL DN HN	4,98	3,50	8,48	3,48
	SVM,dmc	todos	6,00	3,00	9,00	2,00
(CP, EC, SC, CE)	SVM,dmc	todos	6,80	2,00	8,80	0,05
	votação,dmc	ES GL HN	6,52	2,75	9,27	1,02
(CP, SC, CE)	SVM,dmc	todos	6,60	2,67	9,27	-0,90
	produto	ES TN BN	6,06	3,67	9,73	1,73
(EC, SC, CE)	SVM,dmc	todos	7,74	2,00	9,74	-0,26
	soma	ES HN	8,26	1,67	9,93	-0,57
(SC, CE)	SVM,dmc	todos	7,72	2,50	10,22	-1,28
	soma	ES DN	7,54	3,50	11,04	1,54

Tabela 5.4: Resultados dos classificadores para o projeto (CP, EC, SC, CE). A coluna *estratégia* contém os classificadores usando várias características (de *produto* até *SVM*) ou apenas aquelas do extrator indicado (de *ES* até *DS*), onde (P) e (NP) indicam o uso de um classificador paramétrico e não paramétrico respectivamente. A coluna *prgs* indica o número de extratores utilizados pela seleção de subconjunto de características para o classificador correspondente. A coluna *dif* contém a diferença entre o segundo e primeiro valor de *pFPFN* para cada linha.

estratégia	conjunto de avaliação					conjunto de produção				dif
	prgs	*Ac*	*pFP*	*pFN*	*pFPFN*	*Ac*	*pFP*	*pFN*	*pFPFN*	
produto	3	96,67	6,00	2,00	8,00	93,41	6,90	2,75	9,65	1,65
soma	3	96,67	7,00	1,50	8,50	92,11	8,36	2,00	10,36	1,86
pluralidade	1	96,00	8,00	2,00	10,00	92,37	8,04	2,50	10,54	0,54
Borda	1	96,00	8,00	2,00	10,00	92,37	8,04	2,50	10,54	0,54
votação,dmc	3	96,50	6,00	2,25	8,25	93,76	6,52	2,75	9,27	1,02
votação,rmc	2	95,00	**2,00**	6,50	8,50	**95,89**	**3,78**	8,25	12,03	3,53
vet.bin,rmc	7	96,83	**2,00**	3,75	**5,75**	94,81	5,00	7,50	12,50	6,75
vet.bin,dmc	7	**97,17**	3,00	2,75	**5,75**	93,54	6,30	8,50	14,80	9,05
SVM,dmc	7	96,50	7,00	1,75	8,75	93,56	6,80	2,00	**8,80**	0,05
SVM,rmc	7	95,33	6,50	3,75	10,25	93,07	7,24	3,00	10,24	**-0,01**
ES (P)	1	94,67	3,50	6,25	9,75	94,81	5,16	5,50	10,66	0,91
ES (NP)	1	96,00	8,00	2,00	10,00	92,37	8,04	2,50	10,54	0,54
GL (P)	1	92,50	13,50	4,50	18,00	87,56	13,08	4,50	17,58	-0,42
GL (NP)	1	90,33	22,00	3,50	25,50	79,80	21,70	**1,50**	23,20	-2,30
BN (P)	1	76,67	10,00	30,00	40,00	86,30	12,46	29,25	41,71	1,71
BN (NP)	1	79,00	40,50	11,25	51,75	65,44	36,46	10,75	47,21	-4,54
DN (P)	1	90,33	15,00	7,00	22,00	84,67	15,76	10,00	25,76	3,76
DN (NP)	1	90,17	18,00	5,75	23,75	82,65	18,04	8,75	26,79	3,04
TN (P)	1	91,50	9,00	8,25	17,25	89,70	10,28	10,50	20,78	3,53
TN (NP)	1	92,67	16,00	3,00	19,00	84,63	16,16	5,50	21,66	2,66
HN (P)	1	95,50	6,50	3,50	10,00	92,04	8,10	6,25	14,35	4,35
HN (NP)	1	92,17	22,00	**0,75**	22,75	79,94	21,54	**1,50**	23,04	0,29
DS (P)	1	79,50	22,00	19,75	41,75	77,78	22,36	20,50	42,86	1,11
DS (NP)	1	79,67	26,00	17,50	43,50	75,74	24,78	17,75	42,53	-0,97

O classificador baseado em vetores binários (*vet.bin* na Tabela 5.4) apresentou um comportamento interessante, pois conseguiu os menores valores de *pFPFN* para o conjunto de avaliação, mas os piores valores para o conjunto de produção. Isso porque ele se torna bastante específico para o conjunto de avaliação, perdendo a capacidade de classificar bem as novas seqüências apresentadas pelo conjunto de produção, possuindo assim uma capacidade de generalização inadequada.

Assim como ocorreu na Tabela 5.3, a maioria dos classificadores (21 de 24) obteve valores de *pFN* menores que *pFP*. Com exceção dos classificadores baseados em vetores binários (rmc) e TN (usando votação), todos os outros mantiveram a relação $pFN < pFP$ ou $pFN > pFP$ entre os conjuntos de avaliação e produção. A conservação dessa relação é importante quando se avalia os conjuntos de possíveis legítimas e contaminantes gerados.

Para analisar quais os organismos contaminantes que mais dificultaram a classificação no projeto (CP, EC, SC, CE), foram contadas quantas seqüências de cada contaminante foram classificadas incorretamente como provenientes de DM (portanto falsos positivos). Os resultados para três dos melhores classificadores (SVM, votação e soma) indicaram que, conforme previsto, os organismos classificados mais vezes como DM foram na ordem CE (6 falsos positivos em média), SC (3) e EC e CP (ambos com 0 falsos positivos). Dentre as seqüências falsas positivas de cada organismo existiram algumas comuns aos três classificadores empregados. Tais seqüências merecem maior atenção, pois elas provavelmente contém características intrínsecas capazes de diferenciá-las de DM que não foram pegas pelos extratores empregados; caso se descobrisse um extrator capaz de obtê-las, o desempenho dos classificadores poderia ser melhorado.

5.1.3 Comparação de desempenho com diferentes tamanhos de conjunto de treinamento

Para avaliar a importância do número de seqüências no conjunto de treinamento no desempenho dos classificadores, foi utilizado novamente o projeto (CP, EC, SC, CE). Ao conjunto de treinamento foram atribuídas por vez quantidades q crescentes de seqüências, variando entre 2 e 750 para cada organismo. Foram usadas 100 seqüências no conjunto de avaliação e outras 100 no conjunto de produção, também para cada organismo. Todos os classificadores foram testados, e seus gráficos estão na Figura 5.2.

Como pode ser observado, o erro (*pFPFN*) para treinamentos com $q < 50$ é elevado e tem grande variância. Com $q \geq 50$, o erro fica estável ou diminui lentamente para a maioria dos classificadores, fato ilustrado pela Tabela 5.5, construída com os mesmos valores usados na Figura 5.2. Essa tabela exibe os valores médios dos erros e desvios padrão para $q < 50$ e $q \geq 50$. O erro médio e desvios padrão são sempre menores quando $q \geq 50$, sugerindo então $q = 50$ como mínimo para que este projeto (CP, EC, SC, CE)

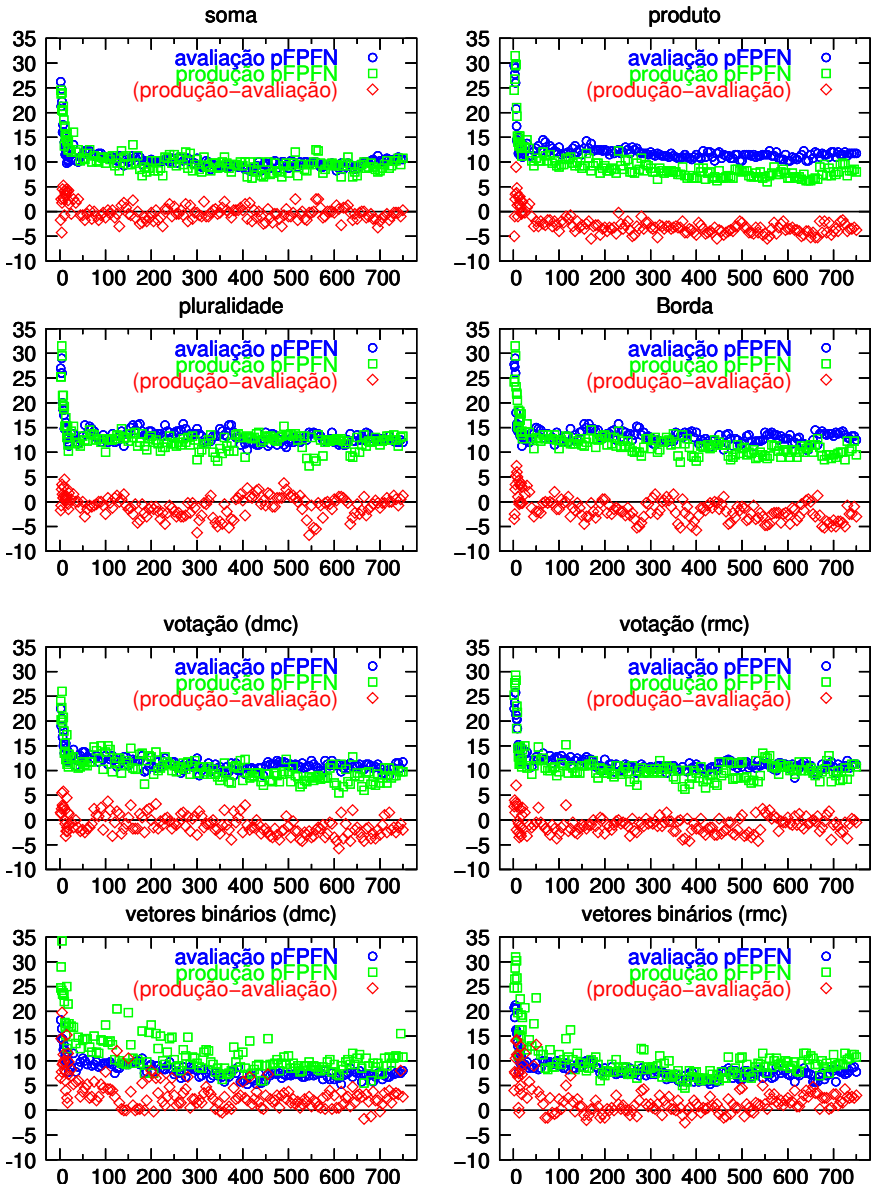

Figura 5.2: Efeito de diferentes tamanhos de conjuntos de treinamento na performance dos classificadores não paramétricos (quatro primeiros) e paramétricos (quatro últimos). O eixo x indica a quantidade q de seqüências de cada organismo empregadas no treinamento, e o eixo y contém respectivamente o erro ($pFPFN$) dos classificadores nos conjuntos de avaliação e produção, além da diferença entre esse dois erros. Um resumo qualitativo destes gráficos está na Tabela 5.5.

apresente taxas de erros menores e com menor variância.

Tabela 5.5: Valores médios para *pFPFN* de cada classificador com menos de 50 e com pelo menos 50 seqüências de treinamento por organismo. As abreviações *vot.* e *vet.bin* referem-se aos classificadores da votação e vetores binários. Os valores da tabela estão na forma $\mu \pm 2\sigma$, onde μ é o valor médio dos percentuais usados na Figura 5.2, e σ é o desvio padrão. O intervalo $[\mu-2\sigma, \mu+2\sigma]$ concentra 95% dos percentuais utilizados, constituindo assim um intervalo de confiança sobre eles. Com 50 ou mais seqüências, os desvios padrão de *pFPFN* para os conjuntos de avaliação e produção são significativamente menores.

classificador	avaliação (*pFPFN*)		produção (*pFPFN*)		diferença	
	< 50	≥ 50	< 50	≥ 50	< 50	≥ 50
soma	$14,3 \pm 9,49$	$9,9 \pm 1,73$	$16,5 \pm 8,20$	$9,5 \pm 2,66$	$2,2 \pm 4,52$	$-0,5 \pm 2,37$
produto	$15,7 \pm 11,92$	$11,7 \pm 1,62$	$16,9 \pm 12,97$	$8,3 \pm 2,70$	$1,2 \pm 5,40$	$-3,4 \pm 2,08$
pluralidade	$16,5 \pm 10,52$	$13,1 \pm 2,22$	$17,3 \pm 11,41$	$12,2 \pm 2,89$	$0,9 \pm 3,08$	$-1,0 \pm 3,92$
Borda	$16,4 \pm 10,38$	$13,1 \pm 2,15$	$18,2 \pm 10,64$	$11,1 \pm 2,95$	$1,8 \pm 5,66$	$-2,0 \pm 3,51$
vot.,dmc	$14,9 \pm 6,12$	$11,1 \pm 1,98$	$15,6 \pm 9,28$	$9,9 \pm 3,96$	$0,7 \pm 4,92$	$-1,2 \pm 3,76$
vot.,rmc	$16,2 \pm 9,33$	$11,1 \pm 1,77$	$16,5 \pm 13,11$	$10,0 \pm 3,08$	$0,3 \pm 5,44$	$-1,1 \pm 2,93$
vet.bin,dmc	$11,5 \pm 5,36$	$7,7 \pm 2,37$	$19,9 \pm 10,89$	$10,3 \pm 5,43$	$8,3 \pm 8,23$	$2,7 \pm 4,39$
vet.bin,rmc	$12,6 \pm 8,39$	$7,6 \pm 2,25$	$20,0 \pm 12,78$	$9,0 \pm 4,35$	$7,4 \pm 7,90$	$1,3 \pm 3,92$

Também através da Tabela 5.5 pode-se ver que o classificador que mais se aproxima da capacidade de generalização ideal (diferença zero entre os erros no conjunto de avaliação e produção) é o método da *soma*, apresentado diferença média de $-0,5 \pm 2,37$. Por outro lado, o classificador baseado em *vetores binários* apresentou a pior generalização, variando entre +1,3 e +2,7% de acordo com a função de consolidação usada, e com os maiores desvios padrão. O classificador baseado no método do *produto* mostrou-se o mais conservador nas estimativas de erros, apresentando o maior valor negativo na diferença entre o erro nos conjuntos de avaliação e produção ($-3,4 \pm 2,08$). Ainda pela Tabela 5.5 observa-se que diferença entre *pFPFN* de produção e avaliação é sempre maior para $q < 50$, indicando que $q \geq 50$ leva a estimativas de erro mais conservadoras e portanto mais apropriadas.

Dado o valor mínimo de $q = 50$ para o projeto (CP, EC, SC, CE), procurou-se usar sempre valores superiores a esse para os conjuntos de treinamentos dos organismos em todos os outros projetos analisados.

5.1.4 Análise dos valores de confiabilidade atribuídos a cada seqüência

Um dos objetivos da metodologia proposta é fornecer valores de confiabilidade para a predição de cada seqüência pedida. Cada um desses valores é dado pelo classificador utilizado, e quanto maior for o valor, mais confiável é a predição dada. Dessa maneira, as seqüências com os maiores valores de confiabilidade terão maior chance de terem sido classificadas corretamente, enquanto as seqüências com os menores valores deverão ser olhadas com maior cuidado. A Figura 5.3 traz exemplos da relação entre valores de confiabilidade e predições corretas dados por quatro classificadores (soma, produto, votação e SVM) para o projeto (CP, EC, SC, CE).

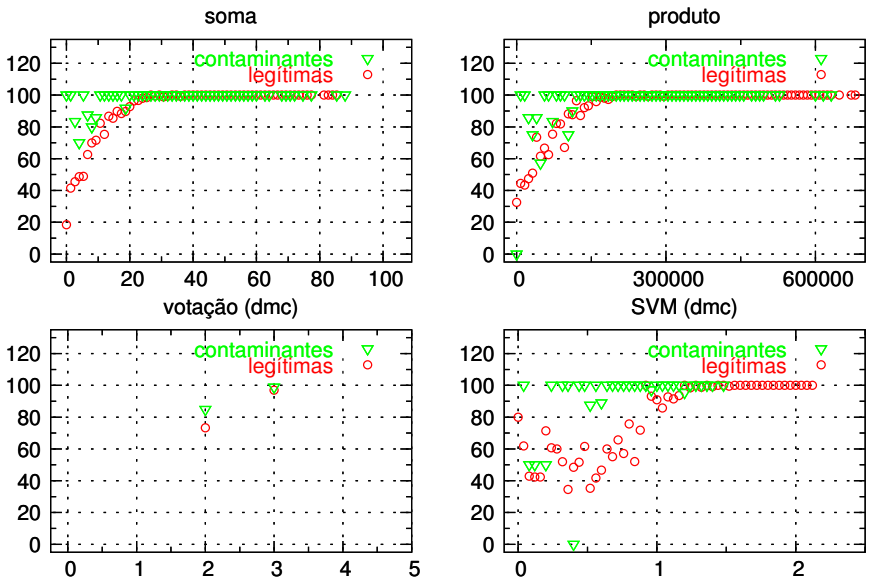

Figura 5.3: Relação entre os valores de confiabilidade e o percentual de seqüências preditas corretamente. O eixo x contém os valores de confiabilidade dados por cada classificador, e o eixo y contém o percentual de predições corretas (pTP para legítimas e pTN para contaminantes) associadas com cada valor no eixo x.

Nesta figura pode-se notar que o percentual de acertos (pTP para seqüências legítimas e pTN para contaminantes) tende a aumentar junto com os valores de confiabilidade, apresentando portanto o comportamento esperado. Alguns gráficos no entanto devem ser comentados. O classificador da votação utilizou três extratores (ver Tabela 5.4) e definiu seu parâmetro n como 2, e por isso seus valores de confiabilidade foram sempre 2 ou 3. No classificador baseado em SVM a relação entre a confiabilidade na predição e o percentual de acertos só fica mais clara a partir do valor 1 de confiabilidade.

Os valores de confiabilidade podem ser usados para se estabelecer conjuntos de seqüên-

cias com diferentes graus de confiabilidade, particionando-se apropriadamente o conjunto original de seqüências classificadas. Poderia haver, por exemplo, três conjuntos: um contendo predominantemente seqüências legítimas, outro com contaminantes, e o último contendo seqüências de origem incerta. Essa divisão possibilita que se trabalhe com conjuntos mais confiáveis, por exemplo quando as taxas de erro globais não forem adequadas. Embora essa separação seja importante, a atual versão do QUEST não a implementa; no entanto a discussão sobre esse particionamento é bastante relevante e é feita a seguir.

A Figura 5.4 ilustra o processo. A idéia é que o usuário defina os percentuais de erros tolerados, fornecendo *pFP* e *pFN*. Com o conjunto de avaliação, obtém-se os valores de confiabilidade das seqüências dos conjuntos de legítimas e contaminantes gerados. A implementação deve então definir quais os dois valores de confiabilidade mínimos para que se atinja *pFP* e *pFN* esperados, ordenando em separado os valores dos conjuntos de legítimas e contaminantes e testando qual o melhor *valor de corte* em cada um deles, de maneira similar à feita na Seção 4.4.3 para os classificadores paramétricos. As seqüências com valores de confiabilidade inferiores aos valores de corte formam um terceiro conjunto de seqüências com classificação incerta.

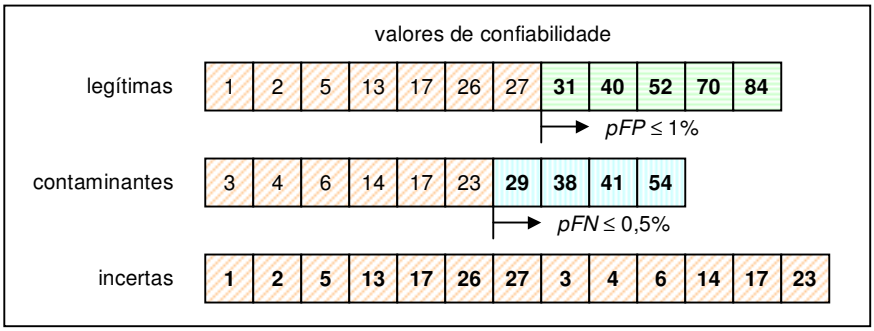

Figura 5.4: Exemplo de valores de corte usados para atingir *pFP* e *pFN* pedidos nos conjuntos de legítimas e contaminações gerados inicialmente. Neste caso, foi pedido $pFP \leq 1\%$ e $pFN \leq 0{,}5\%$, sendo definidos os valores de corte de 31 e 29 respectivamente nos conjuntos de legítimas e contaminantes. Todas as seqüências com valores inferiores aos valores de corte formam o conjunto de classificação incerta. Os três conjuntos resultantes (legítimas, contaminantes e incertas) são formados pelos valores em destaque.

Os valores de confiabilidade dados por certos classificadores podem no entanto não serem adequados para esta tarefa, por causa da granularidade de seus valores. Um exemplo disto está na própria Figura 5.3. No caso do classificador da votação, o único particionamento possível (além do particionamento padrão que engloba todas as seqüências) seria o das seqüências com valor de confiabilidade igual a três, que pode não resultar nos valores de *pFP* e *pFN* esperados pelo usuário. A ação mais apropriada nestes casos é informar ao usuário os diferentes valores de *pFP* e *pFN* possíveis.

5.1.5 Resumo

Os testes realizados com os projetos fictícios mostraram que a separação de seqüências entre um organismo e um outro conjunto de organismos é possível, embora o erro resultante desta separação ainda precise ser diminuído. O uso de diversos extratores para a extração de características e seu uso com os classificadores baseados na votação, soma e SVM apresentou os melhores resultados, sendo portanto preferível o uso de tais classificadores. O tamanho mínimo para o conjunto de treinamento dos organismos usados (alvo e contaminantes) é de 50 seqüências; valores superiores (por exemplo de 100 a 500 seqüências) são recomendáveis. Um maior número de organismos contaminantes empregados não foi determinante na piora das taxas de erro obtidas.

5.2 Teste com um conjunto real

O *Schistosoma mansoni* é o agente causador da doença denominada *esquistossomose*, que atinge 200 milhões de indivíduos em 74 países (inclusive o Brasil), e que pode levar a morte. Com o objetivo de se aumentar a quantidade de informações disponíveis para este organismo, a FAPESP[1] iniciou em abril de 2001 o que foi chamado *projeto EST de Schistosoma mansoni* [29]. A escolha por um projeto EST foi devido ao tamanho relativamente grande do genoma do organismo, com aproximadamente 270 milhões de bases. Este projeto gerou 163.586 ESTs, que resultaram em 30.988 seqüências montadas, representando uma amostra de 92% do total de 14.000 genes estimados para o organismo.

Dentre estas 163.586 ESTs apareceram diversas sugestões de contaminações. Para lidar com este problema, o projeto original utilizou uma abordagem baseada em similaridade. Apesar de se ter conseguido separar diversas seqüências com grande probabilidade de serem contaminantes, as deficiências e problemas apresentados por essa abordagem motivaram o desenvolvimento deste trabalho. Não houve entretanto tempo suficiente para a execução do QUEST antes da publicação do artigo decorrente do projeto; ainda assim, a sua avaliação é importante para analisar o comportamento do QUEST em situações reais. Esta seção apresenta os resultados do QUEST para as seqüências geradas pelo projeto EST de *Schistosoma mansoni*, referenciado nesta seção simplesmente por projeto.

Descrição das seqüências do projeto Foram obtidas as 163.586 seqüências, denominadas aqui de *reads*. Através de similaridade, esse total foi separado pelo projeto em três conjuntos disjuntos:

1. contaminantes: 9.506 reads considerados supostos contaminantes.

[1] Fundação de Amparo à Pesquisa do Estado de São Paulo

2. sem interesse: 29.399 reads classificados como não sendo de interesse para o propósito original do projeto, como por exemplo ribossomais, mitocondriais, repetições e transposons.

3. legítimas: os 124.681 reads restantes considerados como originários do organismo alvo *S.mansoni*.

Esse conjunto de 124.681 reads foi montado pelo projeto usando o programa CAP3, dando origem a um conjunto de 30.988 seqüências denominadas SmAEs (*Schistosoma mansoni* assembled EST sequences). Essas SmAEs e todos os reads originais foram obtidos a partir da página do projeto na internet (*http://bioinfo.iq.usp.br/schisto*) e através de comunicação pessoal com o coordenador do projeto. O QUEST foi usado para uma nova análise sobre o conjunto de reads considerados contaminantes e sobre as SmAEs (consideradas legítimas) geradas.

Seleção dos contaminantes para uso com o QUEST Diversas contaminações eram previstas para este projeto, pois foram obtidas amostras do parasita em humanos e camundongos, nos seus vetores (caramujos), além de suas formas livres. Além delas, outras contaminações (predominantemente bacterianas) surgiram ao se comparar as seqüências do projeto com as de bancos públicos através da ferramenta BLAST. A lista de contaminantes utilizada pelo QUEST foi construída com base nessas discussões surgidas dentro do projeto, e sua relação está na Tabela 5.6.

Tabela 5.6: Descrição dos organismos possivelmente contaminantes usados no projeto de *Schistosoma mansoni*. Os motivos da escolha são: (1) seqüências contaminantes ou ortólogos encontrados através do BLAST; (2) utilizados como hospedeiros do *S.mansoni*.

organismo	abreviação	tipo	motivo da escolha
Bacillus subtilis	BS	bactéria	1
Bradyrhizobium japonicum	BJ	bactéria	1
Pseudomonas fluorescens	PF	bactéria	1
Xanthomonas axonopodis pv. citri	XA	bactéria	1
Xylella fastidiosa	XF	bactéria	1
Homo sapiens	HS	ser humano	2
Mus musculus	MM	camundongo	2

Para compor os conjuntos de treinamento e avaliação, as seqüências dos organismos contaminantes escolhidos e do organismo alvo (*S.mansoni*) foram obtidas em 05/04/2004 a partir do genbank [6], utilizando-se somente aquelas representado a região codificadora

(CDS) de genes desses organismos. As seqüências foram amostradas aleatoriamente dentro do total obtido, de forma a tornar as análises o mais homogêneas possível. Informações gerais sobre as seqüências utilizadas estão na Tabela 5.7.

Tabela 5.7: Informações sobre os conjuntos rotulados utilizados. *SM* é a abreviação usada para o organismo alvo *Schistosoma mansoni*. Os organismos marcados com "*" já tiveram seus genomas completamente seqüenciados.

organismo	seqs válidas	treinamento	avaliação	total	tam.médio (nt)	%GC
SM	262	150	112	262	797	38,5
BS*	3.773	500	100	600	827	44,4
BJ*	7.741	500	100	600	857	64,9
PF	5.374	500	100	600	845	61,6
XA*	3.819	500	100	600	851	65,2
XF*	2.208	500	100	600	768	54,0
HS*	80.006	500	100	600	903	53,2
MM	68.584	500	100	600	918	52,3
total	171.767	3.650	812	4.462	-	-

Escolha dos classificadores Dentre os classificadores descritos neste trabalho, foram escolhidos aqueles que apresentaram melhores resultados para projetos fictícios de acordo com a Tabela 5.3: soma, votação (dmc) e SVM (também dmc).

5.2.1 Avaliação das seqüências classificadas como legítimas pelo projeto

A primeira análise foi feita com as 30.988 SmAEs obtidas pelo projeto, resultado da montagem dos 124.681 reads considerados legítimos. A princípio, todos os contaminantes do projeto deveriam ter sido retirados por um processamento prévio, e sendo assim as 30.988 SmAEs seriam originárias do organismo alvo *Schistosoma mansoni*. O objetivo do QUEST neste caso é a procura por contaminações não identificadas previamente.

Dentre as SmAEs, 3.568 não tinham tamanho adequado para análise através da ferramenta QUEST, sendo utilizadas as 27.430 SmAEs restantes.

Resultados gerais Os resultados gerais da análise das SmAEs estão na Tabela 5.8. Os erros previstos são comparáveis a aqueles encontrados nos projetos fictícios testados anteriormente, variando entre 8,54% e 14,93%, porém utilizando agora um maior número

Tabela 5.8: Análise de 27.430 SmAEs do projeto EST de *S.mansoni* por três classificadores. Estão indicadas as taxas de erros previstas obtidas por cada classificador.

	votação	soma	SVM
SmAEs legítimas	21.293	21.181	23.274
pTP	93,75	87,50	90,18
pFP	2,29	2,43	4,14
SmAEs contaminantes	6.137	6.249	4.156
pTN	97,71	97,57	95,86
pFN	6,25	12,50	9,82
$pFPFN$	8,54	14,93	13,96
extratores usados	ES DN	ES GL TN BN DS	todos

de organismos contaminantes (7 contra 4). Os valores de pFN foram menores que pFP para os classificadores da votação (2,29% contra 6,25%) e soma (2,43% contra 12,50%), sendo este um comportamento desejável conforme discutido.

O número de contaminantes apresentados na Tabela 5.8 deve ser interpretado com cuidado. Por exemplo, no caso da classificação pelo método da votação, das 27.430 seqüências analisadas, 6.137 (22,37%) foram rotuladas como contaminações. No entanto, isso não significa necessariamente que 22,37% das SmAEs analisadas sejam contaminações: na verdade, o conjunto denominado "contaminantes" contém teoricamente 97,71% (que é o valor de pTN) das contaminações presentes no conjunto de SmAEs inicial. No entanto, uma vez que não se conhece *a priori* a proporção entre contaminantes e seqüências legítimas, esse conjunto poderia ser composto por exemplo por uma maioria de seqüências legítimas classificadas incorretamente. Há contudo indícios de que realmente existem contaminantes dentre as SmAEs, conforme será visto a seguir.

Estimativa do percentual de contaminações no conjunto de SmAEs original O percentual de contaminações presentes no conjunto de SmAEs do projeto original analisadas pode ser estimado segundo o método descrito na Seção 4.5.1. A estimativa foi feita para cada um dos classificadores empregados, e os resultados estão a seguir:

- Votação: 6.023 SmAEs possivelmente contaminantes (21,96% do total)
- Soma: 6.562 (23,92% do total)
- SVM: 1.699 (6,20% do total)

Isso resulta, na estimativa mais conservadora, em um percentual de 6,20% de contaminações dentre as 27.430 SmAEs analisadas.

Avaliação das predições do QUEST para algumas SmAEs A fim de avaliar as predições feitas pelos três classificadores para uma amostra de SmAEs, foram utilizados os resultados de uma classificação baseada em similaridade. Assim como a metodologia proposta neste trabalho, a classificação baseada em similaridade serve apenas como sugestão para a origem das seqüências, uma vez que a origem real só pode ser determinada através de experimentação laboratorial. As predições baseadas em similaridade foram feitas tendo-se ciência de suas limitações.

A amostra de SmAEs avaliada foi escolhida com base em seus valores de confiabilidade. Foram usadas as 25 SmAEs classificadas como legítimas e contaminantes com os maiores valores de confiabilidade de cada um dos três classificadores usados. Essas SmAEs foram comparadas através da ferramenta BLAST[2] com as seqüências de um banco público[3], sendo analisadas apenas as SmAEs com alguma seqüência similar. É importante observar que a classificação por similaridade feita aqui utilizou apenas o banco citado, enquanto a classificação feita pelo projeto original utilizou um número maior de bancos. Seguiu-se uma classificação manual usando os critérios abaixo:

1. se as 10 seqüências mais similares no banco público fossem todas originárias de bactérias, ou se a seqüência mais similar pertencesse a um dos organismos contaminantes usados com alinhamento de pelo menos 99% de identidade, a SmAE era considerada provável contaminação.

2. caso contrário a SmAE era considerada como provável legítima.

Os resultados gerais dessa classificação são apresentados na Tabela 5.9. Assim como na Tabela 5.8 (que apresenta os resultados gerais do QUEST), os melhores classificadores para essa amostra de SmAEs foram, na ordem, o da votação, soma e SVM. No conjunto de SmAEs consideradas legítimas pelo QUEST não houve nenhuma considerada como contaminação pela similaridade, o que indica boa correlação entre altos valores de confiabilidade e predições provavelmente corretas neste conjunto.

Já no conjunto de SmAEs consideradas contaminantes pelo QUEST estavam presentes diversas consideradas legítimas pela similaridade. Uma explicação para isso pode ser o fato do número real de seqüências contaminantes no conjunto de possíveis contaminantes não ser muito grande. Embora o classificador da votação tenha conseguido em tese colocar 97,71% de todos os contaminantes neste conjunto, seu número absoluto pode ainda ser pequeno comparado ao de SmAEs legítimas falsas negativas também presentes neste conjunto. Se a classificação por similaridade estiver correta, mostra-se também

[2]foram usados os mesmos parâmetros do projeto original: programa BLASTX, sem filtro, e-value $\leq$ 1e-5

[3]banco nr (*non redundant*) copiado de *http://www.ncbi.nlm.nih.gov* em 01/04/2004

Tabela 5.9: Verificação da classificação das 25 SmAEs preditas como legítimas e contaminantes com melhores valores de confiabilidade para cada classificador usado. Foram consideradas apenas as SmAEs com alguma seqüência similar em um banco de seqüências de origem conhecida. A coluna *classe* representa a classe atribuída pelo QUEST. As predições *provavelmente corretas* e *provavelmente incorretas* são baseadas na análise via similaridade.

classificador	classe	prov.corretas	prov.incorretas
votação	legítimas	25	0
	contaminantes	22	3
soma	legítimas	25	0
	contaminantes	20	5
SVM	legítimas	25	0
	contaminantes	17	8

que o QUEST pode eventualmente retornar altos valores de confiabilidade para seqüências preditas incorretamente, fato que merece estudos adicionais.

Não existem SmAEs comuns entre as 16 possivelmente preditas de forma incorreta como contaminantes na Tabela 5.9 (3 pela votação, 8 por SVM e 5 pela soma), o que revela um certo grau de complementaridade entre os três classificadores usados. Este fato poderia ser usado na construção de um combinador de classificadores com o objetivo de melhorar a classificação, o que merece mais estudos.

Avaliação de SmAEs possivelmente ortólogas a seqüências contaminantes A maioria das prováveis SmAEs contaminantes utilizadas na Tabela 5.9 são possivelmente ortólogas às seqüências presentes nos bancos públicos. Isso pode ser visto pelos organismos listados na Tabela 5.10, que traz uma amostra das SmAEs com os mais altos valores de confiabilidade de serem contaminantes segundo o classificador da votação. Nela, apenas 3 dos 16 diferentes organismos listados (BJ, BS e HS) foram usados no treinamento do QUEST.

Isso significa que a ferramenta consegue detectar não somente seqüências dos sete organismos contaminantes utilizados, mas também organismos filogeneticamente próximos a eles. Essa é uma característica muito importante para uma ferramenta como o QUEST, que está também presente nas ferramentas baseadas em similaridade.

Como se vê na Tabela 5.10, várias das prováveis SmAEs contaminantes têm um baixo percentual de identidade com a seqüência mais similar encontrada. Isso ressalta outra qualidade da metodologia implementada pelo QUEST: ela detecta contaminações onde a metodologia baseada em similaridade seria de difícil aplicação, pois haveria dificuldade

em distinguir entre uma seqüência ortóloga do organismo alvo e uma ortóloga de um contaminante.

Avaliação das SmAEs sem similaridade No projeto de *S.mansoni*, mais da metade (55%) das SmAEs não tiveram similaridade significativa com nenhuma seqüência de bancos públicos. Cada uma dessas SmAEs poderia ser, portanto, originária do organismo alvo ou de contaminantes até então desconhecidos, dada a metodologia baseada em similaridade usada na determinação da origem das mesmas.

Uma das vantagens da metodologia proposta é conseguir classificar essas seqüências sem similaridade. A Tabela 5.11 mostra que, conforme esperado, o QUEST fornece uma classificação para todas as 10.282 SmAEs sem similaridade existentes dentro do conjunto de 27.430 analisadas. Das SmAEs sem similaridade, o QUEST classificou respectivamente 23,29, 26,24 e 18,70% como contaminações, segundo os classificadores da votação, soma e SVM. Esses valores são bastante altos, mas conforme já discutido, os conjuntos de prováveis legítimas e contaminantes devem ser examinados pelos valores de confiabilidade de suas seqüências.

5.2.2 Avaliação das seqüências classificadas como contaminantes pelo projeto

O QUEST foi executado sobre o conjunto de 9.506 reads considerados contaminantes pelo projeto original para verificar se existiam reads originários de *S.mansoni* presentes, e portanto classificados incorretamente.

Para esta análise foram pegos 9.495 do total de 9.506 reads classificados como contaminantes pelo projeto original. Um conjunto de 11 reads não foi encontrado entre as seqüências do projeto e portanto não foi utilizado. Desses 9.495 reads, 1.297 não tinham o tamanho adequado para utilização pelo QUEST, sendo analisados os 8.198 restantes.

Resultados gerais A Tabela 5.12 contém os resultados gerais dados pelos três classificadores para o conjunto de 8.198 reads. O valor de *pFPFN* variou entre 8,86% para o classificador da votação e 13,90% para o da soma, próximo portanto dos valores obtidos para os projetos fictícios testados.

A quantidade de reads considerados contaminantes pelos três classificadores foi sempre maior que a de legítimas, sendo este um indicativo de que há de fato um maior número de contaminantes reais dentre os 8.198 reads testados. Contudo, o número de reads considerados legítimos foi considerável, variando entre 11% e 46%, o que sinaliza também a existência de falsos negativos no conjunto.

Tabela 5.10: Amostra de 25 SmAEs que têm ao menos uma seqüência similar no banco público usado e que tiveram o maior valor possível de confiabilidade de serem contaminantes segundo o classificador da votação. A coluna *%id* contém o percentual de identidade entre a SmAE e a seqüência mais similar do banco público. As SmAEs marcadas com "*" são provavelmente falsas sugestões de contaminações, uma vez que as 10 seqüências mais similares ou não vêm de bactéria ou não têm identidade $\geq$ a 99% no caso dos organismos usados pelo QUEST, como o *Homo sapiens*. Os organismos usados no treinamento do QUEST e presentes nesta lista estão marcados com "**". O organismo *Cricetulus griseus* na linha 11 é hamster e provavelmente é uma contaminação, dada a proximidade entre hamster e camundongo e o alto percentual de identidade (97%).

#	SmAE	organismo com seqüência mais similar	%id
1	C600010.1 *	*Cervus nippon*	33
2	C600031.1	*Cytophaga hutchinsonii*	39
3	C600040.1	*Bradyrhizobium japonicum***	44
4	C600098.1	*Streptomyces coelicolor*	57
5	C600104.1	*Bacillus anthracis*	65
6	C600111.1	*Synechococcus sp.*	35
7	C600127.1	*Bacillus subtilis***	77
8	C600166.1	*Bacillus halodurans*	65
9	C600172.1	*Bacillus halodurans*	52
10	C600182.1	*Bacillus anthracis*	40
11	C600270.1	*Cricetulus griseus*	97
12	C600277.1 *	*Bos taurus*	76
13	C600338.1 *	*Homo sapiens***	50
14	C600339.1	*Pseudomonas putida*	51
15	C600374.1	*Streptomyces coelicolor*	52
16	C600438.1	*Pseudomonas syringae*	83
17	C600485.1	*Bacillus subtilis***	70
18	C600497.1	*Bacillus subtilis***	38
19	C600525.1	*Bacillus halodurans*	47
20	C600540.1	*Pseudomonas putida*	76
21	C600571.1	*Oceanobacillus iheyensis*	77
22	C600746.1	*Homo sapiens***	99
23	C600823.1	*Streptomyces coelicolor*	42
24	C600847.1	*Corynebacterium efficiens*	44
25	C600889.1	*Corynebacterium glutamicum*	49

Tabela 5.11: SmAEs com e sem similaridade analisadas pelos classificadores da votação, soma e SVM. As linhas *% similares a mais* indicam o percentual de seqüências similares a mais com relação às não similares.

total de SmAEs	27.430		
com similar	17.148 (62,52%)		
sem similar	10.282 (37,48%)		
	votação	soma	SVM
SmAEs legítimas	21.293	21.181	23.274
com similar	13.406	13.597	14.915
sem similar	7.887	7.584	8.359
% similares a mais	70%	79%	78%
SmAEs contaminantes	6.137	6.249	4.156
com similar	3.742	3.551	2.233
sem similar	2.395	2.698	1.923
% similares a mais	56%	32%	16%

Avaliação das predições do QUEST para alguns reads Assim como foi feito para as SmAEs, foi usada a abordagem baseada em similaridade como forma de avaliar as predições feitas pelo QUEST. Deve-se lembrar que tanto a metodologia proposta quanto a da similaridade servem apenas para fornecer indícios e não provas de legitimidade ou não para qualquer seqüência.

Foi analisada uma amostra de 25 reads considerados legítimos e com máximo valor de confiabilidade segundo o classificador da votação. Os critérios de similaridade usados na classificação foram os mesmos da seção anterior. Para os 25 reads analisados, a classificação por similaridade considerou 13 como contaminantes e 12 como possíveis legítimos, conforme visto na Tabela 5.13.

Uma possível explicação para o alto número de prováveis contaminantes pode ser o fato da classificação original do projeto (baseada em similaridade) ter sido bastante rigorosa, contando inclusive com supervisão humana. Com isso, conseguiu-se separar um conjunto de contaminantes com um número muito pequeno de falsos negativos. Caso isso seja verdade, existem poucos reads legítimos dentro do conjunto considerado pelo QUEST como tal, o que explica a discordância entre as duas metodologias.

Outro fato que pode ter atrapalhado a classificação foi o uso de reads (não montados) em contraste com a análise das SmAEs (compostas por reads montados) usadas na Seção 5.2.1. Os reads apresentam em geral um maior número de erros em suas seqüências. Isso prejudica o desempenho do QUEST, visto que ele usa apenas as informações contidas

Tabela 5.12: Análise feita por três classificadores de 8.198 reads considerados contaminações pelo projeto EST de *S.mansoni*. Os valores de confiabilidade gerais dados por cada classificador estão indicados, assim como os extratores utilizados.

	votação	soma	SVM
SmAEs legítimas	3.770	874	1.639
pTP	96,43	88,39	92,86
pFP	5,29	2,29	2,57
SmAEs contaminantes	4.428	7.324	6.559
pTN	94,71	97,71	97,43
pFN	3,57	11,61	7,14
pFPFN	8,86	13,90	9,71
extratores usados	ES GL HN	ES GL DN TN	todos

nas próprias seqüências para trabalhar.

5.2.3 Discussão

Um projeto EST real apresenta algumas dificuldades adicionais quando comparado com os projetos fictícios discutidos na Seção 5.1. Por exemplo, esses projetos podem amostrar ESTs de partes não traduzidas de genes (regiões UTR), sendo que a metodologia implementada pelo QUEST trabalha com as regiões traduzidas dos mesmos (regiões CDS). No caso do projeto EST de *S.mansoni*, a maioria das seqüências foi obtida através da metodologia ORESTES, que gera ESTs preferencialmente da região codificadora dos genes [10], o que diminuiu a presença e impacto das regiões UTR no funcionamento do QUEST. No entanto, projetos EST que não são baseados na metodologia ORESTES (e que amostram regiões UTR com maior freqüência) são muito mais comuns, fazendo com que o impacto dessas regiões seja provavelmente mais acentuado.

Para o projeto EST de *S.mansoni*, o QUEST conseguiu manter os valores de erros em patamares compatíveis com os apresentados para projetos de teste. Uma análise mais detalhada de uma amostra SmAEs revelou diversas delas como possíveis contaminantes, possibilidade reforçada pelo uso da similaridade como segunda opinião. Essa mesma análise revelou também que o QUEST classificou diversas SmAEs como contaminantes enquanto a similaridade classificou-as como legítimas. Nestes casos há maior probabilidade do QUEST ter errado, o que é de certa forma esperado. Como o número de SmAEs legítimas é maior que o de contaminantes, e o QUEST busca construir um conjunto de contaminantes o mais concentrado possível, isso implica também na presença de um número de falsos negativos, proporcional a relação (desconhecida) entre legítimas e contaminantes

Tabela 5.13: Amostra de 25 reads com ao menos uma seqüência similar no banco público usado e que tiveram o maior valor possível de confiabilidade de serem legítimas segundo o classificador da votação. A coluna *%id* contém o percentual de identidade entre o read e a seqüência mais similar do banco público. Os reads marcados com "*" são provavelmente falsas sugestões de legítimas, uma vez que as 10 seqüências mais similares ou vêm de bactéria ou têm identidade $\geq$ a 99% no caso dos organismos usados pelo QUEST, como o *Homo sapiens*. O organismo *Biomphalaria glabrata* marcado com "**" é um caramujo (vetor do *S.mansoni*) e é também uma provável falsa sugestão de legítima.

#	ident. read		organismo com seqüência mais similar	%id
1	ME1-0098T-D348-G05-U.B		*Pisum sativum*	80
2	MS1-0123U-A366-B01-U.G		*Homo sapiens*	83
3	MS1-0094P-V358-B04-U.B	*	*Clostridium perfrigens*	58
4	MG1-0056U-A270-G10-2.B		*Nostoc punctiforme*	47
5	MG1-0056U-A270-G10-1.B		*Leptospira interrogans*	51
6	MA3-9999U-M343-A01-U.B	*	*Clostridium perfringens*	63
7	MA3-0001U-M334-E06-U.G	*	*Brucella suis*	81
8	MG1-0036U-A253-B01-U.B	**	*Biomphalaria glabrata*	94
9	MS1-0049G-V346-C05-U.B		*Homo sapiens*	91
10	MS1-0094P-V340-F04-U.B	*	*Clostridium perfringens*	59
11	MA3-9999U-M329-F02-U.G		*Mus musculus*	93
12	MS1-0094G-V307-G04-U.B	*	*Fusobacterium nucleatum*	58
13	MG1-0067P-V303-B06-U.B	*	*Aster yellows phytoplasma*	58
14	MG1-0064T-D213-G10-U.G	*	*Thermotoga maritima*	54
15	ML1-0017T-R270-A12-U.G		*Rattus norvegicus*	62
16	MG1-0026U-V290-H12-U.B	*	*Streptomyces coelicolor*	69
17	MS1-0074T-L308-F06-U.B	*	*Homo sapiens*	100
18	MG1-0013G-V284-A01-U.B		*Homo sapiens*	60
19	ML1-0087T-R250-B03-U.G	*	*Streptococcus pneumoniae*	96
20	MA3-9999U-V238-G09-U.B	*	*Chlorobium tepidum*	46
21	MA3-0001U-L244-G08-U.B		*Mus musculus*	91
22	MA3-0001U-L243-D11-U.B		*Mus musculus*	83
23	ML1-0091T-L259-D04-U.B	*	*Microbulbifer degradans*	84
24	ML1-0007T-V221-H07-U.B		*Eremothecium gossypii*	44
25	MA3-9999U-L243-D11-U.G		*Mus musculus*	86

existentes. A metodologia baseada em similaridade não apresenta este tipo de problema.

A maior contribuição do QUEST na análise das SmAEs foi a classificação daquelas sem seqüências similares em bancos públicos, que poderiam ser tanto legítimas como contaminações. Essa classificação foi a primeira atribuída às SmAEs sem similares, e pode servir como base para a escolha de SmAEs para diferentes trabalhos.

A execução do QUEST com os reads classificados pelo projeto original como contaminantes revelou a presença de alguns possivelmente legítimos, e que podem portanto ter sido excluídos indevidamente. A análise através da similaridade no entanto discordou da classificação do QUEST em quase 50% dos casos trabalhados. A causa disto é que a classificação do projeto original foi bastante rigorosa, classificando como contaminantes apenas os reads com ótima similaridade contra um banco de seqüências possivelmente contaminantes, e que teve portanto uma pequena taxa de falsos negativos. Assim, o conjunto de possíveis legítimas criado pelo QUEST pode concentrar de fato a maioria das legítimas classificadas incorretamente pelo projeto, embora esse número seja pequeno comparado com o total de contaminantes.

5.3 Tempo de execução

Para avaliar o tempo de execução do QUEST, foram contabilizados os tempos de cada módulo (ver descrições no apêndice B) na classificação das 27.430 SmAEs do projeto EST de *S.mansoni*. Para os testes, foi utilizado um computador pessoal com um processador Pentium (2.4GHz), 256Mb de RAM e sistema operacional linux.

O tempo gasto por cada módulo é apresentado na Tabela 5.14. Apesar do QUEST usar linguagens de programação interpretadas (PERL e BASH), em geral mais lentas que as linguagens compiladas, o tempo de execução do QUEST é considerado razoável para o tipo de problema abordado. O QUEST gastou a maior parte das 9 horas de execução na extração de características (quase 95%), sendo os extratores GL e HN os mais lentos.

Para comparação com o tempo de execução da metodologia baseada em similaridade, as 27.430 SmAEs foram comparadas com o banco nr (copiado em 01/04/2004 de *http://www.ncbi.nlm.nih.gov*), contendo aproximadamente 2,7 milhões de seqüências. Utilizando o mesmo equipamento, a abordagem da similaridade levou por volta de quatro dias para terminar as análises, ou seja, demorou 10 vezes mais que o QUEST para o mesmo trabalho de classificação. Além disso, conta a favor do QUEST o fato dele poder ser reimplementado em linguagens mais velozes, e contra a metodologia baseada em similaridade o fato dos bancos públicos de seqüências terem crescimento muitas vezes exponencial.

Tabela 5.14: Tempos de execução em segundos de cada módulo do QUEST na classificação de 27.430 SmAEs, usando o classificador da votação. A descrição de cada módulo encontra-se no apêndice B. Os tempos dos módulos 1.2, 1.3, 1.4 e 1.5 são a soma dos tempos de seus módulos constituintes 1.x.y. O módulo principal (1) mostra a soma dos tempos dos módulos 1.2, 1.3, 1.4 e 1.5.

nº	módulo	tempo (s)
1	quest	32.664 ($\approx$ 9 horas)
1.1	load_pars	1
1.2	proc_sets	659
1.2.1	valid_seqs	26
1.2.2	rename_seqs	613
1.2.1	split_sets	20
1.3	programs	31.012
1.3.1	es_run	376
1.3.2	gl_run	6.376
1.3.3	binomial	136
1.3.4a	multinomial(DN)	4.976
1.3.4b	multinomial(TN)	4.928
1.3.5	hexamers	9.384
1.3.6	dn_signature	4.808
1.3.7	norm_caracs	28
1.4	classifiers	709
1.4.1	run_classifiers	475
1.4.2	gist_proc	234
1.5	classification	283
1.5.1	classify (votação)	283

Capítulo 6

Conclusão

A motivação para este trabalho surgiu das dificuldades encontradas no trato das contaminações no projeto EST de *Schistosoma mansoni*. Neste projeto, o autor se envolveu com vários dos processamentos aplicados às seqüências obtidas, como por exemplo:

- Adaptação das ferramentas para visualização de seqüências utilizadas em outros projetos [22].
- Montagem e execução de diversas ferramentas de análise sobre as seqüências geradas.
- Desenvolvimento de um sistema completo baseado em uma ontologia de genes [4] para categorização das funções das seqüências montadas, assim como a visualização e edição dessa categorização.
- Ajuda na detecção das contaminações presentes entre as seqüências do projeto.

O auxílio neste projeto resultou na co-autoria em um dos artigos resultantes deste projeto, publicado na revista *Nature Genetics* em 2003 [29]. Contudo, não houve tempo para a aplicação da metodologia desenvolvida no presente trabalho aos dados do projeto *S. mansoni* antes dessa publicação. Ainda assim, a metodologia é genérica o suficiente para ser aplicada diretamente a outros projetos EST, e pode ainda ser usada para outros problemas onde a identificação do organismo de origem de seqüências se faça necessária, como nos exemplos vistos anteriormente. As maiores contribuições desta dissertação são:

- Uma metodologia para se determinar a origem de seqüências de DNA alternativa à similaridade, contornando diversos de seus problemas.
- Uma metodologia que utiliza diversos extratores de características e classificadores. Não há conhecimento de outros trabalhos com essas qualidades.

Os resultados parciais deste trabalho foram apresentados e publicados nos anais do 2º Workshop Brasileiro sobre Bioinformática [23] (WOB, 2003). Esses resultados serão publicados também em uma futura edição da Revista Tecnologia da Informação (RTInfo). Além disso, este trabalho resultou em um artigo enviado ao décimo segundo *Intelligent Systems for Molecular Biology* (ISMB, 2004). Embora o artigo tenha sido rejeitado, os revisores consideraram o trabalho bom e promissor, apesar de terem considerado as taxas de erros ainda muito elevadas.

Apesar da metodologia ter resultado em valores de erros relativamente altos para o problema, ela provê a estrutura básica para que se consiga diminuir estes valores. Isso poderia ser conseguido com as seguintes melhorias:

- Desenvolvimento de uma maneira de combinar os resultados do QUEST com os resultados da classificação dada pela similaridade. Isso deve trazer grandes melhorias às classificações feitas.
- Uso de mais extratores de características mais efetivos, como o ESTSCAN.
- Uso de classificadores mais poderosos, como o baseado em SVM já utilizado neste trabalho.

Com a finalidade de melhorar a aplicabilidade do QUEST, alguns estudos adicionais poderiam ser feitos:

- Análise da viabilidade de se trabalhar com seqüências de qualquer tamanho, eliminando a restrição imposta atualmente pelo QUEST.
- Expansão do escopo para inclusão de análise de seqüências não codificantes
- Estudo da possibilidade de se usar um sistema de aprendizado não supervisionado, o que poderia eliminar a necessidade da definição prévia dos contaminantes, situação mais próxima da ideal para o problema.

O desempenho do QUEST poderia ser melhorado caso ele fosse reimplementado com o uso de linguagens compiladas. Como forma de melhorar sua usabilidade, poderia ser implementada também uma ferramenta para instalação automática.

Apêndice A

Interface do programa QUEST

A.1 Arquivos de configurações de projetos

Um arquivo de configurações do QUEST é um arquivo texto simples organizado em *seções*. Em cada arquivo deve haver ao menos três seções principais, nesta ordem:

1. uma seção *project*, contendo dados gerais do projeto.
2. uma seção com os dados do organismo alvo.
3. uma seção para cada organismo contaminante utilizado.

As seções são identificadas pelo seu nome entre colchetes, como por exemplo `[project]`. Cada seção contém um ou mais campos para sua definição, organizados na forma de um par `chave=valor`.

A seção *project* contém o campo *name*, obrigatório, contendo o nome do projeto. Se o projeto não for para testes, o campo *infile* indica o arquivo no formato *fasta* com as seqüências não rotuladas a serem classificadas.

As seções subseqüentes contêm informações sobre os organismos utilizados, contendo os mesmos campos tanto para o organismo alvo quanto para os contaminantes. A identificação da seção indica a maneira como o organismo será referenciado pelo QUEST. O campo *name* contém o nome ou qualquer outra informação sobre o organismo. O campo *infile* indica o arquivo (também no formato fasta) com as seqüências originárias do organismo desta seção. Os campos *ptrain*, *peval* e *ptest* (este último apenas no caso de projetos de teste) indicam como o arquivo de seqüências deve ser particionado para gerar os diferentes conjuntos de seqüências necessários ao QUEST. Se for passado um valor inteiro, a mesma quantidade de seqüências será utilizada; se for passado um valor com decimais, ele será interpretado como um percentual de seqüências a ser usado.

Um exemplo de arquivo de configurações para um projeto de teste para tentar separar seqüências de *Drosophila melanogaster* de outros dois contaminantes é dado a seguir.

```
[project]
name=Separar DM de CP e EC

[DM]
name=Drosophila melanogaster
infile=db/dm.fasta
ptrain=200
peval=200
ptest=5000

[EC]
name=Escherichia coli
infile=db/ec.fasta
ptrain=500
peval=100
ptest=100

[CP]
name=Clostridium perfrigens
infile=db/cp.fasta
ptrain=500
peval=100
ptest=100
```

Abaixo é dado um outro exemplo de arquivo de configurações para um projeto real. Notar a presença do campo *infile* na seção *project* e a ausência dos campos *ptrain* nas seções dos organismos, indicando tratar-se de um projeto real.

```
[project]
name=Schistosoma mansoni EST project
infile=db/sm.project.fasta

[SM]
name=Schistosoma mansoni
infile=db/sm.fasta
ptrain=150
```

```
peval=112

[XA]
name=Xanthomonas axonopodis pv. citri
infile=db/xa.fasta
ptrain=500
peval=100

[XF]
name=Xylella fastidiosa
infile=db/xf.fasta
ptrain=500
peval=100
```

A.2 Saída do programa para os projetos

A saída do programa QUEST é dividida em duas partes principais. A primeira parte (com linhas precedidas por `#`) contém informações gerais sobre o projeto e sobre o processo de classificação, tais como o nome do projeto, o classificador pedido pelo usuário, os programas geradores das características utilizadas pelo classificador, e as estimativas gerais de falsos e verdadeiros positivos e negativos.

A segunda parte é uma lista contendo para cada seqüência analisada a sua identificação (pe, `C600002.1`), a sua classificação (`_leg_` significa proveniente do organismo alvo e `_cont_` contaminação), um valor de confiabilidade para a predição (quanto maior mais confiável, usado para comparação entre duas ou mais seqüências dessa mesma lista), e para os classificadores não paramétricos o organismo de origem mais provável para a seqüência correspondente. Um exemplo de saída é dado abaixo.

```
# QUEST v1.0
# Projeto: Schistosoma mansoni EST project
# Classificador: Soma
# Programas: ES,GL,TN,BN,DS
# Estimativas: pTP=87,50%, pFP=12,50%, pTN=97,57%, pFN=2,43%
C600002.1       _leg_   30,98   SM
C600003.1       _leg_   59,01   SM
C600004.1       _leg_   59,07   SM
C600005.1       _leg_   32,39   SM
C600006.1       _leg_   52,84   SM
```

```
C600007.1       _cont_  1,26    MM
C600008.1       _cont_  1,70    BS
C600009.1       _cont_  5,03    BS
C600010.1       _cont_  1,76    MM
C600011.1       _leg_   13,42   SM
C600012.1       _leg_   66,89   SM
C600013.1       _leg_   47,57   SM
C600014.1       _leg_   29,08   SM
C600015.1       _leg_   22,44   SM
C600016.1       _leg_   15,65   SM
C600017.1       _leg_   3,98    SM
C600018.1       _leg_   16,11   SM
C600020.1       _leg_   36,95   SM
...
```

Apêndice B

Implementação e disponibilidade do QUEST

O QUEST foi implementado inteiramente através de scripts PERL (versão 5.8.0) e BASH (versão 2.00), rodando sobre sistemas operacionais linux e unix. A escolha do PERL e BASH deve-se principalmente aos recursos oferecidos por essas linguagens e pela facilidade na implementação quando comparadas com linguagens como C. Para execução do QUEST são também necessários os programas ESTSCAN e GLIMMER. Eles podem ser obtidos gratuitamente nos endereços:

http://www.isrec.isb-sib.ch/ftp-server/ESTScan
ftp://ftp.tigr.org/pub/software/Glimmer

O QUEST está disponível através de solicitação direta aos autores deste trabalho. Embora não haja uma rotina específica para sua instalação, esta resume-se a uma cópia dos scripts necessários, instalação do ESTSCAN e GLIMMER e pequenas modificações no arquivo de configurações gerais do QUEST. O QUEST executa apenas sobre sistemas operacionais aparentados do unix e que possuam PERL e BASH instalados. A Tabela B.1 lista os principais módulos desenvolvidos.

Tabela B.1: Detalhes dos principais módulos desenvolvidos.

nº	módulo	linguagem	linhas	descrição
1	quest	bash	59	módulo principal
1.1	load_pars	bash	229	carga dos parâmetros gerais e do projeto
1.2	proc_sets	bash	229	pré-processamento dos conjuntos de seqüências dos organismos
1.2.1	valid_seqs	perl	106	separação das seqüências válidas
1.2.2	rename_seqs	perl	133	renomeia seqüências para padrão do QUEST
1.2.1	split_sets	perl	221	divisão entre os conjuntos de treino, avaliação e produção
1.3	programs	bash	250	execução dos programas extratores de características
1.3.1	es_run	perl	129	executa o programa ES
1.3.2	gl_run	perl	160	executa o programa GL
1.3.3	binomial	perl	149	implementação de BN
1.3.4	multinomial	perl	361	implementação de DN e TN
1.3.5	hexamers	perl	339	implementação de HN
1.3.6	dn_signature	perl	317	implementação de DS
1.3.7	norm_caracs	perl	33	normaliza os valores das características
1.4	classifiers	bash	170	exibição dos erros gerais para todos os classificadores
1.4.1	run_classifiers	perl	1.350	escolha do subconjunto de características e execução de cada classificador
1.4.2	gist_proc	perl	86	transforma saída do GIST para o padrão do QUEST
1.5	classification	bash	84	inicia classificação do conjunto de produção
1.5.1	classify	perl	160	classifica conjunto de produção

Referências Bibliográficas

[1] MD Adams, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, AR Kerlavage, WR McCombie, and JC Venter. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252:1651–6, 1991.

[2] JE Allen, Pertea M, and Salzberg SL. Computational gene prediction using multiple sources of evidence. *Genome Research*, 14:142–8, 2004.

[3] SF Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–402, 1997.

[4] M Ashburner, CA Ball, JA Blake, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9, 2000.

[5] P Baldi, S Brunak, and Chauvin Y. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[6] DA Benson, I Karsch-Mizrachi, DJ Lipman, J Ostell, and DL Wheeler. GenBank: update. *Nucleic Acids Research*, 32(90001):D23–26, 2004.

[7] TA Brown. *Genomes*. John Wiley and Sons, New York, first edition, 1999.

[8] C Cortes and V Vapnik. Support vector networks. *Machine Learning*, 20, 1995.

[9] AL Delcher, D Harmon, S Kasif, O White, and S Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–41, 1999.

[10] E Dias-Neto et al. Shotgun sequencing of the human transcriptome with ORF expressed sequences tags. *Proc. Natl. Acad. Sci. USA*, 97:3491–96, 2000.

[11] RO Duda, PE Hart, and DG Stork. *Pattern classification*. John Wiley and Sons, second edition, 2000.

[12] WJ Ewens and GR Grant. *Statistical methods in bioinformatics*. Springer, New York, 2001.

[13] P Green. Phrap and Cross match Documentation. *www.phrap.org*, 1994.

[14] PT Hraber and JW Weller. On the species of origin: Diagnosing the source of symbiotic transcripts. *Genome Biology*, 2:research0037.1–0037.14, 2001.

[15] X Huang and A Madan. CAP3: A DNA sequence assembly program. *Genome Research*, 9:868–77, 1999.

[16] C Iseli, CV Jongeneel, and P Bucher. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol.*, pages 138–48, 1999.

[17] S Karlin. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology*, 9:335–43, 2001.

[18] L Lam and CY Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, (16):945–54, 1995.

[19] B Lewin. *Genes VII*. Oxford University Press, New York, 2000.

[20] F Mignone, C Gissi, S Liuni, and G Pesole. Untranslated regions of mRNAs. *Genome Biology*, 3:reviews0004.1–0004.10, 2002.

[21] A Muto and S Osawa. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA*, 84:166–69, 1987.

[22] VK Okura. Bioinformática de projetos genoma de bactérias. Master's thesis, UNICAMP, fev 2002.

[23] JP Piazza and JC Setubal. New ways for automatic detection of contaminants in EST projects. *Proc. 2nd Brazilian Workshop on Bioinformatics (WOB)*, 2003.

[24] F Roli and G Giacinto. Design of multiple classifier systems. In H Bunke and A Kandel, editors, *Hybrid methods in pattern recognition*. World Scientific, 2002.

[25] JC Setubal and J Meidanis. *Introduction to Computational Molecular Biology*. PWS, Boston, 1997.

[26] S Theodoridis and K Koutroumbas. *Pattern recognition*. Academic Press, San Diego, first edition, 1999.

[27] M van Erp, L Vuurpijl, and L Schomaker. An overview and comparison of voting methods for pattern recognition. *Proc. 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2002.

[28] JC Venter, K Remington, JF Heidelberg, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66–73, 2004.

[29] S Verjovski-Almeida, R DeMarco, ALE Martins, PEM Guimarães, EPB Ojopi, ACM Paquola, JP Piazza, MY Nishiyama Jr., JPFW Kitajima, et al. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. *Nature Genetics*, 35(2), 2003.

[30] O White, T Dunning, G Sutton, M Adams, JC Venter, and C Fields. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research*, 21(16):3829–3838, 1993.