XVII CIC
X ENPOS
Conhecimento sem fronteiras
XVII Congresso de Iniciação Científica
X Encontro de Pós-Graduação
11, 12, 13 e 14 de novembro de 2008

SSRLOCATOR – FERRAMENTA PARA DETECÇÃO *IN SILICO* DE *SIMPLE SEQUENCE REPEATS* INTEGRADA COM DESENHO DE *PRIMERS* E SIMULAÇÃO DE PCR EM BANCO DE DADOS GENOMICOS

Luciano Carlos da Maia[1], Darío Abel Palmieri[2], Velci Queiroz de Souza[3], Mauricio Marini Kopp[4], Fernando Irajá Félix de Carvalho[5] e Antonio Costa de Oliveira[5]

[1]Engenheiro Agrônomo, Aluno de Doutorado - Centro de Genômica e Fitomelhoramento – FAEM/UFPel;
[2]Geneticista, Doutor em Genética - Pesquisador Departamento de Ciências Biológicas – FCL - UNESP - Campus Assis; [3]Engenheiro Agrônomo, Doutor em Agronomia - Professor Adjunto Departamento de Agronomia CESNORS/FW – UFSM;
[4]Engenheiro Agrônomo, Doutor em Agronomia -Pesquisador da Embrapa Gado de Leite, Laboratório de Biotecnologia e Fisiologia Vegetal- Juiz de Fora-MG;
[5]Engenheiro Agrônomo, PhD em Genética - Professor Adjunto Departamento de Fitotecnia - Centro de Genomica e Fitomelhoramento – FAEM/UFPel.

Introdução

Microssatélites ou SSRs (*Simple sequence repeat*) são seqüências formadas pela repetição em série de nucleotídeos dispostos em arranjos contendo entre um e seis pares de bases (Morgante & Olivieri, 1993). As regiões de DNA repetitivo estão mais propensas à ocorrência de laço ou estruturas conhecidas como *hairpins*, pois, nestes trechos, durante a ação de replicação, a *DNA polimerase* sofre um escorregão (*slippage*), provocando inserção ou deleção de pares de bases e promovendo dessa forma, um aumento ou redução no tamanho da seqüência de repetição (Lyer *et al*. 2000). Estes locos podem ser encontrados em seqüências gênicas (CDS, 5'-UTR, 3'-UTR e íntrons) e intergênicas, podendo conduzir a um ganho ou perda de função dos genes (Li *et al*. 2004). Esta classe de marcadores é poderosa em diversas aplicações da genética e melhoramento, devido a sua reprodutibilidade, natureza multi-alélica, característica co-dominante e abundância genômica, proporcionando aos geneticistas e melhoristas uma estratégia para ligar variações do genótipo às variações fenotípicas (Varshney *et al*. 2005). Os primeiros protocolos para o isolamento microssatélites demandavam intenso trabalho de laboratório, entretanto, com o acúmulo de informações geradas por inúmeras iniciativas de seqüenciamento de genomas no mundo inteiro, a utilização da bioinformática facilitou a obtenção de novos microssatélites (Maia *et al*. 2008). Nos primeiros trabalhos com identificação *in silico* de microssatélites foram utilizados os pacotes FASTA e BLAST, seguidos pelos programas SPUTNICK, REPEATMASKER, TRF-Tandem Repeat Find e TROLL. Posteriormente foram desenvolvidos em linguagem Perl os scripts MISA e SSRIT (Maia *et al*. 2008). A utilização destes programas geralmente necessita da conexão com outros programas para o desenho de *primers*, e, numa terceira etapa, a conexão com o e-PCR (Schuler 1997), para verificar a redundância dos *primers*. Os procedimentos que integram as etapas e os vários programas são denominados de *pipeline* e, em muitas ocasiões, o uso dessas rotinas são pouco acessíveis para biólogos pouco experientes em computação (Maia *et al*. 2008).

O presente trabalho teve como objetivo desenvolver uma ferramenta computacional em bioinformática que integre as funções de localização e caracterização de microssatélites, desenho de *primers* específicos para cada loco identificado, simulação da reação de PCR, amplificando da lista de *primers* desenhados contra diferentes arquivos fasta, alinhamento global entre os *amplicons* gerados no PCR virtual e cálculo do *score* e identidade do alinhamento global dos *amplicons*.

METODOLOGIA

Busca por SSRs

O módulo foi escrito em linguagem Perl e consiste na geração de uma matriz, que, a partir da combinatória entre A, T, C e G, cria todos os possíveis arranjos compostos entre 1 e 10 nucleotídeos. O algoritmo utilizado foi uma adaptação do algoritmo aplicado nos *scripts* MISA (Thiel *et al*. 2003) e SSRIT (Temnykh *et al*. 2001).

Desenho de Primers

Um algoritmo escrito em linguagem *Delphi®,* conecta o programa Primer3 (Rozen & Skaletsky, 2000), responsável pelo desenho dos *primers* e os resultados são gravados em um banco de dados *FireBird*, onde são gravadas as identificações de cada loco microssatélite, as seqüências dos *primers foward* e *reverse* e o trecho do DNA contido entre o conjunto de *primers* (*amplicon* origem).

Virtual-PCR

O módulo para simular a reação de PCR (*Polymerase chain reaction*) foi escrito em *Delphi*®. O algoritmo utiliza dados do módulo anterior (locus microssatélite, *primer foward, primer reverse e amplicom origem*), seguido pela busca de *contigs* contendo sítios de ligação dos *primers foward* e *reverse*. Quando são encontrados sítios de ancoragem o trecho de DNA abrangendo desde a posição de inicio do *primer foward* até a posição terminal do *primer reverse* é copiada para uma variável chamada *amplicon homólogo*.

Alinhamento global

Para o alinhamento global entre o *amplicon* de origem e o *amplicon* homólogo foi utilizado o algoritmo descrito por Needleman & Wunsch (1970) e Smith & Waterman (1981). Ainda, no mesmo módulo, é efetuado o cálculo da identidade entre os *amplicons* alinhados, conforme descrito por Waterman (1994) e Vingron & Waterman (1994).

Validação - Seqüências para análise

Foram utilizadas 28.469 seqüências completas de *cDNA* (*full length-cDNA*) de arroz (*Oryza sativa* ssp.-cv. Nipponbare), seqüenciados pelo *The Rice Full-Length cDNA Consortium*, não redundantes e mapeados nos bancos de dados derivados do seqüenciamento das subespécies japonica (japonica draft genome, BAC/PAC clones - IRGSP) e indica (*indica draft genome*) (Kikush *et al*. 2003).

RESULTADOS E DISCUSSÃO

Os resultados encontrados revelaram que, das 28.469 seqüências de cDNA analisadas, 3.899 apresentaram a ocorrência de microssatélites, o que corresponde a 13.60% das seqüências. Na Tabela 1 é mostrado o total de ocorrências para cada um dos motivos analisados (monômeros, decâmeros, trímeros, tetrâmeros, pentâmeros, hexâmeros, heptâmeros, hexâmeros, nonâmeros e decâmeros) e a porcentagem correspondente para cada grupo. Os resultados para desenho de *primers* mostrados na Tabela 2 indicam que dos 3.899 locos microssatélites localizados, somente 3.399 dessas regiões possibilitaram o desenho de *primers*.

Tabela 1. Total de ocorrências e porcentagem de microssatélites, distribuídos para cada um dos motivos analisados (Maia *et al*. 2008).

N-mer	Mono	Di	Tri	Tetra	Penta	Hexa	Hepta	Octa	Nona	Deca	Total
Ocorrências	124	596	1994	251	426	390	82	6	25	5	3,899
%	3.18	15.29	51.14	6.44	10.93	10.00	2.10	0.5	0.64	0.13	-

Tabela 2. Distribuição do número de cDNAs, número de microssatélites e *primers* localizados (Maia *et al*. 2008).

Descrição							**Total**
N°SSRs	1	2	3	4	5	6	**-**
N°Primers	1	2	3	4	5	6	**-**
N°cDNAs	2.813	242	29	1	1	1	**3,087**
Total *Primers*	**2.813**	**484**	**87**	**4**	**5**	**6**	**3,399**
%	**82,76**	**14,24**	**2,56**	**0,12**	**0,15**	**0,18**	**-**

Simulação da Virtual-PCR - Análise da redundância de *primers*

Na simulação do PCR virtual um total de 4.744 fragmentos foram amplificados, dos quais, 3.399 *amplicons* foram gerados nas regiões de origem dos *primers* e 1.345 *amplicons* foram gerados em outras regiões diferentes, onde estes conjuntos de *primers* encontraram sítios de ligação, gerando *amplicons rundantes*.

CONCLUSÃO

Foi desenvolvido com sucesso um programa denominado SSRLocate, por meio do qual foi possível a sistematização dos algoritmos conectando as etapas de 1) localização dos microssatélites, 2) desenho de *primers*, 3) simulação da PCR entre a lista de *primers* derivados dos loci microssatélites e as seqüências contidas nos outros arquivos fasta, 4) obtenção dos diversos relatórios com os tipos e freqüências de cada um dos motivos identificados, lista de *primers* desenhados com todas as informações padrões, relatórios sobre a PCR, alinhamento global entre as seqüências e redundâncias de *primers* e *amplicons*.

Finalmente, sugerimos o uso desta ferramenta como uma nova estratégia no contexto da mineração (*data mining*) de *primers* derivados de locos microssatélites, sejam eles oriundos de seqüências genômicas ou de regiões expressas (ESTs/cDNAs). Apontamos ainda, uma segunda estratégia para a utilização deste programa, onde ele poderá ser utilizado na localização de *primers*-microssatélites no banco de seqüências de uma espécie e prever a transferibilidade destes através da simulação da PCR (*Virtual-PCR*) contra bancos de outras espécies, fornecendo resultados da transposição de marcadores, ancorando conjuntos de *primers* em regiões ortólogas (*amplicons* homólogos com identidade igual a 100) ou regiões parálogas (*amplicons* homólogos com diferentes níveis consideráveis de identidade).

Referências bibliograficas:

Iyer RR, Pluciennik A, Rosche WA, Sinden RR, Wells RD (2000) DNA polymerase III proofreading mutants enhance the expansion and deletion of triplet repeat sequences in Escherichia coli. **The Journal of biological chemistry** 21: 2174-2184

Kikuchi S et al. (2003) Collection, Mapping, and Annotation of over 28,000 cDNA Clones from japonica Rice - The Rice Full-Length cDNA Consortium. **Science** 301: 376-379

Li B, Xia Q, Lu C, Zhou Z, Xiang Z (2004) Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. **Genomics Proteomics Bioinformatics** 2: 24-31

Maia LC, Palmieri DA, Souza VQ, Kopp MM, Carvalho FIF e Costa de Oliveira, A (2008) SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. **Int J Plant Genomics**. 412696.

Morgante M and Olivieri AM (1993) PCR-amplified micro satellites as markers in plant genetics. **The Plant Journal** 3: 175-182

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology** 48: 443-453

Rozen S and Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. **Methods in molecular biology** 132: 365-386
Schuler GD (1997) Sequence mapping by electronic PCR. **Genome Research** 7: 541-550

Smith TF and Waterman MS (1981) Identification of Common Molecular Subsequences. **Journal of Molecular Biology** 147: 195-197

Temnykh S, Declerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. **Genome Research** 11: 1441-1452

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). **TAG Theoretical and Applied Genetics** 106: 411-422

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. **Trends in Biotechnology** 23: 48-55

Vingron M and Waterman MS (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. **Journal of molecular biology** 7: 1-12

Waterman M (1994) Estimating statistical significance of sequence alignments, Philosophical transactions of the Royal Society of London. **Biological sciences** 29: 383-390