

ISSN 1677-8464

VPRO - Um Identificador de Padrões de Seqüências

Edgard Henrique dos Santos¹
Paula Regina Kuser Falcão²
Goran Neshich³

Quando se deseja identificar a função, homólogos ou a família de uma proteína é essencial o uso de informações contidas no banco de dados de padrões PROSITE (Hulo et al., 2006).

O banco de dados PROSITE consiste em uma grande coleção de padrões com significado biológico que são descritos como expressões regulares usados para a detecção de seqüências específicas e também genéricas, como uma matriz de distribuição de nucleotídeos para a busca de grandes domínios. Todos os padrões são construídos usando alinhamentos feitos manualmente e possuem uma extensiva documentação sobre funções, ocorrências taxonômicas entre outras informações. Os padrões do PROSITE fazem parte do processo de anotação de proteínas da UniProtKB/Swiss-Prot (Swiss Institute of Bioinformatics, 2006).

Para cada grupo de informação do PROSITE existe um template de regras chamado ProRule (de Castro et al., 2006). Essas regras são usadas internamente pelo grupo Swiss-Prot para automatizar a inclusão de domínios no PROSITE. As regras definem o significado biológico da informação específica dos resíduos com

seus respectivos domínios. Essa informação, derivada do mapeamento de resíduos significativos dentro de uma seqüência, é encontrada na forma de blocos de dados que, em conjunto com o padrões na forma de expressões regulares, permitem a identificação de características de intra-domínios, como sítios ativos, locais de ligação ou pontes de dissulfeto.

Como forma de facilitar a análise de seqüências do banco de dados de estruturas PDB que possuam padrões identificados no banco de dados PROSITE, foi criado pelo grupo de Bioinformática, da Embrapa Informática Agropecuária (2006), o programa VPRO.

Programa VPRO

O VPRO foi desenvolvido em Java (Sun Microsystems, 2007) e faz parte do contexto das rotinas de cálculo de parâmetros estruturais de proteínas do Sting (Higa et al., 2003).

A necessidade de se obter ganho computacional e a iniciativa da transição das rotinas de controle de execução de cálculo, originalmente em Perl (Perl

¹ Bacharel em Ciência da Computação, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: edgard@cnptia.embrapa.br)

² Ph.D. em física Aplicada - Cristalografia de Proteínas, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: paula@cnptia.embrapa.br)

³ Ph.D. em Biofísica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (E-mail: neshich@cnptia.embrapa.br)

Foundation, 2006), para Java, foram os principais motivos para a criação do VPRO.

O programa transforma seqüências de aminoácidos encontradas nos arquivos PDB para o formato FASTA (Fasta..., 2006) e as analisa relativamente a cada padrão do PROSITE, gerando como saída os dados posicionais da seqüência correspondentes.

As seqüências de proteínas do arquivo PDB são transformadas para o formato FASTA para facilitar a comparação dos padrões com as expressões regulares do PROSITE.

Como exemplo mostra-se parte do arquivo PDB 1C30:

```

CISPEP 23 ARG G 998 PRO G 999 0 -0.35
CISPEP 24 SER H 357 PRO H 358 0 2.31
CRYST1 152.600 164.600 332.700 90.00 90.00 90.00 P 21 21 21 16
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.006553 0.000000 0.000000 0.000000
SCALE2 0.000000 0.006075 0.000000 0.000000
SCALE3 0.000000 0.000000 0.003006 0.000000
A T O M 1 N M E T A 1 28.271 23.163 102.440 1.00 85.67
A T O M 2 C A M E T A 1 29.110 21.961 102.260 1.00 100.00
A T O M 3 C M E T A 1 30.268 22.085 101.213 1.00 100.00
A T O M 4 O M E T A 1 30.124 22.736 100.163 1.00 100.00
A T O M 5 C B M E T A 1 28.252 20.663 102.118 1.00 100.00
A T O M 6 C G M E T A 1 28.553 19.553 103.150 1.00 100.00
A T O M 7 S D M E T A 1 27.139 18.847 104.095 1.00 100.00
A T O M 8 C E M E T A 1 26.012 18.235 102.790 1.00 100.00
A T O M 9 N P R O A 2 31.430 21.446 101.520 1.00 100.00
A T O M 10 C A P R O A 2 32.638 21.436 100.681 1.00 33.21
A T O M 11 C P R O A 2 32.577 20.330 99.622 1.00 44.21
A T O M 12 O P R O A 2 31.506 19.784 99.281 1.00 55.17
A T O M 13 C B P R O A 2 33.795 21.158 101.642 1.00 50.30
A T O M 14 C G P R O A 2 33.178 20.575 102.914 1.00 72.34
A T O M 15 C D P R O A 2 31.682 20.842 102.860 1.00 65.46
A T O M 16 N L Y S A 3 33.732 19.943 99.100 1.00 32.46
A T O M 17 C A L Y S A 3 33.681 18.934 98.073 1.00 42.71
A T O M 18 C L Y S A 3 33.096 17.595 98.416 1.00 31.22
A T O M 19 O L Y S A 3 33.140 17.105 99.539 1.00 68.37
A T O M 20 C B L Y S A 3 34.955 18.759 97.300 1.00 57.99
A T O M 21 C G L Y S A 3 35.933 17.904 98.058 1.00 70.49
A T O M 22 C D L Y S A 3 37.360 18.029 97.539 1.00 100.00
A T O M 23 C E L Y S A 3 37.907 16.738 96.925 1.00 100.00
A T O M 24 N Z L Y S A 3 39.309 16.428 97.264 1.00 100.00
A T O M 25 N A R G A 4 32.564 17.016 97.352 1.00 62.09
A T O M 26 C A A R G A 4 31.966 15.719 97.372 1.00 32.33
A T O M 27 C A R G A 4 33.098 14.682 97.480 1.00 100.00
A T O M 28 O A R G A 4 34.191 14.775 96.889 1.00 41.30
A T O M 29 C B A R G A 4 31.052 15.452 96.165 1.00 82.70
A T O M 30 C G A R G A 4 29.559 15.660 96.450 1.00 37.99
A T O M 31 C D A R G A 4 28.629 15.420 95.246 1.00 41.55
A T O M 32 N E A R G A 4 27.981 14.099 95.194 1.00 29.84
...

```

A primeira coluna, cujos valores são ATOM, contém em cada linha informações sobre seqüências e cadeias da proteína 1C30. No formato FASTA esta informação se apresenta da seguinte forma:

```

>CHAINA - (CRYSTAL STRUCTURE OF CARBAMOYL PHOSPHATE SYNTHETASE: SMALL 2 SUBUNIT MUTATION C269S)
MPKRTDIKSILILGAGPIVIGQACEFDYSGAQACKALREEGYRVILVNSNPATIMTDPEDATYIEPIHWEVVRKIIKERPDVAVLPTM
GGQATALNCALELERQGVLEEFVMTMIGATADADAIDKAEDRRRFVAMKKIGLETARSGIAHTMEEALVAADVGFPCIIHRSFTMGGS
GGGIAYNREEFEEICARGLDLSPTKELLIDESLIGWKEYEMEVVRDKNDCIIVCSIENFDAMGIHTGDSITVAPAQTLTDKEYQIMRN
ASMAVLREIGVETGGSNVQFAVNPKNRGLIVIMNPRVSRSSALASKATGFPIAKVAAKLAVGYTLDELMNDITGGRTPASFEPSIDYV
VTKIPRFNFEKFAANDRLTTQMKSVEVMAIGRTQQESLQKALRGLVGGATGFDPKVSLDDPEALTKIRRELKDGADRIWYIADA
FRAGLSVDGVFNLTNIDRWFLVQIEELVRLEEKVAEVGITGLNADFLRQLKRKGFADARL
...

```

Após comparação com os dados do PROSITE, explicada mais detalhadamente adiante, tem-se a saída apresentada a seguir:

```

1C30
A 014 024 PS00013
A 164 178 PS00866
A 297 304 PS00867
A 839 846 PS00867
C 014 024 PS00013
C 164 178 PS00866
C 297 304 PS00867
C 839 846 PS00867
E 014 024 PS00013
E 164 178 PS00866
E 297 304 PS00867
E 839 846 PS00867
G 014 024 PS00013
G 164 178 PS00866
G 297 304 PS00867
G 839 846 PS00867

```

A primeira linha mostra a cadeia e em seguida as posições inicial e final e o código descritivo do PROSITE. Este código faz parte do início do grupo de informações sobre família e domínio de proteínas dentro do arquivo prosite.dat. Os grupos são separados por linhas contendo duas barras. A primeira coluna contém para cada linha dois caracteres que identificam as informações descritas na segunda coluna de dados:

```

//
ID PROKAR_NTER_METHYL; PATTERN.
AC PS50075;
DT NOV-1997 (CREATED); NOV-1997 (DATA UPDATE); DEC-2005 (INFO UPDATE).
DE Acyl carrier protein phosphopantetheine domain profile.
MA /GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=71;
MA /DISJOINT: DEFINITION=PROTECT; N1=6; N2=66;
MA /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=2.3; R2=.02281121; TEXT='-LogE';
MA /CUT_OFF: LEVEL=0; SCORE=271; N_SCORE=8.5; MODE=1; TEXT='!';
MA /CUT_OFF: LEVEL=-1; SCORE=184; N_SCORE=6.5; MODE=1; TEXT='?';
MA /DEFAULT: D=-20; I=-20; B1=-80; E1=-80; MI=-105; MD=-105; IM=-105; DM=-105; MM=1; M0=-1;
MA /I: B1=0; B1=-105; BD=-105;
MA /M: SY='T'; M=-5,-15,-20,-17,-12,-10,-22,-18,2,-13,-1,0,-13,-6,-10,-13,-5,4,1,-23,-9,-12;
MA /M: SY='E'; M=-6,-6,-22,-6,9,-13,-21,-9,-11,0,-8,-7,-7,-13,1,1,-4,-3,-8,-24,-10,4;
MA /M: SY='E'; M=-5,-9,-24,11,15,-24,-12,-3,-23,3,-20,-15,6,-9,5,1,4,-2,-19,-29,-16,9;
MA /M: SY='E'; M=-5,2,-26,4,8,-22,-13,-7,-21,7,-17,-12,0,-13,3,7,-2,-6,-16,-22,-12,5;
MA /M: SY='L'; M=-6,-27,-19,-30,-23,4,-30,-23,26,-25,28,17,-25,-27,-21,-20,-19,-5,23,-23,-3,-23;
MA /M: SY='R'; M=-3,-10,-10,-11,2,-16,-19,-11,-13,-1,-8,-7,-8,-17,-1,3,-5,-6,-9,-26,-13,-1;
MA /M: SY='E'; M=-1,3,-23,4,9,-24,-11,-7,-22,8,-19,-13,2,-11,5,6,2,-2,-17,-26,-15,7;
...
DR Q12397, STCA_EMENI, T; P09095, TYCA_BREPA, T; O30408, TYCB_BREPA, T;
DR O30409, TYCC_BREPA, T; P74965, VENB_VIBVU, T; O07900, VIBB_VIBCH, T;
DR Q03149, WA_EMENI, T;
DR P80916, ACP_ACICA, P; P80917, ACP_ALCFA, P; P80919, ACP_ERYLO, P;
DR Q8GR69, DLTC_ABIDE, N; Q81T99, DLTC_BACAN, N; P61398, DLTC_BACC1, N;
DR Q81G41, DLTC_BACCR, N; Q63E04, DLTC_BACCZ, N; Q6HLH9, DLTC_BACHK, N;
DR P39579, DLTC_BACSU, N; P61399, DLTC_LACJO, N; Q9CG51, DLTC_LACLA, N;
DR Q92D49, DLTC_LISIN, N; Q5HHF0, DLTC_STAAC, N; P0A018, DLTC_STAAM, N;
DR P0A019, DLTC_STAAAN, N; Q6GIF4, DLTC_STAAR, N; Q6GAZ2, DLTC_STAAS, N;
DR P0A021, DLTC_STAAU, N; P0A020, DLTC_STAAW, N; Q5HQM8, DLTC_STAEQ, N;
DR Q8CPW0, DLTC_STAES, N; Q4L4U7, DLTC_STAHJ, N; Q49W73, DLTC_STAS1, N;
DR Q9X2N6, DLTC_STAXY, N;
DR Q9CIS9, CBIO1_LACLA, F;
3D 1ACP; 1AF8; 1DNY; 1DV5; 1F80; 1HQB; 1HY8; 1KLP; 1LOH; 1L0I; 1N8L; 1NNZ;
3D 1NQ4; 1OP7; 1T8K; 2AF8;
DO PDOC00012;
//
ID PROKAR_LIPOPROTEIN; RULE.
AC PS00013;
DT APR-1990 (CREATED); JUL-2003 (DATA UPDATE); DEC-2005 (INFO UPDATE).
DE Prokaryotic membrane lipoprotein lipid attachment site.
PA {DERK}(6)-[LIVMFVWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C.
RU Additional rules:
RU (1) The sequence must start with Met.
RU (2) The cysteine must be between positions 15 and 35 of the sequence in
RU consideration.
RU (3) There must be at least one charged residue (Lys or Arg) in the first
RU seven residues of the sequence.

```

```

NR /RELEASE=48.8.205780;
NR /TOTAL=1344(1332); /POSITIVE=945(937); /UNKNOWN=119(117);
NR /FALSE_POS=280(278); /FALSE_NEG=49; /PARTIAL=4;
CC /TAXO-RANGE=AB?P?; /MAX-REPEAT=2;
CC /SITE=5,lipid;
CC /VERSION=1;
DR P50927, 17KD_RICAM , T; P50928, 17KD_RICAU , T; P0A3N4, 17KD_RICCN , T;
DR Q52764, 17KD_RICJA , T; P50929, 17KD_RICMO , T; P50930, 17KD_RICPA , T;
DR P16624, 17KD_RICPR , T; P50931, 17KD_RICRH , T; P0A3N5, 17KD_RICRI , T;
DR P22882, 17KD_RICTY , T; O83142, 5NTD_TREPA , T; Q9KQ30, 5NTD_VIBCH , T;
DR P22848, 5NTD_VIBPA , T; Q8DFG4, 5NTD_VIBVU , T; Q46837, ACFD_ECOLI , T;
DR Q9KTQ4, ACFD_VIBCH , T; P0AE07, ACRA_ECO57 , T; P0AE06, ACRA_ECOLI , T;
DR P24180, ACRE_ECOLI , T; O05703, ADCA_STRPN , T; Q8CWN2, ADCA_STRR6 , T;
DR P94202, ALGK_AZOVI , T; P96956, ALGK_PSEAE , T; Q88NC7, ALGK_PSEPK , T;
...

```

A linha AC (ACcession number) mostra o número de inserção associado a cada grupo, ou código descritivo. O formato é sempre:

AC Psnnnnn;

Onde 'PS' se refere a PROSITE e 'nnnnn' um número com cinco dígitos.

A linha PA (PAttern) contém a definição de um padrão PROSITE. Os padrões são descritos usando as convenções do Swiss-Prot group ([http://ca.expasy.org/prosite/prosuser.hAla-any-\[Ser or Thr\]-\[Ser or Thr\]-\(any or none\)-Vatml#convent32](http://ca.expasy.org/prosite/prosuser.hAla-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Vatml#convent32))

Um exemplo de padrão:

PA <A-x-[ST](2)-x(0,1)-V.

Este padrão, que precisa estar em um N-terminal (a extremidade de uma proteína ou polipeptídeo terminada por um aminoácido com um grupo amino livre) de uma seqüência (representado pelo sinal '<'), é traduzido como: Ala-qualquer-[Ser or Thr]-[Ser or Thr]-(qualquer ou nenhuma)-Val

A linha RU (Rules) contém regras adicionais que não são possíveis de descrever apenas com expressões regulares, mas que devem ser observadas quando os padrões são comparados.

O trecho de código a seguir mostra um exemplo de regra:

```

1: String ps13rule = "[^DERK]{6}[LIVMFWSTAG]{2}[LIVMFYSTAGCQ][AGS]C";
2: String part;

3: regex = Pattern.compile(ps13rule);

4: matcher = regex.matcher(seq);

5:     if (matcher.find()) {

6:         part = seq.substring(0, matcher.start());

7:         if ((matcher.end() >= 15) && (matcher.end() <= 35)) {

8:             regex = Pattern.compile("[^M][A-Z]*[KR][A-Z]*");

```

Na linha 1, a regra obtida do PROSITE é uma expressão regular. Caso essa expressão seja encontrada na seqüência (linha 5), a análise passa a um próximo nível de verificação (linha 7) e assim por diante.

O VPRO usa as linhas PA e RU do PROSITE como base para a análise das seqüências, juntamente com os aminoácidos no formato FASTA.

As regras e as expressões regulares são traduzidas para a sintaxe do Java. Essas informações são previamente inseridas na memória pelo programa para reduzir o tempo de processamento. Em seguida é feita a comparação das seqüências por cadeia de cada um dos 41.000 arquivos PDB com as regras do PROSITE (PA e RU).

A forma simplificada do algoritmo:

```

abre o PROSITE;
carrega-o na memória;

para cada arquivo pdb {
    para cada cadeia {
        transforma sequencia para formato FASTA;
        verifica_regras- >compara_sequencia_PROSITE(sequencia fasta,prosite hash);
    }
}

gera saida em arquivo text

```

Posteriormente essa informação pode ser obtida pelo STING no módulo JavaProteinDossier (Neshich et al., 2004) (Fig. 1 e 2).

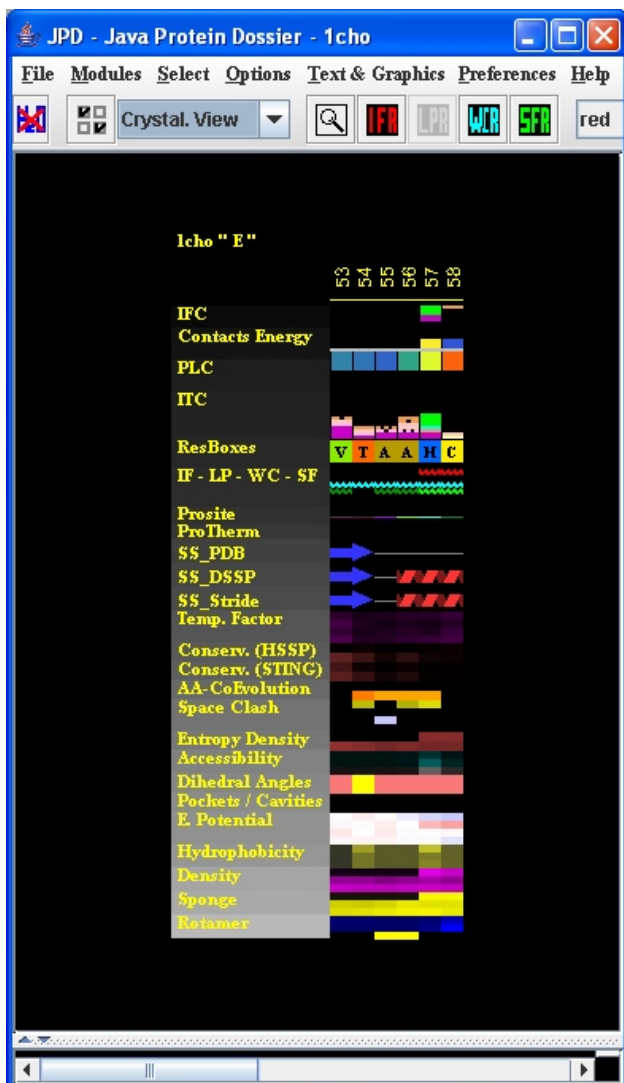


Fig.1. Parte da sequência de aminoácidos da proteína 1cho representados no JPD.

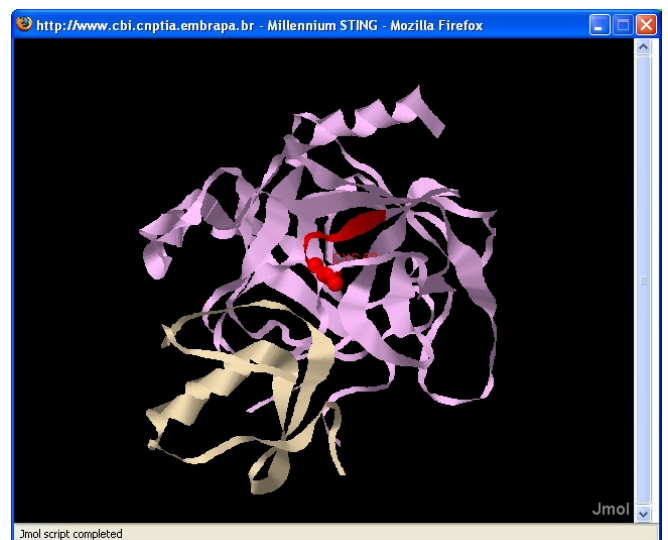


Fig. 2. Representação das seqüências de aminoácidos correspondentes ao PROSITE em 3D.

O programa pode ser acionado por linha de comando, passando como parâmetros o banco PROSITE, o caminho dos arquivos PDB e o local de armazenamento dos arquivos de saída; exemplo: "\$ java vpro /db/prosite /db/pdb/db/pro".

Pode ser também usado como API (Application Programming Interface) por outros programas Java.

Conclusões

A busca de padrões de seqüências do PROSITE é simples e permite concluir informações importantes sobre a estrutura das proteínas. Mas para isso a escolha e o uso de ferramentas computacionais é fundamental.

O VPRO é uma alternativa para tornar mais eficiente a análise de seqüências no contexto de gerenciamento de cálculos do Sting usando Java.

Outras vantagens são a facilidade de uso, a opção de interagir com os arquivos PDB sem a necessidade de programas auxiliares e poder ser instanciado por outros programas Java.

Referências Bibliográficas

DE CASTRO, E.; SIGRIST, C. J. A.; GATTIKER, A.; BULLIARD, V.; PETRA, S.; LANGENDIJK-GENEVAUX, P. S.; GASTEIGER, E.; BAIROCH, A.; HULO, N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, v. 34, (Web Server issue), p. W362-W365, July 2006.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. *Embrapa Informática Agropecuária [home page]*. Disponível em: <<http://www.cnptia.embrapa.br>>. Acesso em: 06 nov. 2006.

FASTA format description. Disponível em: <<http://compbio.mcs.anl.gov/blast/docs/fast.html>>. Acesso em: 06 nov. 2006.

HIGA, R. H.; TOGAWA, R. C.; MANCINI, A.; FALCÃO, P. R. K.; YAMAGISHI, M. E. B.; PAPPAR JUNIOR, G.; TORRES, V. W.; CAMOS, T. F. E.; BAUDET, C.; NESHICH, G. STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, v. 31, n. 13, p. 3386-3392, 2003.

HULO, N.; BAIROCH, A.; BULLIARD, V.; CERUTTI, L.; DE CASTRO, E.; LANGENDIJK-GENEVAUX, P. S.; PAGNI, M.; SIGRIST, C. J. A. The PROSITE database. *Nucleic Acids Res.*, v. 34, (Database issue), p. D227-D230, Jan. 2006.

NESHICH, G.; MANCINI, A.; BAUDET, C.; YAMAGISHI, M. E. B.; FALCÃO, P. R. K.; FILETO, R.; ROCCHIA, W.; PINTO, I.; MONTANER, A.; PALANDRANI, J.; KRAUCHENKO, J.; TORRES, R. C.; SOZA, S.; TOGAWA, R. C.; HIA, R. H. Java Protein Dossier: a novel web based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Res.*, v. 32, p. W595-W601, 2004.

THE PERL FOUNDATION. *The Perl Directory - perl.org*. Disponível em: <<http://www.perl.org/>>. Acesso em: 06 nov. 2006.

SUN MICROSYSTEMS. *Java SE - overview - at a glance*. Disponível em: <<http://java.sun.com/javase/>>. Acesso em: 06 maio 2007.

SWISS INSTITUTE OF BIOINFORMATICS. *UniProtKB/Swiss-Prot*. Disponível em: <<http://www.ebi.ac.uk/swissprot/>>. Acesso em: 06 nov. 2006.

Comunicado Técnico, 81

Ministério da Agricultura, Pecuária e Abastecimento



Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone: (19) 3789-5743
Fax: (19) 3289-9594
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2007

© Todos os direitos reservados.

Comitê de Publicações

Presidente: Kleber Xavier Sampaio de Souza.
Membros Efetivos: Adriana Farah Gonzalez (secretária), Ivanilde Dispatto, José Iguelmar Miranda, Marcia Izabel Fugisawa Souza, Sílvio Roberto Medeiros Evangelista, Stanley Robson de Medeiros Oliveira.

Suplentes: Laurimar Gonçalves Vandrúsculo, Maria Goretti Gurgel Praxedes.

Expediente

Supervisor editorial: Ivanilde Dispatto
Normalização bibliográfica: Marcia Izabel Fugisawa Souza
Editoração eletrônica: Área de Comunicação e Negócios