Documentos

ISSN 1677-9274
Dezembro, 2011
115

Eutils-search versão 2.0 – Manual do Usuário

Embrapa

ISSN 1677-9274
Dezembro, 2011

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 115

Eutils-search versão 2.0 - Manual do usuário

Roberto Hiroshi Higa
Rafael Seiji Tanaka

Embrapa Informática Agropecuária
Campinas, SP
2011

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição on-line 2011

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Tanaka, Rafael Seiji.

Eutils-search versão 2.0 – manual do usuário / Rafael Seiji Tanaka, Roberto Hiroshi Higa. - Campinas : Embrapa Informática Agropecuária, 2011.

23p. : il. – (Documentos / Embrapa Informática Agropecuária ; ISSN 1677-9274, 115).

1. Software eutils-search. 2. Biotecnologia. 3. Biologia molecular. 4. Mineração de texto. I. Higa, Roberto Hiroshi. II. Embrapa Informática Agropecuária. III. Título. IV. Série.

005.15 CDD (21. ed.)

Autores

Rafael Seiji Tanaka
Estagiário da Embrapa Informática Agropecuária
e-mail: tanaka.rafael@gmail.com

Roberto Hiroshi Higa
Doutor em Engenharia Elétrica
Pesquisador da Embrapa Informática Agropecuária
Av. André Tosello, 209, Barão Geraldo
Caixa Postal 6041 - 13083-970 - Campinas, SP
Telefone: (19) 3211-5862
e-mail: roberto@cnptia.embrapa.br

Apresentação

O software Eutils-search tem por objetivo trazer do banco de dados PubMed informações sobre artigos relacionados a genes de um organismo específico, de acordo com as regras referentes à taxa de acesso impostas pelo site. As informações trazidas são, então, armazenadas localmente em um banco de dados para acesso rápido. Além disso, o software também gera documentos XML correspondentes às informações do organismo requisitado.

O eutils-search é uma ferramenta de apoio ao desenvolvimento de aplicações de mineração de textos voltadas para os domínios de biotecnologia e biologia molecular, baseada em informações textuais obtidas do banco de dados PubMed.

Este documento apresenta os pré-requisitos e a descrição dos parâmetros necessários para utilização do software, bem como uma descrição de alguns aspectos internos do software, para melhor entendimento do processo que ele automatiza, além de alguns exemplos e uso.

Kleber Xavier Sampaio de Souza
Chefe-geral
Embrapa Informática Agropecuária

Sumário

Eutils-search versão 2.0 – Manual do Usuário

Rafael Seiji Tanaka
Roberto Hiroshi Higa

Pré-requisitos

Banco de dados PostgreSQL e a biblioteca JDBC correspondente.

Espaço mínimo disponível em disco: 3GB ou de acordo com o organismo requisitado.

Novas funcionalidades - versão 2.0

Na versão 2.0 foram adicionados 5 novos parâmetros:

- `updatelist`: faz a atualização dos 4 arquivos (arquivos gene2pub med, generifs_basic, gene_info e nodes.dmp).
- `force`: usado para realizar o download artigos de varias espécies de uma só vez.
- `search`: faz uma busca no site do NCI e baixa os artigos retornados como resposta.
- `createqueryxml`: Cria arquivos XML para os artigos vinculados à

query ID de busca especificada.

- `listqueries`: lista as queries e correspondentes IDs armazenados no banco de dados.

Além disso, o parâmetro `-fetchdata` agora trata qualquer taxid e não apenas aqueles em nível de organismos (*species*).

Parâmetros de chamada

O eutils-search é executado por meio da seguinte linha de comando:

```
java -jar eutils-search.jar [parâmetros] tax_id
```

onde tax_id é o ID do tax requisitado e pode ser encontrado no site do NCBI por meio do link http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root. O eutils-search considera tax_ids de forma genérica, embora seja mais comum a utilização de tax_ids que se referem a organismos específicos, *species* na terminologia utilizada pelo NCBI.

Para usar o eutils-search é necessário que os arquivos gene2pubmed, gene_info e generifs_basic e nodes.dmp estejam no mesmo diretório do arquivo eutils-search.jar. Esses arquivos podem ser obtidos, via ftp, dos sítios <ftp://ftp.ncbi.nih.gov/gene/> e <ftp://ftp.ncbi.nih.gov/pub/taxonomy/>. O parâmetro `--updatelist` faz com que esses arquivos sejam atualizados.

O eutils-search possui os seguintes parâmetros:

`--createdb`

Essa opção carrega os dados do arquivo gene2pubmed, que contém as relações entre taxonomias, genes e artigos publicados, para o banco de dados PostgreSQL, eliminando qualquer tabela anteriormente criada. Essa operação deve ser executada uma única vez com o intuito de criar as tabelas ou se o usuário realmente quiser remover as tabelas existentes.

`--populate`

Essa opção indica que todos os dados do arquivo gene2pubmed, que contém as relações entre taxonomias, genes e artigos publicados, devem ser carregados para o banco de dados PostgreSQL. Essa operação é computacionalmente muito custosa e deve ser executada uma única vez. Para atualizar o banco de dados, deve-se utilizar o parâmetro `--update-pubmed`.

`--updatepubmed`

Essa opção indica que o banco de dados deve ser atualizado com novas informações do arquivo gene2pubmed.

`--updaterif`

Essa opção indica que o banco de dados deve ser atualizado com informações do arquivo generifs_basic, que contém a descrição RIF dos artigos do PubMed.

`--updategeneinfo`

Essa opção indica que o banco de dados deve ser atualizado com informações do arquivo gene_info, que contém informações sobre os genes dos organismos.

`--updatelist`

Faz com que os arquivos gene2pubmed, gene_info, generifs_basic e nodes.dmp sejam atualizados. Ao ser executado, a primeira ação realizada pelo eutils-search é verificar as datas da download dos arquivos gene2pubmed, generifs_basic, gene_info e nodes.dmp. Se a diferença exceder a 7 dias (período padrão), todos os arquivos são atualizados, ou seja, baixados novamente.

`--search "expressão de busca"`

Essa opção realiza uma busca no site do NCBI com a expressão especificada e faz o download de todos os artigos que a satisfazem.

`--creategenexml`

Essa opção indica que os arquivos XML referentes ao taxon requisitado devem ser criados. Cada arquivo XML criado agrupa as informações referentes a um gene por taxon requisitado (um exemplo de arquivo XML criado é apersentado no anexo I).

`--createarticlexml`

Essa opção indica que os arquivos XML referente ao taxon requisitado devem ser criados. Neste caso, as informações não são agrupadas por gene, cada novo documento criado corresponde a um artigo.

`--fetchdata`

Essa opção indica que o download dos títulos, datas de publicação e resumos dos artigos relacionados ao organismo requisitado devem ser executados. Devido às restrições impostas para acesso do PubMed, essa operação faz o download de informações com 100 artigos a cada requisição. O intervalo entre cada requisição é de 1 hora de Segunda a Sexta-feira. Não existe intervalo entre requisições das 23h às 05h ou durante Sábados e Domingos. Essa operação pode demorar muito tempo e deve, preferencialmente, ser feita durante os finais de semana. Além disso, ela não deve ser interrompida. Sua interrupção acarretará no seu reinício, ou seja, todo o download feito até o momento será feito novamente.

`--force`

Quando taxid especificado refere-se a uma espécie, o parâmetro `–fetchdata` realiza o download dos artigo relacionados. Se o taxid refere-se a um taxon superior, a especificação deste parâmetro resulta no download de todos os artigos relacionados a espécies abaixo do taxon (taxid) especificado.

Estrutura interna

A utilização do eutils-search é seguida de um processo de mineração de textos, em que os dados obtidos do PubMed são tratados. Dependendo de como este processo está estruturado, os dados obtidos por meio do eutils-search podem ser acessados de diferentes formas. Por isso, é interessante que o usuário conheça alguns detalhes do processo que o eutils-search automatiza e de sua estrutura interna.

O processo automatizado pelo eutils-search pode ser dividido em quatro etapas:

- Criar as tabelas no banco de dados local;
- Ler os arquivos nodes.dmp, gene2pubmed, gene_info e generifs_basic, nessa ordem;
- Obter os dados do PubMed (dado o tax_id referente a um organismo específico);
- Gerar os documentos XML;

Alguns passos não precisam ser executados, caso já o tenham sido anteriormente. Além disso, o passo 2 pode consistir apenas na atualização das informações já existentes. Caso o usuário queira atualizar as informações existentes, basta utilizar um dos parâmetros `update`, apresentados na sessão anterior.

O eutils-search foi desenvolvido utilizando a linguagem java e, internamente, está organizado em 7 classes: Article, Gene, myXMLParser, Searcher, SQLConn, Tax e configHandler.

Classes Article, Gene e Tax

As classes Article, Gene e Tax refletem a estrutura dos dados tratados. Cada uma dessas classes é responsável por armazenar um artigo, um gene e uma taxonomia, respectivamente, criando uma hierarquia na qual os Organismos estão na raíz da árvore, os Genes são filhos dos Organismos e os Artigos (*Article*) são filhos dos Genes. Resumidamente, um conjunto de artigos está elacionado a um gene que, por sua vez, está

relacionado a um organismo.

Essas três classes representam cada tipo específico de dado. Por exemplo, a classe Article armazena o PMID (ID do artigo no PubMed), o título do artigo, seu resumo e outras informações. A classe Gene armazena dados como a descrição do gene e seu símbolo e uma coletânea de artigos referentes a esse gene. A classe Tax armazena apenas o tax_id (ID do organismo) e uma lista de genes que o compõe.

Além disso, a classe Tax é responsável por carregar os dados de 3 arquivos: generifs_basic, gene2pubmed e gene_info. Esses arquivos serão explicados a seguir, sendo que mais informações podem ser obtidos em <ftp://ftp.ncbi.nih.gov/gene/README>.

Arquivos genrifs_basic, gene2pubmed e gene_info

O arquivo generifs_basic pode ser obtido em ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz e deve ser descompactado no mesmo diretório onde o arquivo executável eutils-search.jar está localizado. Ele possui anotações sobre alguns dos genes do PubMed, sendo estruturado em 5 colunas: TaxID, GeneID, PubmedID, Last Update, GeneRIF Text, sendo que as colunas PubmedID e Last Update não são lidas pelo eutils-search. Cada linha desse arquivo contém a informação RIF para uma tripla TaxID, GeneID, PubmedID. A informação GeneRIF Text é armazenada no banco de dados local e é, posteriormente, utilizada para gerar os documentos XML. Note que, apesar da informação PubmedID estar presente nesse arquivo, ela não indica que o GeneRIF é uma informação específica de um artigo, dado que o PubmedID presente nesse arquivo é o primeiro artigo encontrado que está relacionado com o GeneID em questão. Vale ainda notar que o RIF é uma informação entrada manualmente no banco de dados do PubMed através de um processo que pode ser realizado através do link http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html. No banco de dados utilizado pelo eutils-search, a informação sobre o RIF é armazenada em uma tabela específica, pois a quantidade de genes que possuem anotação é muito pequena, comparada com a quantidade de genes no PubMed.

O arquivo gene2pubmed pode ser obtido em ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz e deve ser descompactado no mesmo diretório onde o executável eutils-search.jar está localizado. Ele contém a relação entre artigos, genes e organismos, sendo dividido em três colunas: TaxID, GeneID, PubmedID. Toda a hierarquia entre artigo, gene e taxonomia é definida nesse arquivo que, portanto, tem um papel central no processo automatizado pelo eutils-search.

O arquivo gene_info pode ser obtido em ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz e deve ser descompactado no mesmo diretório em que o executável eutils-search.jar está localizado. Ele contém as informações sobre genes, como descrição e símbolo. Nesse arquivo, que possui diversas colunas, apenas 5 colunas são relevantes para o eutils-search: TaxID, GeneID, Symbol, Description e Other_Designations.

As classes SQLConn, myXMLParser, configHandler Searcher

A classe SQLConn é responsável por realizar a comunicação entre o eutils-search e o servidor de banco de dados local (PostgreSQL). Ela é a responsável pela realização de consultas, leitura dos arquivos generifs_basic, gene2pubmed e gene_info e inserção de dados no banco de dados, armazenamento dos dados obtidos do PubMed, bem como pela criação das tabelas do banco de dados. Os parâmetros de conexão com o servidor de banco de dados são obtidos da classe configHandler. O diagrama de tabelas do banco de dados utilizado por eutils-search é apresentado na Figura 1.

A classe myXMLParser é responsável por tratar os arquivos obtidos do PubMed no formato XML e criar os correspondentes arquivos XML. Para a leitura de arquivos XML, essa classe faz uso do JavaX e do módulo Xpath e para criação dos arquivos XML, ela faz uso da classe SQLConn para obtenção de dados do banco de dados.

A classe configHandler tem como único propósito ler o arquivo de configurações eutils.conf que armazena informações de conexão com o banco de dados. O arquivo de configuração armazena os parâmetros database, username e password, utilizados para conexão com o banco de dados

local. Esse arquivo deve estar no mesmo diretório do programa executável eutils-search.jar. No arquivo eutils.conf, esses parâmetros são definidos utilizando a sintaxe <parâmetro>=<valor>. Por exemplo:

- database=mydb
- username=myuser
- password=mypass

A classe Searcher contém o procedimento principal do software, onde os parâmetros de entrada são lidos e as funções necessárias chamadas, de acordo com o que foi requisitado pelo usuário. Além disso, nessa classe é feito o controle de requisições ao PubMed, de acordo com os limites propostos pelo sistema. Entre o período de 21h e 5h durante a semana e aos sábados e domingos não há limite de requisições. Com exceção desses períodos, existe um limite de 3 requisições por segundo e 100 requisições por dia.

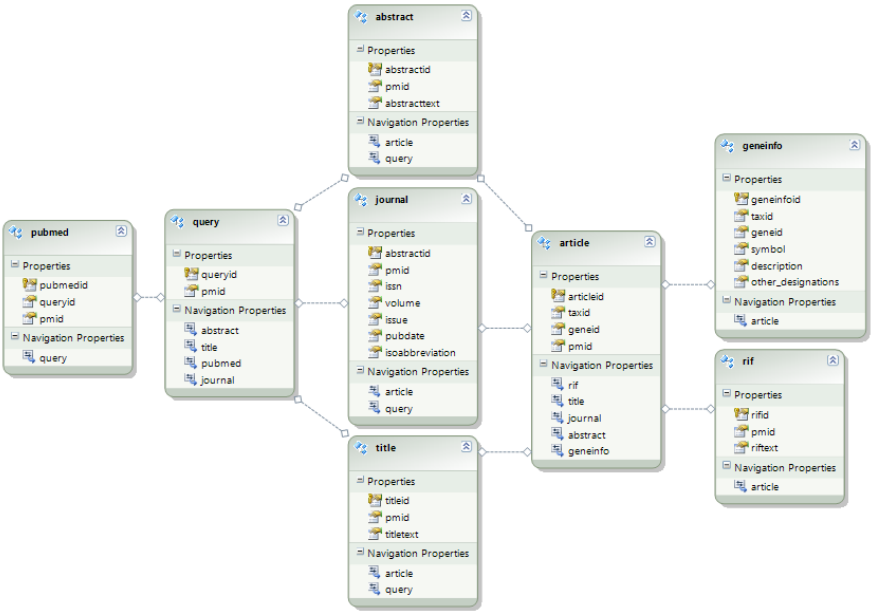

Figura 1. Diagrama de tabelas do banco de dados.

Exemplos

```
--createdb --populate --updaterif --updategeneinfo
--fetchdata --createarticlexml 9913
```

Para realizar todo o processo, ou seja, criar o banco de dados, ler os arquivos generifs_basic, gene2pubmed e gene_info, fazer o download dos dados do PubMed e obter os documentos do organismo Bos taurus (tax_id = 9913).

```
--populate --updaterif –updategeneinfo
```

Supondo que o banco de dados já foi criado, para populá-lo com os dados contidos nos arquivos arquivos generifs_basic, gene2pubmed e gene_info.

```
--updatelist
```

Se esse parâmetro não for passado e o programa estiver desatualizado, a mensagem será impressa:

```
aO(s) seguinte(s) arquivo(s) está(ão)
desatualizado(s):
gene2pubmed
gene_info
generifs_basic
nodes.dmp
Para atualiza-los, rode o programa com o parametro
'--updatelist'.
```

Se esse parâmetro for passado e o programa estiver desatualizado, será feita a atualização dos arquivos necessários:

```
Baixando gene2pubmed.gz (22MB)...
Descompactando... Pronto.

Baixando gene_info.gz (125MB)...
Descompactando... Pronto.
```

```
Baixando generifs_basic.gz (29MB)...
Descompactando... Pronto.

Baixando taxdump.tar.gz (18MB)...
Descompactando... Pronto.
```

```
--search "science[journal] AND breast cancer AND
2008[pdat]"
```

Exemplo com uma expressão de busca:

```
length = 6
Fetching from 0 to 6 out of 6
Fetched 88319 bytes.
```

```
--fetchdata
```

Exemplo com um taxid(70448) de uma specie:

```
70448 species
Searching database for tax_id = 70448
Fetching data for tax_id = 70448
length = 2
Fetching from 0 to 2 out of 2
Fetched 30167 bytes.
```

Exemplo com um taxid acima de specie, sem o parâmetro –force:

```
__627479 no rank
____484896 species
____551790 species
____627480 species
____693582 species
____718008 species
____881915 species
____943335 species
```

```
Para baixar os artigos de todas as species, rode o
programa com o parametro '--fetchdata –force'.
```

```
--force --fetchdata
```

Se executado com um taxid acima de specie, com o parâmetro –force:

```
__627479 no rank
____484896 species
____551790 species
____627480 species
____693582 species
____718008 species
____881915 species
____943335 species
Searching database for tax_id = 484896
Fetching data for tax_id = 484896
length = 1
Fetching from 0 to 1 out of 1
Fetched 6848 bytes.
Searching database for tax_id = 551790
Fetching data for tax_id = 551790
length = 1
Fetching from 0 to 1 out of 1
Fetched 13628 bytes.
Searching database for tax_id = 627480
Fetching data for tax_id = 627480
length = 1
Fetching from 0 to 1 out of 1
Fetched 6803 bytes.
Searching database for tax_id = 693582
Fetching data for tax_id = 693582
length = 1
Fetching from 0 to 1 out of 1
```

```
Fetched 10048 bytes.
Searching database for tax_id = 718008
Fetching data for tax_id = 718008
length = 0
Searching database for tax_id = 881915
Fetching data for tax_id = 881915
length = 0
Searching database for tax_id = 943335
Fetching data for tax_id = 943335
length = 0
```

```
--createarticlexml 9913
```

Supondo que os dados já foram obtidos do pubmed, para criar os documentos XML.

Anexo – Exemplo de arquivo XML criado pelo eutils-search

```
<geneDoc>
      <taxId>9913</taxId>
      <geneId>280691</geneId>
      <geneSymbol>H4</geneSymbol>
      <geneDescription>histone H4</geneDescription>
      <geneOtherDesignations>H4.1</geneOtherDesigna-
tions>
      <articleSet>
            <article pmid=”4321977”>
                  <title>Structure of the glycine-rich,
arginine-rich histone of the Novikoff hepatoma.</title>
                  <abstract/>
                  <rif/>
```

```
            </article>
            <article pmid="5780842">
                  <title>Structural analysis of the
glycine-rich, arginine-rich histone. II. Sequence of
the half of the molecule containing the aromatic amino
acids.</title>
                  <abstract/>
                  <rif/>
            </article>
            <article pmid="5817359">
                  <title>Structural analysis of the
glycine-rich, arginine-rich histone. 3. Sequence of the
amino-terminal half of the molecule containing the mo-
dified lysine residues and the total sequence.</title>
                  <abstract/>
                  <rif/>
            </article>
            <article pmid="9524213">
                  <title>Poly A-containing histone H4
mRNA variant (H4-v. 1): isolation and sequence determi-
nation from bovine adrenal medulla.</title>
                  <abstract>A histone H4 cDNA variant
(H4-v.1) was cloned from a bovine adrenal medullary
phage library using PCR as a method of detection. The
isolated clones contained a short 5' untranslated re-
gion (UTR) followed by the histone H4 coding region and
a long atypical 3'UTR. The 3'UTR comprised the palin-
dromic and purine-rich sequences typical of cell-cycle
dependent histone mRNAs, and a 1.1 kb extension downs-
tream of the palindromic sequence ending with a poly(A)
track typical of cell-cycle independent histone mRNAs.
Northern blot and RT-PCR analyses indicate that the
transcript is fully expressed in bovine adrenal me-
dulla. Thus, bovine histone H4-v.1 mRNA represents the
first example of a histone H4 transcript that contains
both 3'UTR characteristics of cell-cycle dependent and
cell-cycle independent histone mRNAs.</abstract>
                  <rif/>
            </article>
```

```
            <article pmid="17825834">
                  <title>A nucleosome interaction module is required for normal function of Arabidopsis thaliana BRAHMA.</title>
                  <abstract>The BRAHMA (BRM) gene encodes the SNF2-type ATPase of the putative Arabidopsis thaliana SWI/SNF chromatin remodelling complex. This family of ATPases is characterized by the presence of a conserved catalytic domain and an arrangement of auxiliary domains, whose functions in the remodelling activity remains unclear. Here, we characterize, at the molecular and functional level, the carboxy-terminal part of Arabidopsis BRM. We have found three DNA-binding regions that bind various free DNA and nucleosomal probes with different specificity. One of these regions contains an AT-hook motif. The carboxy terminus also contains a bromodomain able to bind histones H3 and H4. We propose that this array of domains constitute a nucleosome interaction module that helps BRM to interact with its substrate. We also characterize an Arabidopsis mutant that expresses a BRM protein lacking the last 454 amino acid residues (BRM-DeltaC), encompassing the bromodomain and two of the three DNA-binding activities identified. This mutant displays an intermediate phenotype between those of the wild-type and a null allele mutant, suggesting that the nucleosome interaction module is required for the normal function of BRM but it is not essential for the remodelling activity of BRM-containing SWI/SNF complexes.</abstract>
                  <rif>characterization at the molecular and functional level, the carboxy-terminal part of Arabidopsis BRM. We have found three DNA-binding regions that bind various free DNA and nucleosomal probes with different specificity</rif>
            </article>
      </articleSet>
</geneDoc>
```

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 9752