

**Servizi bioinformatici del sito web
dell'ormone Responsive Breast Cancer
Genomics Network**
(<http://www.hrbc-genomis.net/>)

Paolo Romano
Istituto Nazionale per la Ricerca sul Cancro
(paolo.romano@istge.it)

Sommario

- ❑ Obiettivi del sito HRBC
- ❑ L'integrazione di dati in biologia
- ❑ SRS: strumento per l'integrazione
- ❑ Esempio: CABRI

Motivazioni

- ❑ strumenti bioinformatici distribuiti su siti diversi:
 - 👉 difficoltà nella ricerca e nella scelta degli strumenti,
- ❑ interfacce, metodi di ricerca, strutture dati eterogenei:
 - 👉 difficoltà nell'utilizzo degli strumenti disponibili
- ❑ post-genomica in continua evoluzione:
 - 👉 strumenti bioinformatici poco numerosi e interfacce primitive,
- ❑ elevata partecipazione nel progetto:
 - 👉 difficoltà di coordinamento e messa in comune dei dati
 - 👉 sinergie e ottimizzazione risorse

Obiettivi

Realizzare un portale che:

- ❑ dia visibilità al progetto e ai partner (parte accessibile a tutti)
- ❑ serva come strumento di lavoro e coordinamento per le unità di ricerca (parte ad accesso riservato)
- ❑ ospiti servizi e strumenti di tipo bioinformatico, utili ai ricercatori coinvolti nel progetto e non,
- ❑ rimanga un riferimento alla fine del progetto

Risultati attesi

- ❑ Disponibilità di strumenti di tipo generale (banche dati, sistemi di ricerca, programmi per analisi delle sequenze, strumenti bioinformatici di uso comune)
- ❑ SRS per interrogazione di banche dati di pubblico dominio (GenBank, LocusLink, OMIM, SwissProt, ecc.) e strumenti di analisi di pubblico dominio (BLAST, FASTA, ecc.)
- ❑ Disponibilità di strumenti di analisi e banche dati di specifico interesse ai fini del progetto HRBC

Contenuti (area pubblica)

Presentazione del progetto

- ❑ sintesi del progetto di ricerca
- ❑ responsabili unità operative e altri contatti
- ❑ elenco articoli/documenti vari prodotti nell'ambito del progetto

Link esterni (portale)

- ❑ link a siti unità operative e partner
- ❑ link a siti di interesse scientifico affine al progetto
- ❑ link a corsi di formazione on-line (free) selezionati tra quelli esistenti per la loro attinenza al progetto e agli strumenti del progetto

Contenuti (area pubblica)

- ❑ Sito SRS (con accesso a software d'analisi)
- ❑ Materiale didattico (corsi organizzati dal progetto e dai partner)
 - Sull'accesso e utilizzo degli strumenti
 - Sulle tecnologie (microarray)
- ❑ Mirror di corsi creati da altri ricercatori
 - BioComputing Division VSNS

Integrazione delle banche dati

- L'integrazione delle banche dati è necessaria per
 - Ottenere una visione complessiva delle informazioni disponibili
 - Eseguire in un numero limitato di passaggi interrogazioni e/o analisi che coinvolgono più banche dati e software
 - Effettuare un reale data mining

Integrazione delle banche dati

- L'integrazione delle banche dati comporta
 - L'analisi e la definizione accurata e univoca degli "oggetti biologici" coinvolti
 - L'analisi dei dati disponibili
 - L'identificazione dei collegamenti tra informazioni presenti in banche dati diverse
 - La definizione e l'implementazione di formati comuni per l'interscambio delle informazioni

I metodi dell'integrazione

□ Sintattici

- Riferimenti reciproci (xrefs)
- Descrizioni condivise (vocabolari)

□ Semantici

- Modelli a oggetti
- Schemi relazionali
- Ontologie

Riferimenti reciproci

- Da un record, a un record collegato di un'altra banca dati:
 - Link diretto, univoco, non reciproco
 - ID database remoto
 - Formati standardizzati
 - Life Science ID
 - Standard I3C

- Limitazioni:
 - Annotazione manuale
 - Significato del collegamento
 - Predefiniti

Descrizioni condivise

- Da un record ai record di un'altra banca dati tramite ricerca testuale:
 - Link implicito, reciproco
 - Determinabile automaticamente
 - Termine di vocabolario
 - Vocabolari standardizzati

- Limitazioni:
 - Diffusione di vocabolari condivisi nell'annotazione
 - Significato del collegamento
 - Necessità di definire l'ambito

SRS - Sequence Retrieval Software

- SRS è un motore di ricerca che consente di interrogare in maniera integrata banche dati eterogenee memorizzate localmente, in maniera semplice ed efficiente
- L'approccio originale di SRS consiste in
 - Banche dati disponibili localmente come “flat file”
 - Sintassi specifiche per l'identificazione dei dati
 - Link interni espliciti e impliciti tra banche dati
 - Integrazione trasparente con applicazioni
 - Integrazione esterna tramite link HTML

Flat file

- I “flat file” sono file di solo testo
 - Non possono includere nessun carattere di controllo (formattazione)
 - Non possono includere immagini, altri elementi multimediali, altri contenuti binari
 - Spesso, i caratteri sono limitati al set ASCII base (0 – 127)

Flat file: vantaggi

- I vantaggi derivanti dall'utilizzo di flat file sono:
 - Formato molto diffuso
 - È “leggibile” e adatto anche agli operatori
 - Non necessita di software costosi
 - Possono includere dati complessi, in maniera articolata, utilizzando un'apposita sintassi
 - Sono facili da indicizzare
 - Molte informazioni già disponibili non saranno mai strutturate diversamente (80%, in calo)

Flat file: svantaggi

- Gli svantaggi dell'utilizzo di flat file sono:
 - Difficoltà di gestione e aggiornamento delle banche dati
 - Mancanza di controllo di qualità dei dati
 - Mancanza di un linguaggio di interrogazione
 - Scarsa o assente modellizzazione degli oggetti biologici descritti
 - Scarsa o assente strutturazione dei dati

Flat file e DBMS

□ DBMS per gestire i dati

- Database relazionali o a oggetti consentono di gestire in maniera soddisfacente le banche dati
- Lo sviluppo dei modelli, il controllo di qualità e la gestione dei dati tramite DBMS

□ Flat file per scambiare i dati

- Semplicità nel creare flat file come export, anche con struttura complessa e articolata
- Conservazione della qualità dei dati
- Semplicità di trasferimento

SRS – Dai flat file alle entries

□ Flat file per SRS

- Banche dati in formato flat file/XML
- Ogni db possiede una sua specifica sintassi, corrispondente alla struttura dati o DTD
- Analizzando sintatticamente i flat file, SRS è in grado di identificare tutte le informazioni che si riferiscono a un singolo elemento o record
- Queste costituiscono una entry

Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964;NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium rubiacearum, (ex Knösel 1962) Knösel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D.Knösel
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. See also Agrobacterium sp. LMG 1 (t2)

Strain_number LMG 1(t2)
Other_collection_numbers -
Restrictions Either Biohazard group 1 or Biohazard group 2
Organism_type Bacteria
Name Agrobacterium sp.
Infrasubspecific_names -
Status -
Other_names Phyllobacterium rubiacearum, (ex Knösel 1962) Knösel 1984 VL
History <- D.Knösel (Phyllobacterium rubiacearum)
Conditions_for_growth Medium 16, 28C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks One (t2) out of two stable colony types isolated from the original culture LMG 1.

SRS – Dalle entry ai field

- L'analisi sintattica dei flat file permette a SRS di identificare i field all'interno di una entry
- Un Field (campo) è quella parte dell'entry che si riferisce a una particolare informazione
- I Field possono a loro volta includere subfield, a seconda della complessità della struttura dati e della relativa sintassi
- Elementi DTD possono essere tradotti direttamente in field

Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964; NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium rubiacearum, (ex Knösel 1962)
Knösel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D. Knösel
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. See
also Agrobacterium sp. LMG 1(t2)

SRS – Gli indici

- Qualunque parte della entry può essere indicizzata
 - Un indice speciale viene creato come mezzo d'accesso principale a ciascuna entry
 - Gli indici sono spesso creati sui contenuti dei singoli field, così che la ricerca possa essere fatta selezionandoli in maniera precisa
 - Le chiavi degli indici possono comprendere una o più parole, quando queste hanno un significato nel loro insieme (keywords)

SRS – I link

- I collegamenti (link) tra banche dati possono essere definiti in maniera
 - Esplicita, quando un termine è appositamente inserito in un field come riferimento a una entry di un'altra banca dati
 - Implicita, cercando termini comuni all'interno di field predefiniti di banche dati diverse

SRS – I link espliciti

- Esplicito riferimento a un'altra banca dati

Other_collection_numbers CCUG 34964; NCIB 12128

Literature DSM ref.no. 72; DSM ref.no. 1300

EMBL: X52289

SRS – I link impliciti

- Termini comuni in banche dati diverse

TargetGene: APOE

Constructed_from pMB1, pSC101 and Tn3

Name *Gluconacetobacter xylinus* subsp. *xylinus*, (Brown 1886) Yamada, Hoshino and Ishikawa 1998 VL

Literature *Nucleic Acids Res* 1990;18:4967 [PMID: 2395673]

SRS: operatori link

- SRS consente di utilizzare i link esistenti per le ricerche tramite un apposito operatore: <
 - o swissprot < EMBL
 - o EMBL < swissprot
 - o swissprot < [EMBL-id: X52289]
 - o [EMBL-organism:human]
< [medline-pmid:3137981]

CABRI: obiettivi

Common Access to Biological Resources and Information

- ❑ Distribuzione di materiali biologici di qualità
- ❑ Linee Guida per la conservazione del materiale
- ❑ Centro Risorse Biologiche virtuale
- ❑ Cataloghi integrati tramite SRS
- ❑ Integrazione con db esterni
- ❑ Shopping cart

CABRI: partner e materiali

Partner:

- ❑ BCCM, CABI, CBS, CIP, DSMZ, ICLC, NCCB, NCIMB (collezioni)
- ❑ IST, CERDIC (ITC)

Materiali:

- ❑ Microrganismi (Batteri, lieviti, funghi filiformi)
- ❑ Linee cellulari animali e umane, ibridomi, linee B tip. HLA
- ❑ Plasmidi, fagi, virus, sonde DNA
- ❑ Complessivamente più di 100.000 risorse

CABRI: struttura dati

Per ogni materiale, identificati:

- ❑ Minimum data Set (MDS): dati essenziali, necessari per identificare la risorsa
- ❑ Recommended Data Set (RDS): dati utili per una descrizione precisa della risorsa
- ❑ Full Data Set (FDS): tutti i dati disponibili sulla risorsa

Per ogni informazione, linee guida per l'inserimento dei dati:

- ❑ Descrizione testuale dettagliata
- ❑ Liste di termini e vocabolari di riferimento
- ❑ Sintassi predefinite

CABRI: Data sets

Data set	Field label	Catalogues
MDS	Strain_number	All
MDS	Other_collection_numbers	All
MDS	Name	All
RDS	Race	All
MDS	Organism_type	All
MDS	Restrictons	All
MDS	Status	All
MDS	History	All
RDS	Misapplied_names	All
RDS	Substrate	All
RDS	Geographic_origin	All
RDS	Sexual_state	All
RDS	Mutant	All
FDS	Genotype	DSMZ
.....

CABRI: Name field

Field	Name
Description	<p>Full scientific and most recent name of the strain. It includes:</p> <ul style="list-style-type: none">★ Genus name and species epithet★ Subspecies★ Pathovar★ Authors of the name★ Year of valid publication or validation★ Approbation of the name
Input process	<p>Enter full scientific name as given by depositor and confirmed (or changed) by collection. Names of authors of the name, year of valid publication or validation and approbation are included after a comma.</p> <p>Values for approbation: AL = approved list, c.f.r. IJSB 1980 VL = validation list, in IJSB after 1980 VP = validly published, paper in IJSB after 1980</p> <p>Reference list: <u>DSMZ list of bacterial names</u></p>
Required for	MDS

CABRI: Reference paper field

Field	Reference paper
Description	Original paper [if available]
Input process	<p>New entries: JournalTitle Year; Volume(issue): beginning page#-ending page#</p> <p>The title is abbreviated following international standard rules (ISSN). Abbreviations are without dot. Authors and title of the article are not mentioned.</p> <p>The reference can be followed by the Pubmed ID enclosed within square brackets as follows:</p>
Required for	[PMID: 1234567], where '1234567' is the Pubmed ID of the paper MDS

CABRI: integrazione

Per ogni catalogo:

- Link HTML a db riferimento (media, hazard, etc...)

Per ogni materiale:

- Link SRS tra cataloghi, basati su dati espliciti e impliciti (Other_collection_numbers)

Per tutti i cataloghi:

- Link HTML basati Pubmed ID verso Medline
- Link SRS / HTML per EMBL Data Library