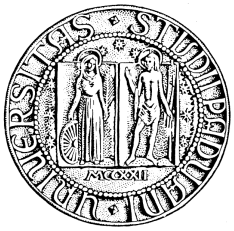
UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE



TESI DI LAUREA

Annotazione funzionale di proteine utilizzando la similarità semantica in Gene Ontology

Relatore:

Ch.mo. prof. MASSIMO MARESCA

Correlatore:

Dott. STEFANO TOPPO

Laureando:

ALESSANDRO PESCAROLO

ANNO ACCADEMICO 2008-2009

Contenuti

Introduzione

Il presente elaborato di tesi nasce dall'esigenza manifestata da un gruppo di ricercatori nel campo della bioinformatica, di rendere maggiormente agevole e quindi produttiva l'annotazione funzionale di proteine anonime utilizzando il concetto di similarità semantica basato sull'ontologia genica (GO).

L'attività lavorativa, in ambito informatico, che da cinque anni svolgo all'interno del Dipartimento di Chimica Biologica, mi ha permesso di entrare in contatto con molteplici tematiche di natura biologica.

Proprio dall'incontro con una di queste è nata l'esigenza di dar vita ad un applicativo web, che consenta al ricercatore di laboratorio la possibilità di utilizzare un strumento per analizzare centinaia di sequenze allo scopo di annotarle rapidamente anziché svolgere il tutto in maniera manuale attraverso la shell del sistema operativo.

Per annotare le sequenze anonime, viene utilizzato un innovativo metodo denominato ARGOT (Annotatio Retrevial of Gene Ontology Terms) [1], nato dalla collaborazione tra il centro di ricerca FEM-IASMA di San Michele all'Adige in provincia di Trento e il Dipartimento di Chimica Biologica dell'Università di Padova.

Tale metodo, che rappresenta il cuore dell'applicativo, è sviluppato utilizzando il linguaggio Java e coinvolge funzioni statistiche, di previsione, di ontologia genica e interrogazione di database esterni pubblici.

1. Bioinformatica

La bioinformatica[1] è una disciplina scientifica che con metodi informatici si occupa della risoluzione di problemi biologici a livello molecolare.

Costituisce un tentativo di descrivere dal punto di vista statistico i fenomeni biologici: storicamente ed epistemologicamente la biologia ha fatto minor ricorso ad un approccio matematico rispetto ad altre discipline scientifiche (come ad esempio fisica e chimica). Lo scopo della bioinformatica è quindi cercare di supplire a questa lacuna fornendo ai risultati tipici della biochimica e della biologia molecolare un corredo di strumenti analitici e numerici. Vengono coinvolte, oltre all'informatica, la matematica applicata, la statistica, la chimica e la biochimica, inoltre nozioni di intelligenza artificiale.

Principalmente la bioinformatica si occupa di:

- fornire modelli statistici validi per l'interpretazione dei dati provenienti da esperimenti di biologia molecolare e biochimica al fine di identificare tendenze e leggi numeriche;
- generare nuovi modelli e strumenti matematici per l'analisi di sequenze di DNA, RNA e proteine, al fine di creare un corpus di conoscenze relative alla frequenza di sequenze rilevanti la loro evoluzione ed eventuale funzione;
- organizzare le conoscenze acquisite a livello globale su genoma e proteoma in basi di dati, al fine di rendere tali dati accessibili a tutti e ottimizzare gli algoritmi di ricerca dei dati stessi per migliorarne l'accessibilità;

1 Definizione tratta da it.wikipedia.org, l'enciclopedia libera

L'evoluzione storica della bioinformatica, che inizialmente si occupava principalmente dello studio del DNA e RNA, ha portato ad un così vasto uso dell'informatica in molti settori della biologia che è stato coniato il nuovo termine, ormai universalmente accettato, di Biologia Computazionale che esplicita con maggior chiarezza e precisione i reali e più vasti contenuti scientifici e disciplinari del connubio tra informatica e biologia nel XXI secolo.

Gli attuali ambiti di ricerca includono l'allineamento di sequenze, la predizione genica, l'allineamento di sequenze proteiche, la predizione di struttura proteica, l'espressione genica e l'interazione proteina-proteina.

1.1 Banche dati

Una delle attività principali dei bioinformatici consiste nella progettazione, costruzione e uso di banche dati di interesse biologico. Una banca dati raccoglie dati e informazioni derivanti da esperimenti di laboratorio, da esperimenti in silico e dalla letteratura scientifica. Il termine in silico al dato informatico che viene utilizzato come punto di partenza per gli esperimenti in vitro. Si dice "in silico", per il semplice fatto che i processori dei computers sono costituiti principalmente dal silicio. Le banche dati sono progettate come contenitori costruiti per immagazzinare dati in modo efficiente e razionale al fine di renderli facilmente accessibili a tutti gli utenti: ricercatori, medici, studenti, etc.

Una banca dati è costituita da voci (in inglese records) ciascuna contenente informazioni sull'oggetto caratteristico della banca dati, per esempio: sequenze nucleotidiche o referenze bibliografiche, che insieme a tutte le altre informazioni si riferiscono a quel record in particolare.

Un record di una banca dati di sequenze nucleotidiche potrebbe contenere, oltre alla sequenza di una molecola di DNA, il nome dell'organismo cui la sequenza appartiene, la lista degli articoli scientifici che riportano dati su quella sequenza, le caratteristiche funzionali (ovvero se si tratta di un gene o di una sequenza non codificante) e ogni altra informazione ritenuta di interesse.

Le banche dati possono essere di due tipi: primarie o specializzate.

Le banche dati primarie contengono informazioni e annotazioni delle sequenze nucleotidiche e proteiche, strutture del DNA e proteine e dati sull' espressione di DNA e proteine.

Le principali banche dati primarie sono: la EMBL [2] datalibrary, la GenBank [3] e la DDBJ [4]. La EMBL datalibrary è la banca dati europea costituita nel 1980 nel laboratorio Europeo di Biologia Molecolare di Heidelberg (Germania). La GenBank è la corrispondente banca americana costituita nel 1982 e la DDBJ è la corrispondente Giapponese. Fra le tre banche dati è stato stipulato un accordo internazionale per cui il contenuto dei dati di sequenza presenti nelle tre banche dati è quasi del tutto coincidente in quanto gli aggiornamenti quotidiani apportati in ciascuna banca dati vengono automaticamente trasmessi alle altre due.

Le banche dati specializzate si sono sviluppate successivamente e raccolgono insiemi di dati omogenei dal punto di vista tassonomico e/o funzionale disponibili nelle Banche dati Primarie e/o in Letteratura, o derivanti da vari approcci sperimentali, rivisti e annotati con informazioni di valore aggiunto.

1.2 Strumenti bioinformatici

Una volta che i dati sono stati archiviati nelle banche date biologiche è necessario utilizzare alcuni strumenti bioinformatici in modo tale da ricavarne informazioni. Essi si sono sviluppati in base a questi tre processi biologici fondamentali:

- la sequenza del DNA codificante, determina la sequenza aminoacidica[2] della proteina (mediante il processo della sintesi proteica);
- la sequenza aminoacidica determina la struttura tridimensionale della proteina;
- la struttura tridimensionale della proteina ne determina la funzione.

La bioinformatica ha focalizzato la sua analisi su dati relativi a questi processi, e di conseguenza le banche dati costituiscono un potente supporto per una vasta gamma di ricerche quali, ad esempio:

- data una sequenza di acidi nucleici o proteica trovare una sequenza simile in banca dati;
- data una struttura proteica trovare, in banca dati, una struttura simile ad essa;
- data una sequenza proteica prevedere una possibile struttura tridimensionale.

2 Gli amminoacidi sono, tra le altre cose, gli elementi costitutivi delle proteine.

I principali strumenti possono essere cosi organizzati:

- *Ricerca di sequenze simili*

  Sequenze omologhe sono sequenze che hanno un gene ancestrale comune. Il grado di similarità fra due sequenze può essere misurato mentre l'omologia è un dato qualitativo.

  Esistono una serie di strumenti, come ad esempio il software BLAST[3] [5], che possono essere utilizzati per identificare similarità fra nuove sequenze con funzione e struttura sconosciuta e sequenze (archiviate nelle banche dati) la cui struttura e funzione sono note.

- *Studio della funzione delle proteine*

  Questo gruppo di programmi permette di utilizzare una sequenza per estrarre informazioni su domini strutturali dalle banche dati specializzate. Questo potrebbe essere di aiuto per avere informazioni sulla funzione di proteine ignote.

- *Analisi delle strutture*

  Questi strumenti permettono di comparare una struttura con una banca dati di strutture note. Molto spesso proteine con struttura simile hanno una stessa funzione, quindi determinare la struttura secondaria/terziaria è cruciale per capire la funzione.

- *Analisi della sequenza primaria*

  Identificare/analizzare l'evoluzione, identificare mutazioni, regioni idrofobiche o altre proprietà che permettano di capire la funzione della proteina.

3 BLAST: acronimo di Basic Local Alignment Search Tool. E' un programma per la ricerca di omologie locali di sequenza. BLAST può eseguire migliaia di confronti fra sequenze in pochi minuti perciò permette di confrontare in poco tempo una sequenza query con l'intero database per ricercare tutte le sequenze simili ad essa.

1.3 Principali applicazioni della bioinformatica

Le applicazioni della bioinformatica sono numerosissime. Si ritiene che molte malattie siano associate ad una componente genetica. La malattia, infatti, può essere ereditaria (sono note circa 3000- 4000 malattie genetiche come la fibrosi cistica, alcune forme di diabete, ecc.) oppure essere il risultato di fattori ambientali che causano alterazioni del genoma (tumori, malattie cardiache, ecc). Una branca della bioinformatica studia quali geni siano associati a diverse malattie per capirne più chiaramente le basi molecolari con lo scopo di migliorarne la prevenzione e la cura.

2. Ontologia informatica

Nel campo dell'informatica, una ontologia è il tentativo di formulare uno schema concettuale esaustivo e rigoroso nell'ambito di un dato dominio; si tratta generalmente di una struttura dati gerarchica che contiene tutte le entità rilevanti, le relazioni esistenti fra di esse, le regole, gli assiomi, ed i vincoli specifici del dominio. L'uso del termine "ontologia" nell'informatica è derivato dal precedente uso dello stesso termine in filosofia, dove ha il significato dello studio dell'essere o dell'esistere, così come dalla filosofia sono tratti i concetti fondamentali di categoria e di relazione.[4]

2.1 Gene Ontology (GO)[5]

I ricercatori spendono moltissimo tempo e sforzo nella ricerca di tutte le informazioni disponibili su ogni piccola area di ricerca. Ciò è ulteriormente ostacolato dalla diversità terminologica utilizzata, che limita l'efficacia del lavoro dei ricercatori stessi e dei computer.

Ad esempio, se si sta cercando un nuovo "target" per un antibiotico, si potrebbe voler trovare tutti i prodotti genici coinvolti nella sintesi di proteine batteriche, che sono significativamente diverse per sequenza o struttura da quelle umane. Se esiste un database nel quale viene descritto come queste molecole vengono coinvolte nella 'traduzione' (translation), e se ne esiste un secondo che descrive la stessa cosa, ma con la frase "sintesi proteica" (protein synthesis), sarà difficile per un essere umano se non addirittura impossibile per un computer trovare termini equivalenti dal punto di vista funzionale.

4 http://it.wikipedia.org/wiki/Ontologia_(informatica)
5 http://www.geneontology.org/

Ecco allora, che per risolvere il bisogno di catalogare e descrivere i prodotti genici e dunque di standardizzare tutta la nomenclatura biologica è stato fondato il progetto GeneOntology (GO) [6] che fornisce una definizione precisa del ruolo svolto dalle singole proteine tramite un vocabolario (delle ontologie) che consente di definire in modo corretto e non arbitrario i processi biologici nei quali una proteina partecipa, le sue funzioni molecolari e la sua localizzazione cellulare.

Il progetto è cominciato nel 1998 come collaborazione tra gli enti che curavano tre diversi database di specie modello:

- FlyBase (Drosophila) [7],
- the Saccharomyces Genome Database (SGD) [8],
- the Mouse Genome Database (MGD) [9].

Da quel momento Il "consorzio GO" è cresciuto includendo molti altri database (piante, animali, ecc.) .

Il database Gene Ontology (GO) è dunque un “vocabolario strutturato”, in continua crescita, che permette un’assegnazione dinamica del significato funzionale delle proteine della cellula. Questo vocabolario è definito controllato e dinamico poiché viene scritto e aggiornato da un gruppo di persone incaricate.

2.2 Come funziona Gene Ontology

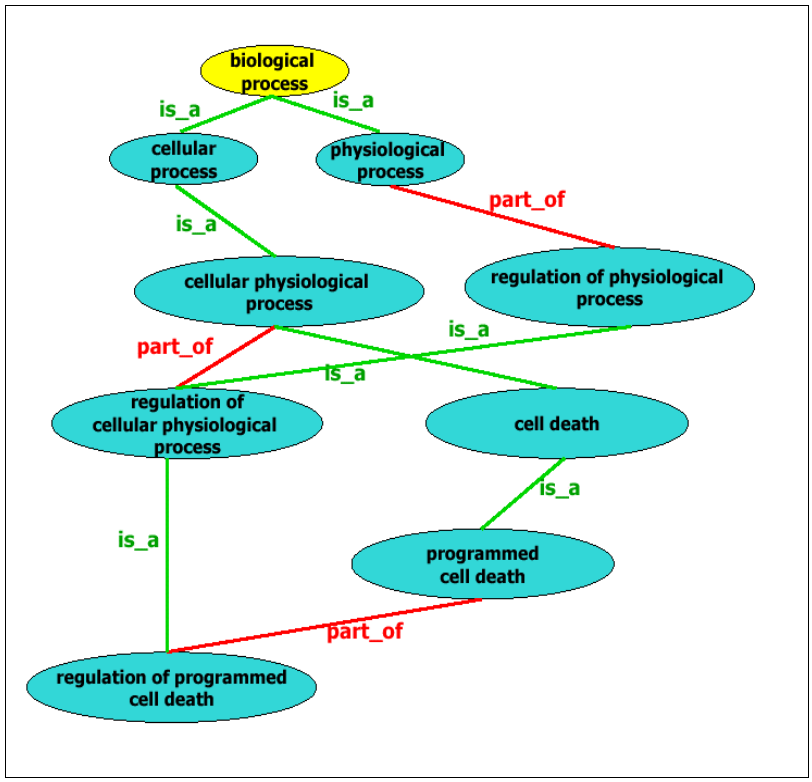
Un gene può avere una o più funzioni cellulari, può essere coinvolto in diversi processi biologici e potrebbe essere associato a diversi compartimenti cellulari.

Per questo motivo la "Gene Ontology" è divisa in tre categorie:

1. *Processo biologico*: un processo biologico è una serie di eventi compiuti da uno o più insiemi ordinati di funzioni molecolari. Esempi di processo biologico sono i processi fisiologici cellulari o di trasduzione del segnale. Non è facile distinguere tra un processo biologico molecolare e funzionale, ma la regola generale è che un processo deve avere più di una procedura distinta.
2. *Funzione molecolare*: si riferisce all'attività biochimica di un gene. Molto spesso questa definizione si riferisce alla capacità di un prodotto genico (o di un complesso proteico) di esprimersi in un determinato contesto metabolico.
3. *Componente cellulare*: si riferisce alla zona specifica della cellula dove un prodotto genico è attivo.

2.3 DAG (Grafo Aciclico Diretto)

Il grafo Aciclico Diretto è la forma rappresentativa usata in GO. Il DAG è una forma di grafico che differisce da una normale gerarchia poiché ogni termine può avere più padri e in esso possono esistere molteplici percorsi da un termine qualsiasi al termine radice. Ciascun vocabolo della GO rappresenta un nodo del DAG al quale vi è associato un identificativo (GO ID).



Esempio di grafico aciclico diretto (DAG)

Un termine di gene ontology molto "generico" contiene al suo interno più termini di gene ontology via via più specifici. Questo fa sì che man mano che si procede "verso il basso" le definizioni diventano sempre più precise mentre i geni che soddisfano quella descrizione sempre meno. Questo albero può quindi essere "letto" a più livelli, da quelli più generali, che stanno alla radice, a quelli via via sempre più specifici che stanno sulle foglie.

3. Il formato FASTA

In bioinformatica, il formato FASTA[6] consiste in un file di testo (*.txt oppure *.fas) per mezzo del quale sia sequenze di nucleotidi (dna), che sequenze di amminoacidi vengono rappresentate tramite lettere dell'alfabeto.

La semplicità del formato FASTA rende semplice analizzare e manipolare le sequenze attraverso linguaggi di scripting come Pyton, Perl, ecc..

Una sequenza in formato FASTA consiste in una prima riga (header) contenente una descrizione, seguita da una o più righe della sequenza. Il primo carattere della riga di descrizione deve cominciare con il simbolo di maggiore “>” e non ci dovrebbero essere spazi tra il segno maggiore ed il primo carattere della descrizione.

Un esempio di sequenza in formato FASTA è:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

I caratteri che compongono la sequenza devono essere nel formato standard di rappresentazione di amminoacidi e acidi nucleici, ovvero il formato IUB/IUPAC[7] con queste eccezioni:

- i caratteri minuscoli sono accettati ma vengono trasformati in caratteri maiuscoli;
- può essere usato un singolo trattino che nel caso, rappresenta un gap;
- in sequenze di amminoacidi sono accettati i caratteri U e *

6 http://www.ncbi.nlm.nih.gov/blast/fasta.shtml
7 http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html

Prima di eseguire una ricerca, tutti i caratteri numerici devono essere rimossi o rimpiazzati dai caratteri appropriati, ad esempio N per le sequenze di nucleotidi sconosciute e X per le sequenze di amminoacidi sconosciute.

La tabella dei caratteri riconosciuti per le sequenze di nucleotidi è:

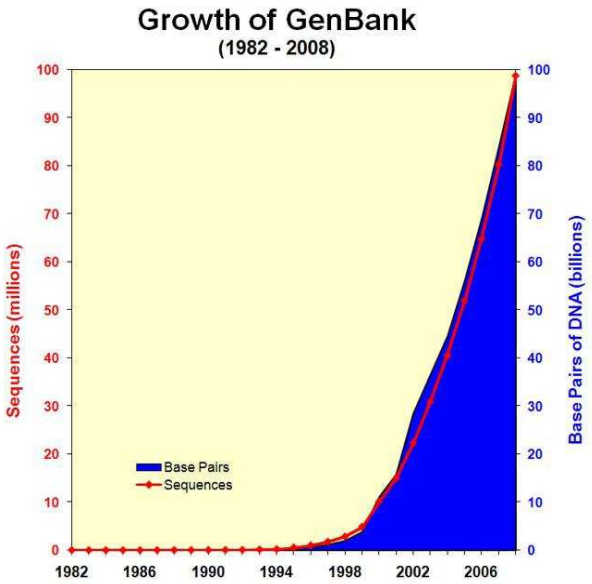
A --> adenosine	M --> A C (amino)
C --> cytidine	S --> G C (strong)
G --> guanine	W --> A T (weak)
T --> thymidine	B --> G T C
U --> uridine	D --> G A T
R --> G A (purine)	H --> A C T
Y --> T C (pyrimidine)	V --> G C A
K --> G T (keto)	N --> A G C T (any)
	- gap of indeterminate length

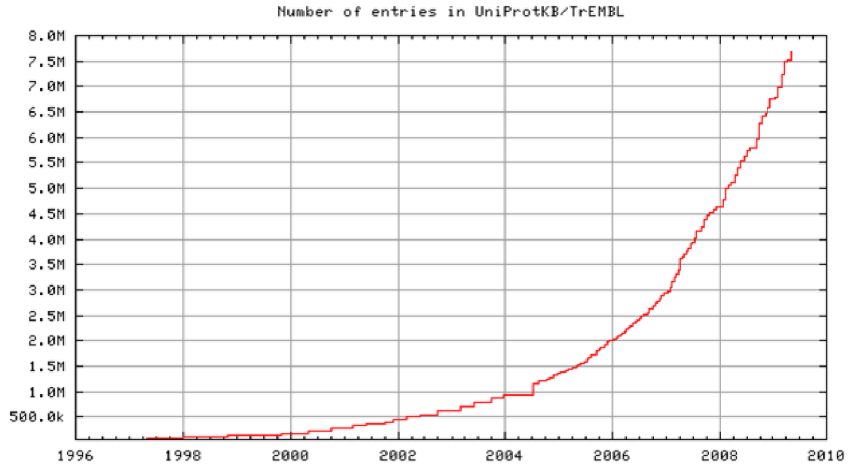
Mentre per gli amminoacidi:

A	alanine	P	proline
B	aspartate or asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate or glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

4. Situazione banca dati di sequenze

Il quantitativo di sequenze ed informazioni associate, disponibili nei database pubblici ha avuto in questi ultimi anni, una crescita esponenziale.



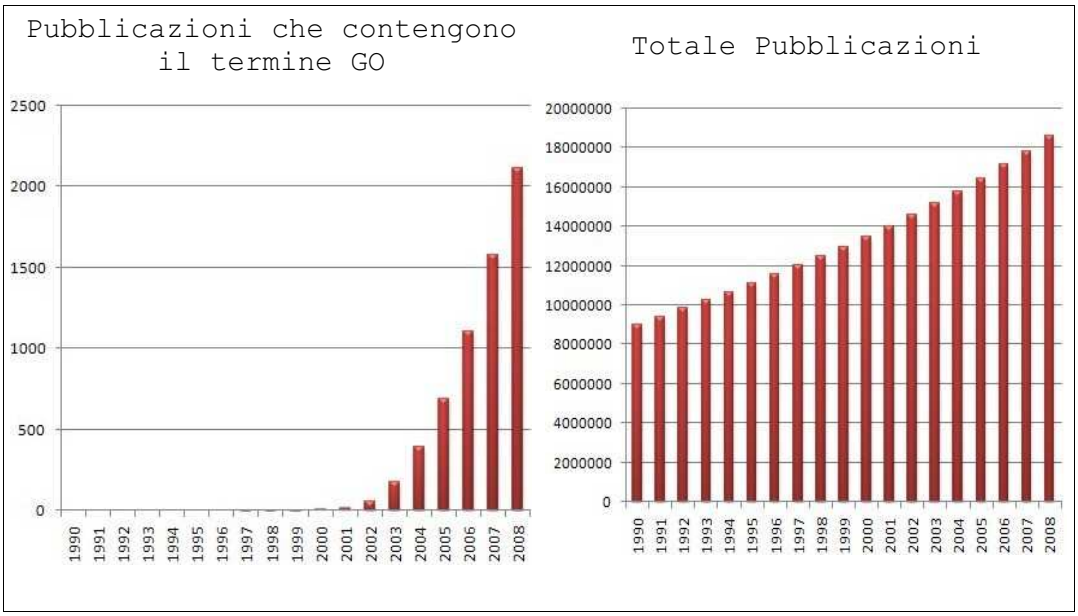


Come si può notare dai grafici qui riportati, sono state inserite nei database oltre cento milioni di sequenze di DNA[8] e oltre otto milioni di proteine[9].

8 Dati da GenBank - http://www.ncbi.nlm.nih.gov/Genbank/
9 Dati da UniProt - http://www.uniprot.org/

Attualmente, sono in corso, e sono già stati ultimati, progetti di sequenziamento sistematico di interi genomi di oltre cinquemila specie, come ad esempio virus, batteri, ecc...

Anche per le pubblicazioni scientifiche in ambito biomedico, assistiamo ad una crescita importante della letteratura disponibile (statistiche riportate da PubMed[10]):



In tale contesto cercare informazioni utili nell'ambito del proprio studio, diventa non facile e soprattutto dispendioso in termini di tempo.

10 Entrez PubMed (o, più spesso, PubMed) è un database bibliografico contenente informazioni sulla letteratura scientifica biomedica dal 1949 ad oggi.
Sito Web: http://www.pubmed.gov/

Per chiarire meglio, facciamo un esempio: se dovessimo cercare l'enzima “glutathione peroxidase” per capire qual è il suo ruolo, come si comporta etc., utilizzando PubMed quello che otterremmo sarebbero oltre 17000 pubblicazioni scientifiche da consultare!

E` certamente impossibile per uno scienziato leggere tutte le pubblicazioni sul tema ricercato. La troppa informazione, è quasi un ostacolo per la ricerca e l'annotazione[11] di sequenze anonime.

La Gene Ontology, illustrata nel capitolo 2, ha sviluppato quello che può essere considerato uno standard riconosciuto per la classificazione funzionale delle proteine, ossia un vocabolario controllato e strutturalmente organizzato in un grafo (DAG), disegnato per essere facilmente sfruttato da metodi computazionali.

11 L'**annotazione** consiste nella localizzazione dei geni, nella comprensione della loro funzione e di eventuali elementi di regolazione.

Supponendo per esempio, che un laboratorio produca circa mille sequenze da analizzare e supponendo che un scienziato impieghi mediamente 3 giorni per annotare manualmente una sequenza (ipotesi reale), si intuisce istantaneamente che impiegherà circa otto anni per completare il lavoro.
Se poi consideriamo il fatto che in 8 anni la banca dati verrà sicuramente aggiornata, modificata e corretta, il nostro scienziato in questione dovrebbe riprendere le prime sequenze annotate per verificarne l'esattezza in base alle nuove informazioni presenti nella banca-dati e il lavoro sarebbe quindi complesso e per nulla efficace.

5. Materiali e metodi

Un buon metodo per l'annotazione di sequenze anonime in tempi ragionevolmente brevi è quello di eseguire una ricerca per similarità di sequenza. Le ricerche di similarità si basano sul confronto sistematico di una sequenza di partenza (*query*) con ognuna delle sequenze contenute nel database.

5.1 BLAST[12]

Alcuni programmi, come BLAST,[13] sono estremamente efficienti e sono in grado di portare a termine una ricerca di similarità in pochi secondi.
BLAST, verosimilmente ad un classico motore di ricerca, come ad esempio Google, Yahoo, eccetera, restituisce una lista di risultati che chiameremo *hits,* ordinati in base al livello di significatività *e-value* rilevato, come visualizzato nell'esempio della tabella seguente:

Sequences producing significant alignments:	**e-value**
Q7YSD4_LOXAF Q7YSD4 Recombination activating gene-1	e-124
Q7YSD3_ELEMA Q7YSD3 Recombination activating gene-1	e-123
RAG1_HUMAN P15918 V(D)J recombination-activating protein	e-121
…	…

L' *e-value*[14] (expectation value) è un indice, che rappresenta la probabilità di quanto l'allineamento ottenuto sia casuale. Più il valore è basso, più il risultato è attendibile. Normalmente l'allineamento viene considerato significativo per un valore soglia (cutoff) di 0.05

12 **B**asic **L**ocal **A**lignment **S**earch **T**ool - http://blast.ncbi.nlm.nih.gov/
13 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
14 http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html

Ad esempio:

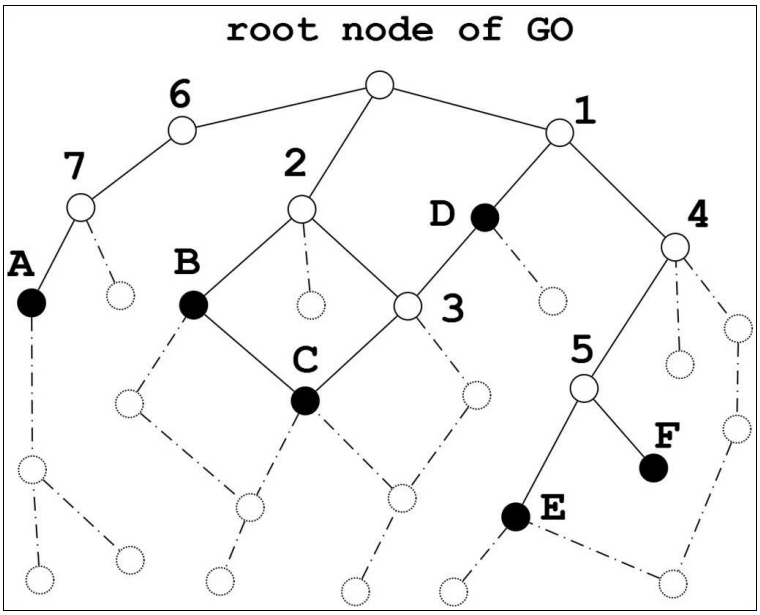
```
Query: 1   HCDIGNAAEFYKIFQLEIGEAYKNPDASKEERKRWQATLDKHLRKRMNLKPIMRMNGNFA 180
           HCDIGNAAEFYKIFQLEIGEAYKNPDASKEERKRWQATLDKHLRKRMNLKPIMRMNGNFA
Sbjct: 1   HCDIGNAAEFYKIFQLEIGEAYKNPDASKEERKRWQATLDKHLRKRMNLKPIMRMNGNFA 60

Query: 181 RKLMTKETVEAVCELIPSEERHEALRELIDLYLKMKPVWRSSCPAKECPESLCQYSFNSQ 360
           RKLMTKETVEAVCELIPSEERHEALRELIDLYLKMKPVWRSSCPAKECPESLCQYSFNSQ
Sbjct: 61  RKLMTKETVEAVCELIPSEERHEALRELIDLYLKMKPVWRSSCPAKECPESLCQYSFNSQ 120

Query: 361 RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM 540
           RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM
Sbjct: 121 RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM 180

Query: 541 NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAHKL 642
           NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAHKL
Sbjct: 181 NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAHKL 214
```

Le sequenze sopra riportate si appaiano esattamente.

Tuttavia, l'utilizzo di BLAST per annotare sequenze di proteine, non è sufficiente, poiché nella bancadati, sono presenti degli errori di annotazione.
Uno studio [10] stima che la percentuale di errori si aggira attorno ad un 30-40% sul totale delle sequenze.
Gli stessi risultati di ricerca di similarità di sequenza con BLAST sono soggetti ad errori. Il valore soglia, scelto per definire tutte quelle sequenze che sono funzionalmente correlate alla query, è un'approssimazione e può dar luogo a falsi positivi o falsi negativi.

5.2 ARGOT

Lo scopo di ARGOT è quello, a partire dai risultati ottenuti con il BLAST, di migliorare la qualità, cercando di estrapolare una maggiore conoscenza.
ARGOT si propone di annotare le sequenze basandosi su un algoritmo che implementa un sistema di pesi e la similarità semantica dei termini GO e non è vincolato dalla soglia imposta dall'e-value del BLAST.
Come concetto, la similarità semantica[15] [11, 12, 13] è ampiamente utilizzata nell'analisi del linguaggio naturale, poiché possiede una metrica, che riflette la distanza o vicinanza tra concetti, ed è dunque gestibile tramite algoritmi informatici.
Il miglioramento proposto dall'algoritmo utilizzato con ARGOT consta in una serie di passi:
1) Viene popolato il grafo dei termini GO estratti dagli hits risultanti dalla ricerca con BLAST .



Ogni nodo contiene l'*e-value* del corrispondente hit e qualora vi siano più

15 La semantica è quella parte della linguistica che studia il significato delle parole (semantica lessicale), degli insiemi delle parole, delle frasi (semantica frasale) e dei testi.

hits che corrispondono ad un nodo in GO, il valore inserito è la somma dei logaritmi di ciascun e-value.

2) Il grafo viene poi potato, eliminando tutti i nodi GO ai quali non corrisponde alcun hit e che non sono nodo padre di un nodo popolato.

Nella figura di pagina precedente, i pallini neri con le lettere alfabetiche rappresentano i nodi GO ai quali corrisponde uno o più hits e quelli vuoti con la numerazione rappresentano i nodi padri. Mentre con la linea spezzata, sono rappresentati i nodi ai quali non è associato alcun hits e che quindi, anche se sono annotati, non sono considerati nelle annotazioni successive..

3) I termini del grafo, sono successivamente clusterizzati in funzione della loro similarità semantica, giungendo così a proporre quella che potrebbe essere l'annotazione più probabile da trasferire alla query inserita.

Al termine dell'elaborazione vengono prodotti 3 indici statistici [1]:

- Total Score (TS)
- Internal Confidence (InC)
- Absolute Confidence (AC)

GO ID	NAME	TOTAL SCORE	INTERNAL CONFIDENCE	ABSOLUTE CONFIDENCE
GO:0003964	RNA-directed DNA polymerase activity	17.6084921709961	0.305652528359139	0.211109779196817
GO:0003723	RNA binding	15.0861809908263	0.368608224185203	0.211109779196817
GO:0004523	ribonuclease H activity	7.65989447112315	0.1489353793007	0.211109779196817
GO:0004190	aspartic-type endopeptidase activity	6.16231405337439	0.133394960067578	0.211109779196817
GO:0003677	DNA binding	3.56615710965915	0.368608224185203	0.211109779196817
GO:0008270	zinc ion binding	1.11862157536	0.0257173391078595	0.0517283115926432

InC e AC sono indici normalizzati nell'intervallo [0,1] e servono a valutare la significatività statistica del termine proposto.

AC è una misura assoluta che fornisce la probabilità di quanto il termine proposto sia lontano dal punteggio teorico massimo.

InC è invece calcolato in relazione a tutti gli hits prodotti da BLAST.
TS è un indice statisticamente più robusto, ricavato da InC, che considera la dispersione delle annotazioni del grafo. Più elevato è il valore dell'indice, maggiore è la significatività dell'annotazione proposta per la query.
ARGOT è implementato in Java ed è liberamente scaricabile per utenti accademici all'indirizzo: http://genomics.research.iasma.it/argot/index.html.

5.3 MySql

MySQL[16] è un database management system (DBMS) relazionale, composto da un client con interfaccia a caratteri e un server, entrambi disponibili sia per sistemi Unix/Linux che per sistemi Windows.
MySQL viene principalmente utilizzato come DBMS, per lo sviluppo di siti e applicazioni web dinamiche. Viene distribuito con licenza GNU-GPL[17] o commerciale

5.4 Apache

Apache HTTP Server[18] o più comunemente Apache è la più diffusa piattaforma web server, con licenza open source a partire dal 1996 .
E' disponibile per i moderni sistemi operativi, come Unix/Linux e WindowsNT.
Apache è un software che realizza le funzioni di trasporto delle informazioni, di internetwork e di collegamento per il protocollo HTTP in pieno rispetto agli standard attuali.[20]

16 http://www.mysql.it/
17 http://www.gnu.org/licenses/gpl.html
18 http://www.apache.org/

5.5 Php

PHP[19,20] (acronimo ricorsivo di PHP Hypertext Preprocessor, preprocessore di ipertesti) è un linguaggio di scripting interpretato, con licenza open-source, originariamente concepito per la realizzazione di pagine web dinamiche. Attualmente è utilizzato principalmente per sviluppare applicazioni web lato server ma può essere usato anche per scrivere script a linea di comando o applicazioni standalone con interfaccia grafica.

Nel gennaio 2005 è stato insignito del titolo di "Programming Language of 2004" dal TIOBE Programming Community Index, classifica che valuta la popolarità dei linguaggi di programmazione sulla base di informazioni raccolte dai motori di ricerca.

Nel 2005 la configurazione LAMP (Linux, Apache, MySQL, PHP) supera il 50% del totale dei server sulla rete mondiale.

19 http://www.wikipedia.it/
20 http://www.php.net/

5.6 Il protocollo SOAP

Soap (Simple Object Access Protocol) è un protocollo di comunicazione per lo scambio di messaggi in ambiente decentralizzato, basato su XML.
Le specifiche SOAP descrivono:

- le convenzioni di formattazione che consentono l'incapsulamento e l'instradamento si un messaggio SOAP;
- regole per come esprimere ed interpretare i dati contenuti nel messaggio;
- un collegamento ad un tipo di trasporto o di protocollo, ad esempio HTTP;
- un meccanismo di RPC;

SOAP utilizza una connessione di tipo *application-to-application* utilizzando teoricamente qualsiasi tipo di protocollo di trasporto, ma per non incorrere in problematiche di firewall e/o proxy viene utilizzato più comunemente il protocollo HTTP.

5.6.1 Struttura di un messaggio SOAP

Un messaggio SOAP altro non è che un documento XML che descrive una richiesta o il risultato di una elaborazione. È costituito dai seguenti elementi:

- Envelope (obbligatorio), rappresenta la radice del documento XML
- Header (opzionale), contiene informazioni globali sul messaggio; ad esempio, nell'header viene specificata la lingua di riferimento del messaggio, la data dell'invio, ecc.
- Body (obbligatorio), contenente il documento vero e proprio, rappresenta la richiesta di elaborazione o la risposta derivata da una elaborazione.

- Fault, se presente nei messaggi di risposta, fornisce informazioni sugli errori che si sono verificati durante l'elaborazione della richiesta.

Figura 5. rappresentazione grafica di un ipotetico messaggio SOAP

5.6.2 Il programma WSNCBIBlast.jar

Nel progetto implementato, la ricerca nella banca dati delle proteine è gestita dallo script java WSNCBIBlast.jar[21] il quale fa parte di una serie di web services messi a disposizione dall'Istituto Europeo di Bioinformatica.[22]

Un esempio di file risposta è il seguente:

```
<?xml version="1.0"?>
<EBIApplicationResult xmlns="http://www.ebi.ac.uk/schema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://www.ebi.ac.uk/schema/ApplicationRes
ult.xsd">
<Header>
      <program name="NCBI-blastp" version="2.2.19 [Nov-02-2008]"
citation="PMID:9254694"/>
      <commandLine command="/ebi/extserv/bin/ncbi-blast-
2.2.19/bin/blastall -p blastp -d $IDATA_CURRENT/blastdb/uniprot -i
/ebi/extserv/blast-work/interactive//blast-20090516-1130038729.input -M
BLOSUM62 -b 100 -v 100 -e 1.0 -X 0 -G 11 -E 1 -a 16 -L 1,391 -m 0 -gt -F
F -C F "/>
```

21 http://www.ebi.ac.uk/Tools/webservices/clients/ncbiblast
22 http://www.ebi.ac.uk/

```
      <parameters>
            <sequences total="1">
                  <sequence number="1" name="Sequence" type="p"
length="391"/>
            </sequences>
            <databases total="1" sequences="8189860"
letters="2688696465">
                  <database number="1" name="uniprot" type="p"
created="2009-05-07T09:45:00+01:00"/>
            </databases>
            <gapOpen>11</gapOpen>
            <gapExtension>1</gapExtension>
      </parameters>
</Header>
<SequenceSimilaritySearchResult>
      <hits total="1">
            <hit number="1" database="uniprot" id="CSK21_HUMAN"
ac="P68400" length="391" description="Casein kinase II subunit alpha
OS=Homo sapiens GN=CSNK2A1 PE=1 SV=1">
                  <alignments total="1">
                        <alignment number="1">
                              <score>2083</score>
                              <bits>806</bits>
                              <expectation>0.0</expectation>
                              <identity>100</identity>
                              <positives>100</positives>
<querySeq start="1" end="391">
MSGPVPSRARVYTDVNTHRPREYWDYESHVVEWGNQDDYQLVRKLGRGKYSEVFEAINITNNEKVVVKILKPV
KKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYMYEILKALDY
CHSMGIMHRDVKPHNVMIDHEHRKLRLIDWGLAEFYHPGQEYNVRVASRYFKGPELLVDYQMYDYSLDMWSLG
CMLASMIFRKEPFFHGHDNYDQLVRIAKVLGTEDLYDYIDKYNIELDPRFNDILGRHSRKRWERFVHSENQHL
VSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSANMMSGISSVPTPSPLG
PLAGSPVIAAANPLGMPVPAAAGAQQ</querySeq>
<pattern>MSGPVPSRARVYTDVNTHRPREYWDYESHVVEWGNQDDYQLVRKLGRGKYSEVFEAINITNNEK
VVVKILKPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFYM
YEILKALDYCHSMGIMHRDVKPHNVMIDHEHRKLRLIDWGLAEFYHPGQEYNVRVASRYFKGPELLVDYQMYD
YSLDMWSLGCMLASMIFRKEPFFHGHDNYDQLVRIAKVLGTEDLYDYIDKYNIELDPRFNDILGRHSRKRWER
FVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSANMMSGIS
SVPTPSPLGPLAGSPVIAAANPLGMPVPAAAGAQQ</pattern>
                              <matchSeq start="1"
end="391">MSGPVPSRARVYTDVNTHRPREYWDYESHVVEWGNQDDYQLVRKLGRGKYSEVFEAINITNNE
```

```
KVVVKILKPVKKKKIKREIKILENLRGGPNIITLADIVKDPVSRTPALVFEHVNNTDFKQLYQTLTDYDIRFY
MYEILKALDYCHSMGIMHRDVKPHNVMIDHEHRKLRLIDWGLAEFYHPGQEYNVRVASRYFKGPELLVDYQMY
DYSLDMWSLGCMLASMIFRKEPFFHGHDNYDQLVRIAKVLGTEDLYDYIDKYNIELDPRFNDILGRHSRKRWE
RFVHSENQHLVSPEALDFLDKLLRYDHQSRLTAREAMEHPYFYTVVKDQARMGSSSMPGGSTPVSSANMMSGI
SSVPTPSPLGPLAGSPVIAAANPLGMPVPAAAGAQQ</matchSeq>
                        </alignment>
                    </alignments>
                </hit>
</hits>
</SequenceSimilaritySearchResult>
</EBIApplicationResult>
```

5.7 GOA e il file OBO

Tutte le annotazioni della Gene Ontology sono inserite nel database "GOA"[14] ovvero UniProt GO Annotation [15].

Le annotazioni sono manutenute da dei curatori, ciascuno dei quali è specializzato in un particolare ambito scientifico.

Nell'immagine seguente un esempio:

ID	ACCID	DB_Object_ID	GO_ID	Evidence	Aspect	DB_Object_Name
11501	A0A7C2	recA	GO:0006259	IEA	P	RecA recombinase
11502	A0A7C2	recA	GO:0008094	IEA	F	RecA recombinase
11503	A0A7C2	recA	GO:0017111	IEA	F	RecA recombinase
11504	A0A7D4	A0A7D4	GO:0006231	IEA	P	Putative thymidylate synthase
11505	A0A7D4	A0A7D4	GO:0008168	IEA	F	Putative thymidylate synthase
11506	A0A7D4	A0A7D4	GO:0016740	IEA	F	Putative thymidylate synthase
11507	A0A7D4	A0A7D4	GO:0050660	IEA	F	Putative thymidylate synthase
11508	A0A7D4	A0A7D4	GO:0050797	IEA	F	Putative thymidylate synthase
11509	A0A7D4	A0A7D4	GO:0050797	IEA	F	Putative thymidylate synthase
11510	A0A7D8	A0A7D8	GO:0003677	IEA	F	Prophage antirepressor

Il file OBO[23], è un file di puro testo che descrive la rappresentazione del Grafo Aciclico Diretto. Con lo scopo di:

- essere leggibile per l'essere umano;
- facilità di parsing per gli algoritmi informatici;
- essere estensibile;
- avere una ridondanza minima;

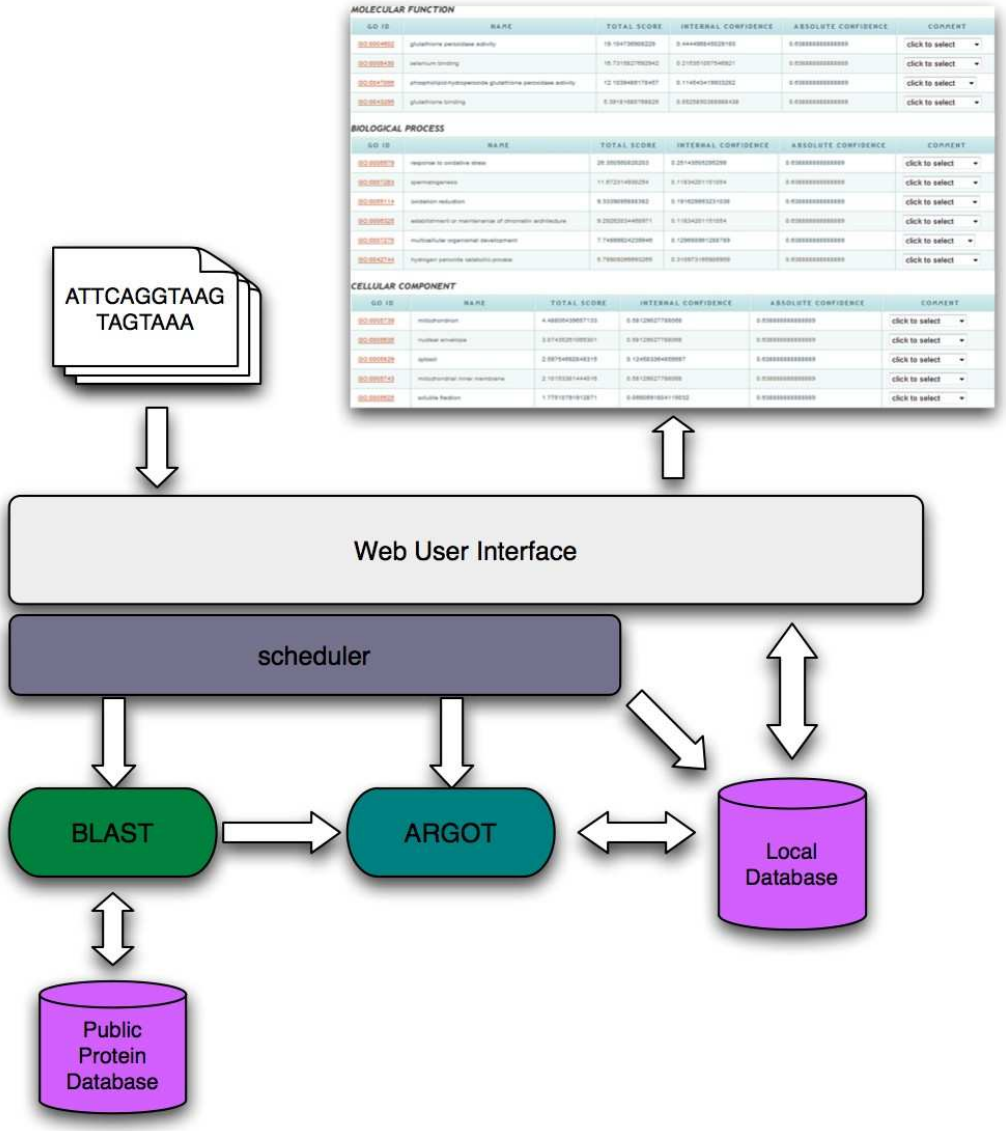
Nella tabella seguente, è visualizzato un termine GO preso dal file OBO:

```
[Term]
id: GO:0016049
name: cell growth
namespace: biological_process
def: "The process by which a cell irreversibly increases in size
over time by accretion and biosynthetic production of matter
similar to that already present." [GOC:ai]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "cell expansion" RELATED []
synonym: "cellular growth" EXACT []
synonym: "growth of cell" EXACT []
is_a: GO:0009987 ! cellular process
is_a: GO:0040007 ! growth
relationship: part_of GO:0008361 ! regulation of cell size
```

23 http://www.geneontology.org/

6. Risultati

6.1 Lo schema logico di funzionamento



Nello schema proposto è riassunto il funzionamento dell'applicativo. Si inizia con inserimento nel sistema, attraverso l'interfaccia utente via web, di una o più sequenze da analizzare. Gli script memorizzano la sequenza e vari parametri nel database locale.

Successivamente lo scheduler fa partire il processo BLAST[24] che riceve in input la sequenza “query” introdotta dall'utente, esegue la ricerca sul database pubblico e restituisce la lista degli hits in formato sia testuale che xml. I files vengono salvati in un'apposita cartella del filesystem.

A questo punto lo scheduler lancia il processo ARGOT, che riceve in input la lista dei risultati del BLAST e utilizzando una copia in locale del database delle annotazioni della Gene Ontology, memorizza i risultati ottenuti nella tabella Argo_res.

Settimanalmente lo scheduler provvede a mantenere aggiornati, la tabella delle annotazioni GO e il file OBO.

L'utente in ogni momento, attraverso l'interfaccia web, può conoscere lo stato di avanzamento dell'elaborazione e infine visualizzare i risultati finali.

24 http://www.ebi.ac.uk/Tools/webservices/services/ncbiblast

6.2 Il database locale

Vengono utilizzate due tabelle di appoggio per le elaborazioni di argot, e altre due per il frontend web, e sono rispettivamente:

La tabella GOA (GeneOntology Annotation) composta, nel momento in cui scrivo, da 40374140 record e contenente tutte le annotazioni GO disponibili per le proteine.

La struttura è la seguente:

- GOA -	
ID	int(11)
ACCID	varchar(20)
DB_Object_ID	varchar(50)
GO_ID	varchar(12)
Evidence	char(3)
Aspect	char(1)
DB_Object_Name	varchar(200)
PRIMARY KEY	(ID)
KEY objectid	(DB_Object_ID)
KEY goid	(GO_ID)
KEY accid	(ACCID)

Nella quale:

- *ID* è la chiave primaria;
- *ACCID* rappresenta un identificatore aggiuntivo per supportare le annotazioni che utilizzano un diverso codice "evidence"
- *DB_Object_ID* è l'identificatore univoco della proteina annotata, ad esempio IPI00188043
- *GO_ID* è l'identificatore nella GO del termine contenuto in DB_Object_ID, ad esempio GO:0005634
- *Evidence* contiene uno di questi codici: EXP, IMP, IC, IGC, IGI, IPI, ISS, IDA, IEP, IEA, TAS, NAS, NR, ND or RCA
- *Aspect* contiene un carattere che rappresenta una delle tre ontologie:

P (processo biologico), F (funzione molecolare), C (componente cellulare)

- DB_Object_Name contiene il nome completo della proteina, se disponibile, altrimenti il campo può essere lasciato vuoto, ad esempio "Cellular tumor antigen p53*"*

La tabella "argo_res", contiene i risultati prodotti da ARGOT per ciascuna elaborazione inviata dall'utente. La tabella potrebbe essere normalizzata, ma per ragioni di carattere operativo si è preferito non procedere.
La struttura è la seguente:

- Argo_res -	
ID	int(11)
GO_ID	varchar(12)
AvgBitscore	double
MinBitscore	double
MaxBitscore	double
IC	double
TotalScore	double
InternalConfidence	double
AbsoluteConfidence	double
RunName	varchar(100)
Comment	text
Aspect	char(1)
PRIMARY KEY(ID) KEY objid (Object_ID) KEY runname (RunName)	

Nella quale:

- *ID* è la chiave primaria;
- *GO_ID* è l'identificatore del termine nella Gene Ontology;
- *AvgBitscore, MinBitscore, MaxBitscore* contengono rispettivamente ill punteggio medio, minimo e massimo ricavato dal BLAST
- *IC* contiene il valore dell'Information Content;
- *TotalScore, InternalConfidence, AbsoluteConfidence* contengono il valore

degli rispettivi indici statistici TS, InC e AC;

- *RunName* contiene l'identificatore del processo inserito dall'utente
- *Comment* contiene la lista degli accession number[25] degli hits inseriti nel nodo GO
- *Aspect* , rappresenta una delle tre ontologie: P , F e C

La tabella di appoggio “jobs”, che contiene in ciascun record, informazioni riguardanti il lavoro ed i parametri inseriti da ciascun utente.

- jobs -
id int(11) ncbi_status enum('waiting','running','retrieving','done','error') argot_status enum('waiting','running','retrieving','done','error') data_ins datetime email varchar(150) ncbi_jobid varchar(30) type enum('protein','dna') abs_confidence float int_confidence float totalScore int(11) description varchar(255)
PRIMARY KEY(id), KEY ncbi_jobid (ncbi_jobid)

Nella quale:

- *id* è la chiave primaria;
- *ncbi_status, blast_status* contengono rispettivamente lo stato dell'avanzamento del processo di ricerca blast e lo stato di alnalisi di argot.

  Di default il campo viene inizializzato allo stato di 'waiting';
- *data_ins* contiene la data di inserimento del lavoro;

25 http://en.wikipedia.org/wiki/Accession_number_(bioinformatics)

- *email* contiene la mail dell'utente;
- *ncbi_jobid* contiene l'identificativo della query inserita utilizzando i webservices dell'Istituto Europeo di Bioinformatica;
- *type* contiene la tipologia di sequenza inserita (proteica o nucleotidica), utilizzato come parametro per cercare su banche dati diverse a seconda della tipologia;
- *abs_confidence, int_confidence, totalScore* contengono i valori di soglia degli indici oltre i quali la visualizzazione verrà colorata;
- *description* contiene la prima riga della sequenza inserita, ovvero secondo il formato FASTA, la descrizione della sequenza;

6.3 Schedulazione ed elaborazione delle attività

L'architettura del sistema è concepita in modo tale da mantenere separato l'aspetto di acquisizione di una o più sequenze da analizzare con l'aspetto di elaborazione, che per alti carichi di lavoro può essere spostata su uno o più server dedicati.

Le operazioni di ricerca presso la banca-dati utilizzando il webservice messo a disposizione da NCBI viene schedulata e gestita nei sistemi GNU/Linux tramite il demone[26] CRON, poiché ad esempio, nel nostro caso non si possono eseguire più di 25 ricerche contemporanee. Anche lo script che gestisce l'elaborazione dei risultati delle ricerche da parte di ARGOT viene schedulato tramite CRON.

6.3.1 Il demone Cron

Cron, permette di eseguire un comando ad intervalli di tempo predefiniti, stabilendo l'ora, i minuti, il giorno della settimana o il mese; l'esecuzione di cron dipende dal file crontab che ne definisce i momenti di attivazione ed i comandi da eseguire.

Ogni riga del crontab ha la seguente forma:

min hour dom month dow user command

dove i primi cinque parametri indicano rispettivamente i minuti, l'ora, il giorno del mese (***d**ay **o**f **m**onth*), il mese ed il giorno della settimana *(**d**ay **o**f **w**eek)* per l'attivazione: se uno di questi parametri viene sostituito con un asterisco, il parametro non verrà considerato. Il parametro “user” indica con quale utente deve essere eseguito il comando specificato nel parametro “command”.

26 I demoni sono tutti quei programmi che sono eseguiti in background; nei sistemi Windows sono chiamati Servizi.

Un caso particolare è specificare il parametro, nella forma */numero che indica di eseguire il comando ogni “numero” cambiamenti dello stato della variabile in questione.

Vediamo un paio di esempi, nei quali il comando da eseguire è stato sostituito con la descrizione dell'operazione:

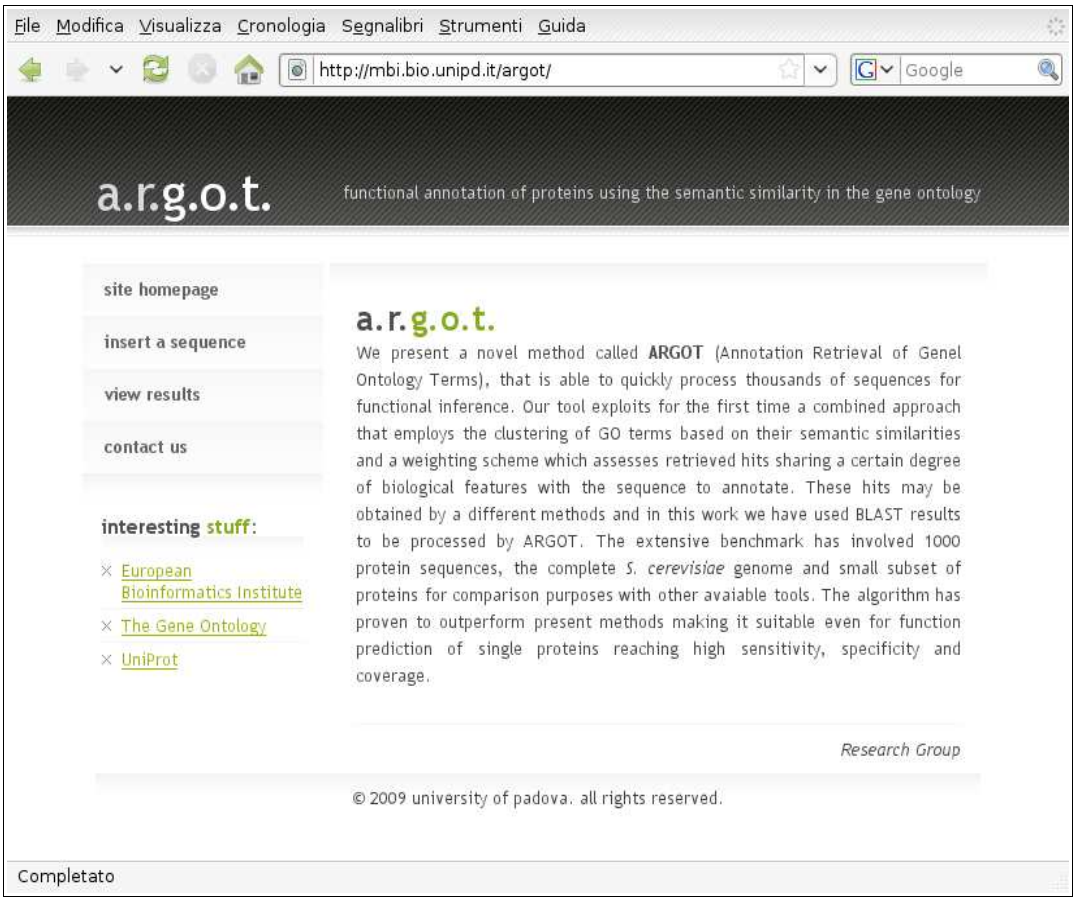
```
#il comando viene eseguito ogni 1 minuto
*/1 * * * * root “controlla se ci sono sequenze appena immesse e fai partire la ricerca BLAST”
```

```
#il comando viene eseguito ogni 2 minuti
*/2 * * * * root “se ci sono risultati pronti analizzali con ARGOT”
```

La flessibilità e le potenzialità di CRON sono elevatissime e ci permettono di effettuare uno scheduling preciso delle nostre attività. Occorre però, prestare attenzione che schedulando troppo spesso un lavoro molto oneroso si rischia, di compromettere le performance della macchina.

6.4. Il front end web

La parte di interazione con l'utente è stata implementata utilizzando la piattaforma L.A.M.P., ovvero GNU/Linux (il sistema operativo), Apache (il web server), MySQL (il database server) e PHP (il linguaggio di scripting).
Il front end è attualmente raggiungibile all'indirizzo:
http://mbi.bio.unipd.it/argot/

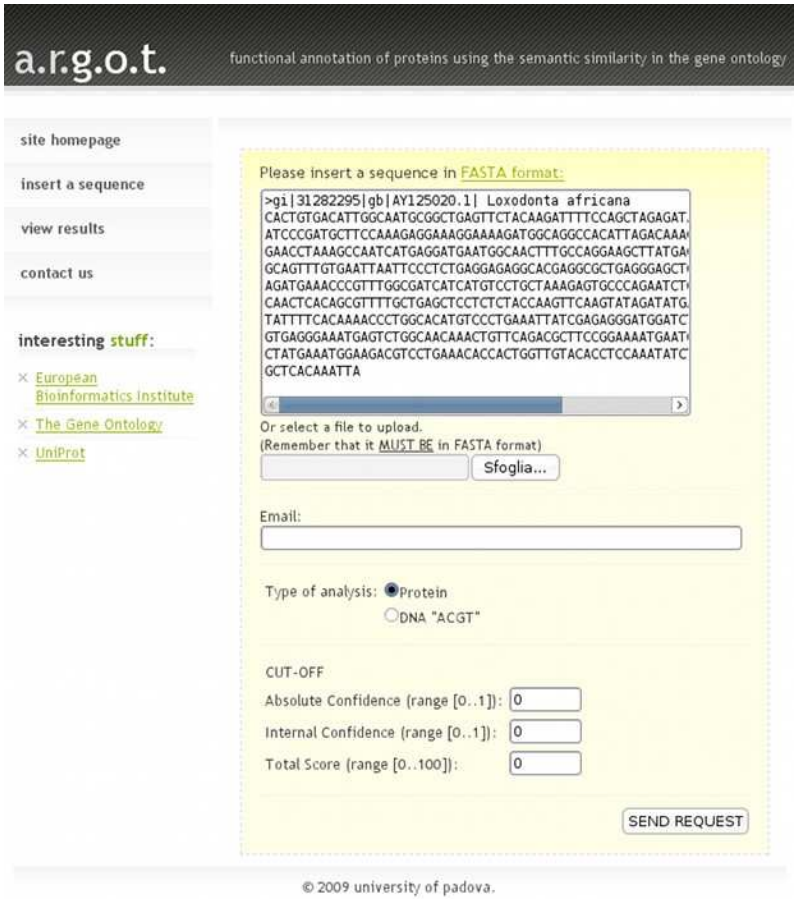


Nell'immagine sopra è visualizzata la pagina principale, tra i link presenti sul menu di sinistra ci soffermeremo su i due di nostro principale interesse, ovvero:

- "*insert a sequence*" per inserire una o più sequenze anonime da elaborare

- "*view results*" per visualizzare i risultati di un'elaborazione

6.4.1 Inserimento di una sequenza



Nella schermata di inserimento di sequenza, è possibile inserire una o più sequenze, tramite il copia-incolla oppure cliccando sul pulsante sfoglia è possibile inserire più sequenze contenute in un file di testo. Le sequenze devono essere obbligatoriamente in formato FASTA.

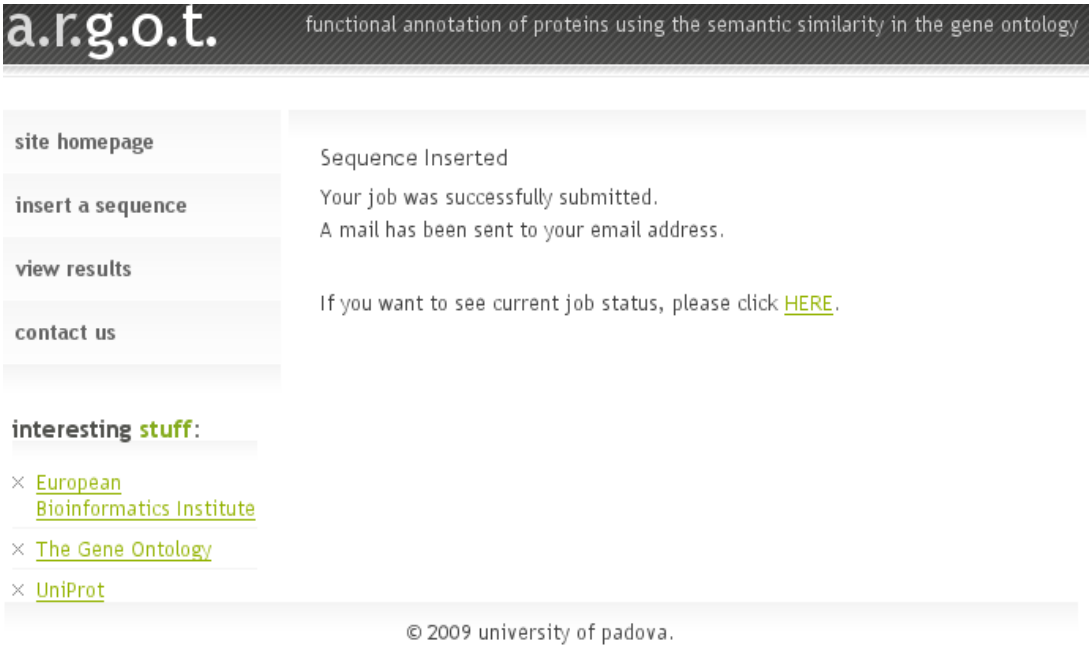
Obbligatoriamente dev'essere inserita poi la email sulla quale una volta terminato il processo si desidera ricevere il messaggio di avvenuta elaborazione. In caso di non inserimento dell'email il sistema visualizza un

messaggio d'errore.

Nella sezione "*Type of analysis*" si specifica la tipologia della sequenza inserita, ovvero se è una proteina oppure una sequenza di DNA.

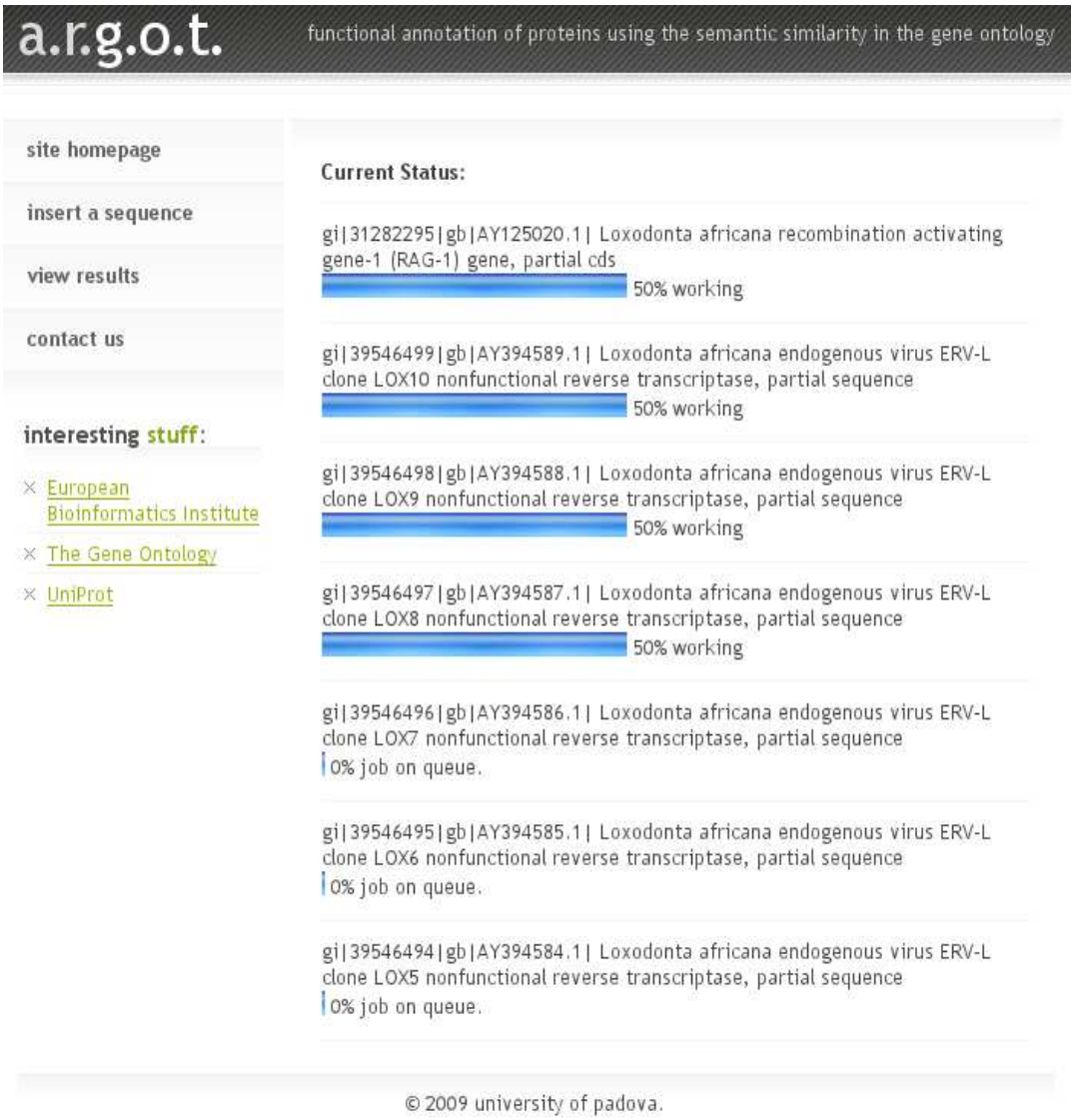
Nella sezione *"Cut-Off"* si possono inserire dei valori soglia, oltre i quali l'applicativo web visualizzerà i dati in colore rosso.

Al termine dell'inserimento della sequenza da analizzare, viene visualizzato il seguente messaggio:



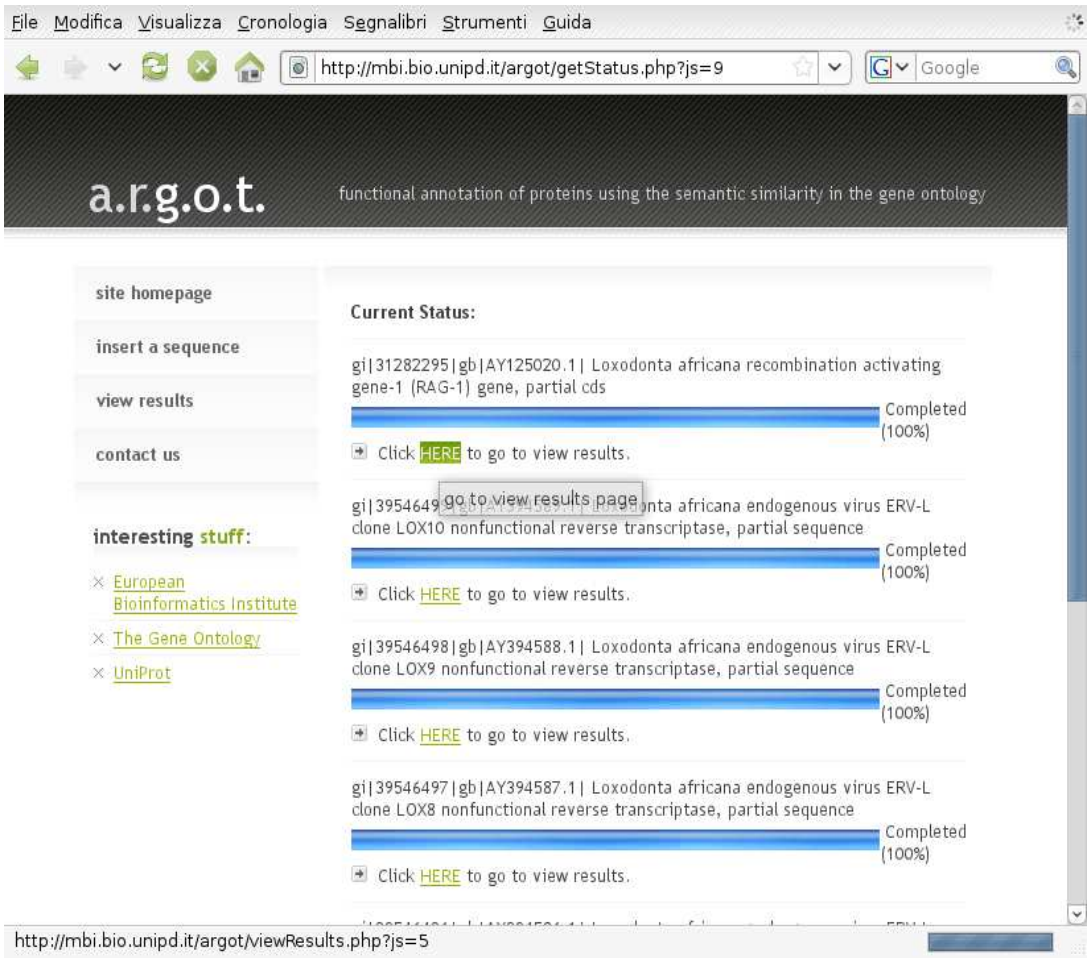
A questo punto l'utente può scegliere di chiudere il browser ed attendere l'email di avvenuta elaborazione, oppure cliccando sul link "*HERE*" può visualizzare in tempo reale l'avanzamento dell'elaborazione.

6.4.2 Elaborazione



Nell'immagine riportata di seguito, viene visualizzato lo stato dell'avanzamento dell'elaborazione per ogni sequenza inserita.

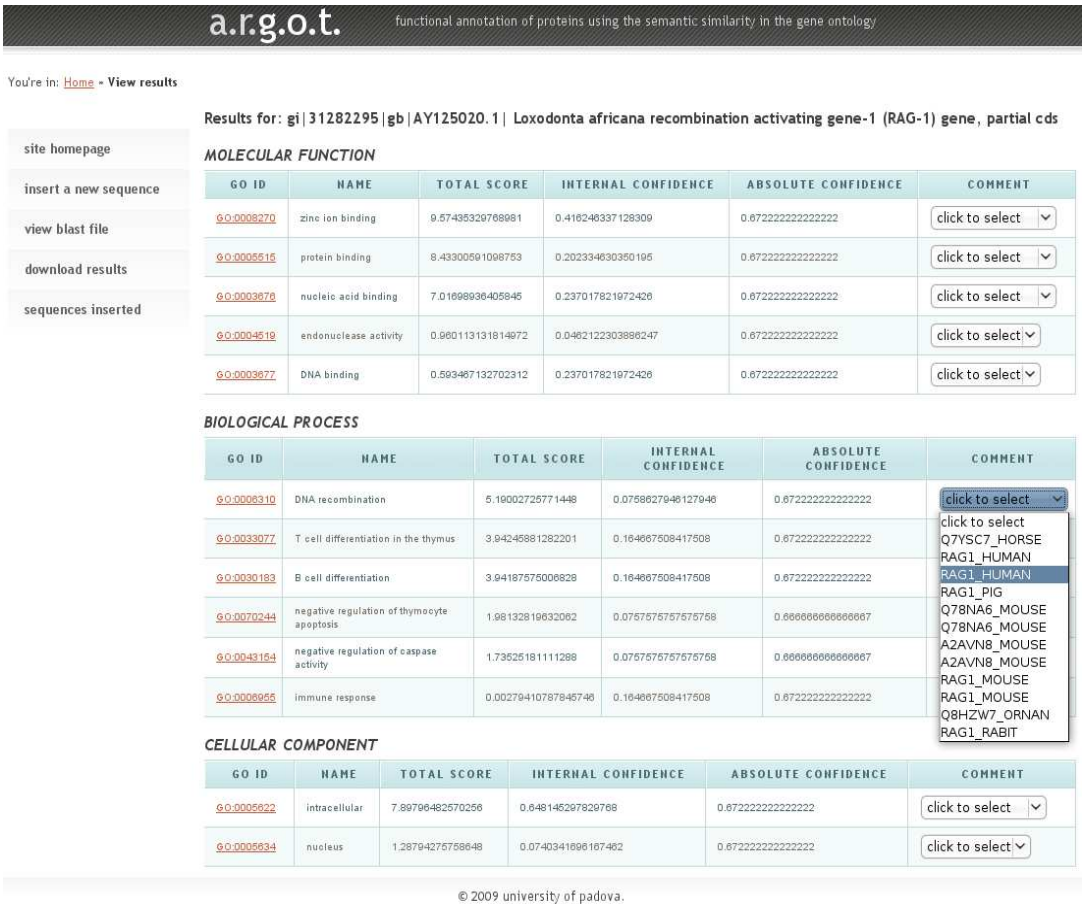
Se è stata inserita più di una sequenza da elaborare verrà visualizzata una schermata intermedia, come riportato di seguito, nella quale è possibile selezionare la sequenza per la quale visualizzare il risultato.

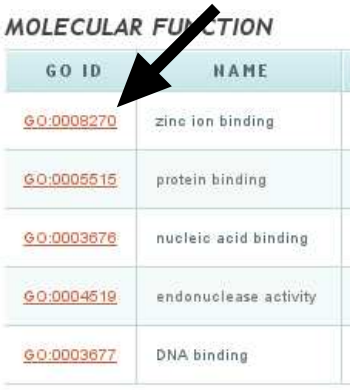


Se invece, è stata inserita una singola sequenza, si andrà direttamente alla pagina di visualizzazione del risultato.

6.4.3 Visualizzazione del risultato

Terminata l'elaborazione ecco di seguito riportata la schermata nella quale sono riportati i risultati raggruppati a seconda della categoria (vedi capitolo 2.2).





Cliccando sul link della colonna "GO ID" verrà aperta una nuova finestra o scheda a seconda del browser, nella quale sarà visualizzato la posizione dell'annotazione nel grafo della Gene Ontology[27]

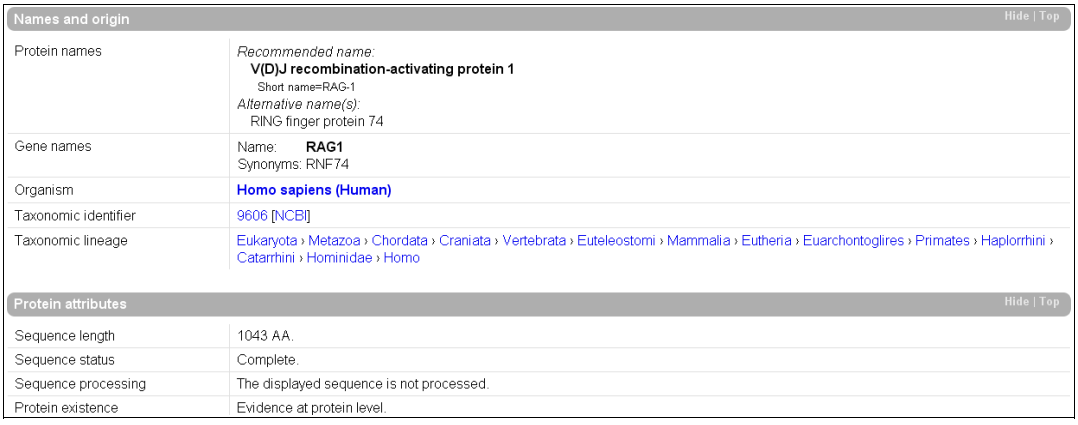
27 http://amigo.geneontology.org/

Nella colonna “COMMENT” è possibile selezionare ciascuna proteina che compone l'annotazione riportata nella colonna GO_ID.

Cliccando sulla voce, verrà visualizzata la pagina (vedi immagine di seguito) che riporta il risultato della prima ricerca con BLAST sul database UNIPROT.

```
>UNIPROT:RAG1_HUMAN P15918 V(D)J recombination-activating protein 1 OS=Homo sapiens
           GN=RAG1 PE=1 SV=1
         Length = 1043

 Score =  436 bits (1122), Expect = e-121
 Identities = 207/212 (97%), Positives = 211/212 (99%)
 Frame = +1

Query: 1    HCDIGNAAEFYKIFQLEIGEAYKNPDASKEERKRWQATLDKHLRKRMNLKPIMRMNGNFA 180
             HCDIGNAAEFYKIFQLEIGE YKNP+ASKEERKRWQATLDKHLRK+MNLKPIMRMNGNFA
Sbjct: 798  HCDIGNAAEFYKIFQLEIGEVYKNPNASKEERKRWQATLDKHLRKKMNLKPIMRMNGNFA 857

Query: 181  RKLMTKETVEAVCELIPSEERHEALRELIDLYLKMKPVWRSSCPAKECPESLCQYSFNSQ 360
             RKLMTKETV+AVCELIPSEERHEALREL+DLYLKMKPVWRSSCPAKECPESLCQYSFNSQ
Sbjct: 858  RKLMTKETVDAVCELIPSEERHEALRELMDLYLKMKPVWRSSCPAKECPESLCQYSFNSQ 917

Query: 361  RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM 540
             RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM
Sbjct: 918  RFAELLSTKFKYRYEGKITNYFHKTLAHVPEIIERDGSIGAWASEGNESGNKLFRRFRKM 977

Query: 541  NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAH 636
             NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAH
Sbjct: 978  NARQSKCYEMEDVLKHHWLYTSKYLQKFMNAH 1009
```

Cliccando invece sul nome della proteina, nell'esempio “P15918”, saranno visualizzati su una nuova finestra i dettagli relativi alla proteina presenti sul database *UniProtKB,*[28] come visualizzato nell'immagine sottostante:



Names and origin Hide | Top

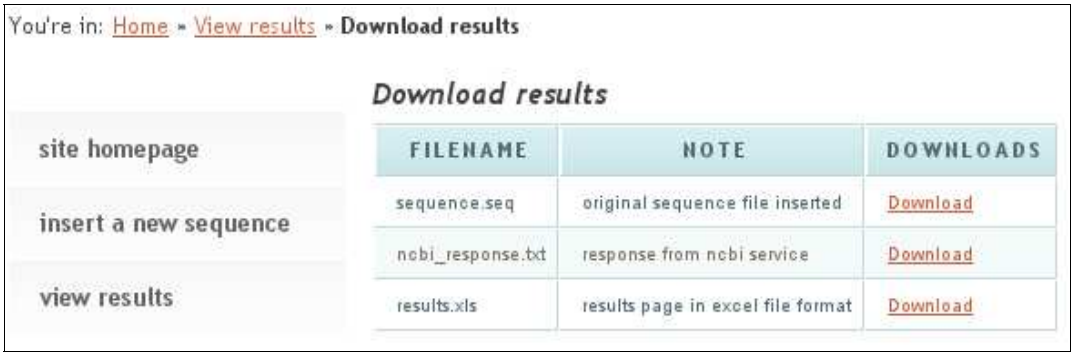
Protein names	*Recommended name:* **V(D)J recombination-activating protein 1** Short name=RAG-1 *Alternative name(s):* RING finger protein 74
Gene names	Name: **RAG1** Synonyms: RNF74
Organism	**Homo sapiens (Human)**
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

Protein attributes Hide | Top

Sequence length	1043 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is not processed.
Protein existence	Evidence at protein level.

28 http://www.uniprot.org/

6.4.4 Download dei risultati



Infine nel menù è presente il link per scaricare nel proprio computer per ulteriori elaborazioni i seguenti files:

- sequence.seq; il file contenente la sequenza query inserita dall'utente
- ncbi_response.txt; il file risposta in formato testo contente il risultato della ricerca sul database Uniprot.
- results.xls; il file in formato excel contenente le tabelle dei risultati visibile nella schermata “Visualizzazione del risultato”

7. Conclusioni

Il lavoro di tesi si è concretizzato nella costruzione di un server web e nella successiva creazione di un'interfaccia utente per l'annotazione funzionale di proteine. Questo perché si è sentita la necessità di procedere nello sviluppo di un servizio aperto e fruibile da tutta la comunità scientifica. Tutto ciò è attualmente, utilizzato per il sequenziamento di vari genomi, tra cui quello della vite [16].
Il cuore del progetto è basato sul software a riga di comando ARGOT, precedentemente sviluppato dai ricercatori del Dipartimento di Chimica Biologica e dall'Istituto di Ricerca FEM-IASMA.

L'originalità del progetto, nel panorama dei sistemi di annotazione, sta nell'implementare per primo un sistema che utilizza la similarità semantica come metodo di selezione della migliore annotazione per la query immessa.
Il tutto è stato costruito e reso disponibile alla comunità scientifica, perché altri progetti, che utilizzano metodi diversi per l'annotazione, come: GOblet[29] [17], PhyDBAC[30] [18], JAFA[31] [19], PFP[32] [20] e GOtcha[33] [21] Blast2GO[34] [22], InterProScan[35] [23], ottengono in un benchmark ristretto ma difficile, di proteine da annotare, dei risultati generalmente peggiori [1]. Inoltre i progetti stessi o sono stati abbandonati, oppure la loro bancadati di riferimento per le annotazioni è aggiornata di rado.
Nel nostro caso, invece, l'aggiornamento avviene settimanalmente in maniera automatica rendendo così più efficace e funzionale allo scopo la ricerca.

29 http://goblet.molgen.mpg.de/cgi-bin/goblet2008/goblet.cgi
30 http://www.igs.cnrs-mrs.fr/phydbac/
31 http://jafa.burnham.org/
32 http://dragon.bio.purdue.edu/pfp/
33 http://www.compbio.dundee.ac.uk/Software/GOtcha/gotcha.html
34 http://blast2go.de/
35 http://www.ebi.ac.uk/Tools/InterProScan/

Vista l'importanza del progetto di tesi e le sopraggiunte richieste di nuove caratteristiche, ne è già stata pianificata una nuova versione, con la possibilità, per l'utente avanzato, di impostare ulteriori parametri di controllo nella fase di ricerca per la migliore annotazione.

8. Bibliografia

1. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S (2009) Rapid Annotation of Anonymous Sequences from Genome Projects Using Semantic Similarities and a Weighting Scheme in Gene Ontology. PLoS ONE 4(2): e4619. Doi:10.1371/ journal.pone.0004619
2. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R. EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res. 2007 Jan;35(Database issue):D16-20.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2009 Jan;37(Database issue):D26-31.
4. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. 2002 Jan 1;30(1):27-30.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402
6. The Gene Ontology project in 2008. Nucleic Acids Res 36: D440–444.
7. Drysdale R; FlyBase Consortium. FlyBase: a database for the Drosophila research community. Methods Mol Biol. 2008;420:45-59.
8. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 1998 Jan 1;26(1):73-9.
9. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT; Mouse Genome Database Group. MGD: the Mouse Genome Database. Nucleic Acids Res. 2003 Jan 1;31(1):193-5.
10. Jones CE, Baumann U, Brown AL (2005) Automated methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics 6:272
11. Lord PW, Stevens RD, Brass A, Goble CA. Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput. 2003:601–612.
12. Lord PW, Stevens RD, Brass A, Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput. pp 601–612.
13. Resnik P (1999) Semantic similarity in a taxonomy: An information-based

measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11: 95–130.

14. Lee V, Camon E, Dimmer E, Barrell D, Apweiler R (2005) Who tangos with GOA? - Use of Gene Ontology Annotation (GOA) for biological interpretation of 'omics' data and for validation of automatic annotation tools. In Silico Biol 5: 5–8.
15. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics. 2009 May 8;10:136.
16. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2: e1326.
17. Groth D, Lehrach H, Hennig S (2004) GOblet: a platform for Gene Ontology annotation of anonymous sequence data. Nucleic Acids Res 32: W313–317.
18. Enault F, Suhre K, Claverie JM(2005) Phydbac ''Gene Function Predictor'': a gene annotation tool based on genomic context analysis. BMC Bioinformatics 6: 247.
19. Friedberg I, Harder T, Godzik A (2006) JAFA: a protein function annotation meta-server. Nucleic Acids Res 34: W379–381.
20. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15: 1550–1556.
21. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5: 178.
22. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.
23. Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396: 59–70.