



UNIVERSITÀ DEGLI STUDI DI CATANIA  
DIPARTIMENTO DI MATEMATICA E INFORMATICA  
CORSO DI LAUREA MAGISTRALE IN INFORMATICA

---

BioCT

---

RELAZIONE PROGETTO BIOINFORMATICA

---

Giuseppe Parasiliti W82/000152  
Giuseppe Sgroi W82/000131

Prof. Alfredo Ferro

---

ANNO ACCADEMICO 2017/2018

# Indice

<b>1</b>	<b>Introduzione</b>	<b>3</b>
1.1	Cos'è la Bioinformatica . . . . .	3
1.2	Librerie utilizzate . . . . .	3
1.3	Tumori analizzati . . . . .	4
1.3.1	Thymoma (THYM) . . . . .	4
1.3.2	Thyroid Cancer (THCA) . . . . .	5
1.3.3	Liver Hepatocellular Carcinoma(LIHC) . . . . .	6
<b>2</b>	<b>BioCT</b>	<b>7</b>
2.1	Introduzione e Funzionamento . . . . .	7
2.2	Struttura . . . . .	8
2.2.1	Dashboard . . . . .	8
2.2.2	Biomarcatori . . . . .	9
2.2.3	MITHrIL . . . . .	11
<b>3</b>	<b>Guida all'utilizzo</b>	<b>14</b>

# Capitolo 1

## Introduzione

### 1.1 Cos'è la Bioinformatica

La bioinformatica è una disciplina scientifica dedicata alla risoluzione di problemi biologici a livello molecolare con metodi informatici e contribuisce alla descrizione dal punto di vista quantitativo dei fenomeni biologici coinvolgendo, oltre alla biologia e all'informatica, altri campi tra cui matematica applicata, statistica, biochimica ed intelligenza artificiale.

La bioinformatica principalmente 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.

### 1.2 Librerie utilizzate

Nello sviluppo della piattaforma BioCT, abbiamo utilizzato molteplici librerie, tra le più importanti ricordiamo:

- Python
  - **Flask**: è un micro-framework scritto in python che ci permette di creare un Web-Server. È basato sullo strumento Werkzeug WSGI con il motore template Jinja2. Ha licenza BSD. Più informazioni su <http://flask.pocoo.org/>
  - **rpy2**: L'interfaccia di alto livello di rpy2 è progettata per facilitare l'utilizzo degli script R in Python. Gli oggetti R sono esposti come istanze di classi implementate in Python, con funzioni R come metodi associati a quegli oggetti.

- R
  - **LIMMA**: Pacchetto fondamentale per la ricerca di biomarcatori fornito dal linguaggio R, in particolare per la ricerca dei geni differenzialmente espressi partendo da valori di espressione determinati per mezzo di esperimenti NGS.
  - **Biobase**: è parte integrante del progetto Bioconductor ed è utilizzato in molti altri pacchetti. Biobase contiene strutture standardizzate per rappresentare dati genomici.
  - **TCGA-Assembler 2**: è un software open-source che permette automaticamente di scaricare, assemblare e processare i dati TCGA (The Cancer Genome Atlas). Per scaricare i dataset ci serviremo del Modulo A che acquisisce i dati pubblici TCGA dal Genomic Data Commons (GDC) dell' U.S. National Cancer Institute. Maggiori informazioni su ( <https://github.com/compgenome365/TCGA-Assembler-2> )

## 1.3 Tumori analizzati

I tumori presi in considerazione sono i seguenti:

- Thymoma (THYM)
- Thyroid Cancer (THCA)
- Liver Hepatocellular Carcinoma(LIHC)

### 1.3.1 Thymoma (THYM)

Il timoma rappresenta in assoluto il più comune dei tumori che colpiscono il timo. In genere si tratta di un tumore che cresce lentamente e solo raramente si diffonde al di fuori dell'organo di origine. Il sistema di classificazione Masaoka è il più ampiamente usato ed è basato sulla estensione anatomica della malattia al momento dell'intervento. Gli stadi Masaoka si suddividono:

- **I**: Completamente incapsulata
- **IIA**: Invasione Microscopica invasione attraverso la capsula all'interno del tessuto adiposo
- **IIB**: Invasione macroscopica all'interno della capsula
- **III**: Invasione macroscopica negli organi adiacenti
- **IVA**: Impianti pleurali o pericardiali
- **IVB**: Formazione di metastasi linfogene o ematogene in siti distanti (di tipo extra-toracico)

### 1.3.2 Thyroid Cancer (THCA)

Il tumore alla tiroide(TC) è causato dall'anomalo sviluppo di alcune cellule di questa ghiandola, simile ad una farfalla, situata alla base del collo appena sotto il pomo d'Adamo. Esso si manifesta molto spesso in una forma benigna e piuttosto raramente in forme maligne (assumendo in questo caso il nome di cancro alla tiroide). Il cancro della tiroide non è molto comune, poiché costituisce l'1-2% di tutti i tumori, con un'incidenza di 4,1 casi ogni 100.000 abitanti per gli uomini e 12,5 nuovi casi ogni 100.000 abitanti per le donne. Secondo stime del Registro tumori italiano, nel 2012 sono stati diagnosticati 3.200 tumori tiroidei nei maschi e 10.900 nelle femmine.

La sopravvivenza è molto elevata (oltre il 90% a 5 anni dalla diagnosi nelle forme differenziate). Il sistema numerico classifica i tumori secondo quattro stadi:

- stadio 1: il tumore è piccolo e circoscritto;
- stadio 2 o 3: il tumore ha invaso i linfonodi adiacenti;
- stadio 4: il tumore si è diffuso ad altri organi.

Secondo il sistema TNM:

- **T**: indica le dimensioni del tumore e comprende quattro stadi, T1 – T4;
- **N**: indica se nei linfonodi adiacenti alla tiroide sono presenti cellule tumorali. Comprende due stadi:
  - **N0**: assenza di cellule tumorali;
  - **N1**: presenza di cellule tumorali (N1a: nei linfonodi nel comparto centrale, N1b nei linfonodi laterali del collo)
- **M** indica se sono presenti metastasi. Comprende due stadi:
  - **M0**: assenza di metastasi;
  - **M1**: presenza di metastasi.

Associando gli stadi T, N e M è poi possibile definire la stadiazione complessiva della malattia.

Tutti i **tumori anaplastici** sono considerati T4, suddivisi in due stadi:

- **T4a**: tumore di qualsiasi dimensione confinato alla tiroide, asportabile chirurgicamente;
- **T4b**: tumore di qualsiasi dimensione esteso oltre la capsula tiroidea, non asportabile chirurgicamente.

### 1.3.3 Liver Hepatocellular Carcinoma(LIHC)

Il carcinoma epatocellulare (epatocarcinoma o HCC) è il più frequente tumore primitivo del fegato e si riscontra, nella maggior parte dei casi, in pazienti con epatopatia cronica (70-90% dei casi di HCC), presentandosi in una forma multifocale alla diagnosi nel 75% dei casi. Per la stadiazione dell'HCC non esiste un sistema universalmente accettato.

Nel corso degli anni sono stati proposti diversi sistemi di stadiazione, quella più diffusa ha l'acronimo TNM che indicano le 3 sottoforme per cui vengono poi raggruppate "Tumore primitivo, linfonodi (node) e metastasi a distanza", è stata accertata essere migliore negli stadi iniziali del tumore. Nella figura seguente si notano i vari stadi della malattia:

Tumore primitivo	Linfonodi regionali	Metastasi a distanza
TX, non determinato	NX, non determinabili	MX, non determinabili
T0, non si evidenzia	N0, assenza metastasi	M0, assenza a distanza
T1, si mostra un nodulo singolo	N1, presenza di metastasi	M1, presenza delle forme a distanza
T2, noduli multipli (dimensioni inferiori ai 5 cm)	*	
T3, noduli multipli (dimensioni superiori ai 5 cm)	*	
T4, interessamento altri organi o con perforazione peritoneo viscerale	*	

Stadio	T-N-M
I	T1 - N0 - M0
II	T2 - N0 - M0
IIIA	T3 - N0 - M0
IIIB	T4 - N0 - M0
IIIC	T (ogni forma)- N1 - M0
IV	T (ogni forma)- N (ogni forma)- M1

## Capitolo 2

# BioCT

### 2.1 Introduzione e Funzionamento

Obiettivo del nostro progetto è quello di realizzare un portale, da noi intitolato “**BioCT**”, che permetta all’utente finale di effettuare le analisi dei biomarcatori e la successiva interazione con Mithril nel modo più semplice e veloce possibile. Il fulcro su cui ruota l’intero progetto è il portale **National Cancer Institute GDC Data Portal** (<https://portal.gdc.cancer.gov/>), dal quale estrapoliamo un set di dati specificatamente ad alcune tipologie tumorali. Per ciascun tumore estraiamo le seguenti tipologie di dati:

- Biospecimen & Clinical
- miRNA Seq.
- RNA Seq.

Una volta salvati i dati, secondo una struttura di path ben organizzata, entriamo nel vivo dell’analisi.

L’intero progetto viene eseguito tramite la libreria **Python** intitolata “**Flask**” la quale ci permette di mettere su un Web-Server, ed è organizzato:

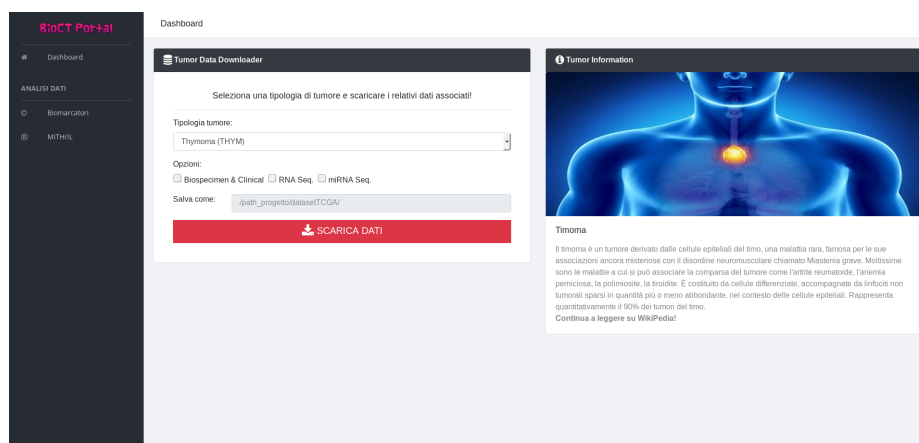
- Lato Client: molteplici pagine HTML e CSS, il cui contenuto è espresso dinamicamente in Javascript che elabora i dati in output provenienti da Python.
- Lato Server: elaborazione dei dati attraverso Python.

## 2.2 Struttura

Il progetto è strutturato in tre sezioni in ognuno delle quali viene effettuata un tipo particolare di analisi dati:

- **Dashboard:** schermata principale che offre la possibilità di selezionare il tumore e scaricare il dataset relativo, permettendo all'utente di scegliere di scaricare i Biospecimen & Clinical, la sequenza RNA e/o la sequenza miRNA.
- **Biomarcatori:** in questa sezione sarà possibile eseguire l'analisi del tumore desiderato, permettendo all'utente di scegliere i vari parametri di input.
- **MITHrIL :** in questa schermata, si potrà avviare l'analisi con il programma MITHrIL selezionando la tipologia dei tumori e i vari contrasti del rispettivo tumore.

### 2.2.1 Dashboard

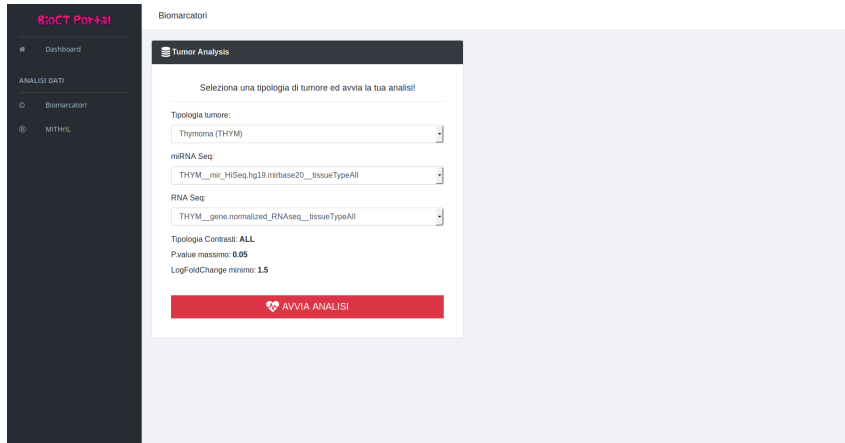


In questa sezione, come già ribadito nel paragrafo precedente, è possibile selezionare il tumore di interesse e scegliere di scaricare il dataset relativo a Biospecimen & Clinical, sequenza RNA e/o sequenza miRNA, infine l'utente potrà scaricare i dataset selezionati cliccando su “SCARICA DATI”.

Alla comparsa del popup, il sistema farà una chiamata POST che eseguirà lo script in python il quale creerà la cartella “datasetTCGA” e la sottocartella relativa al tumore selezionato dall'utente, ed infine, attraverso la libreria rpy2, sopradescritta, eseguirà il codice R per scaricare i dataset attraverso la libreria sopracitata TCGA-Assembler 2. Fatto ciò si potrà procedere al download, durante il quale un popup mostrerà il processo e notificherà l'avvenuto completamento.

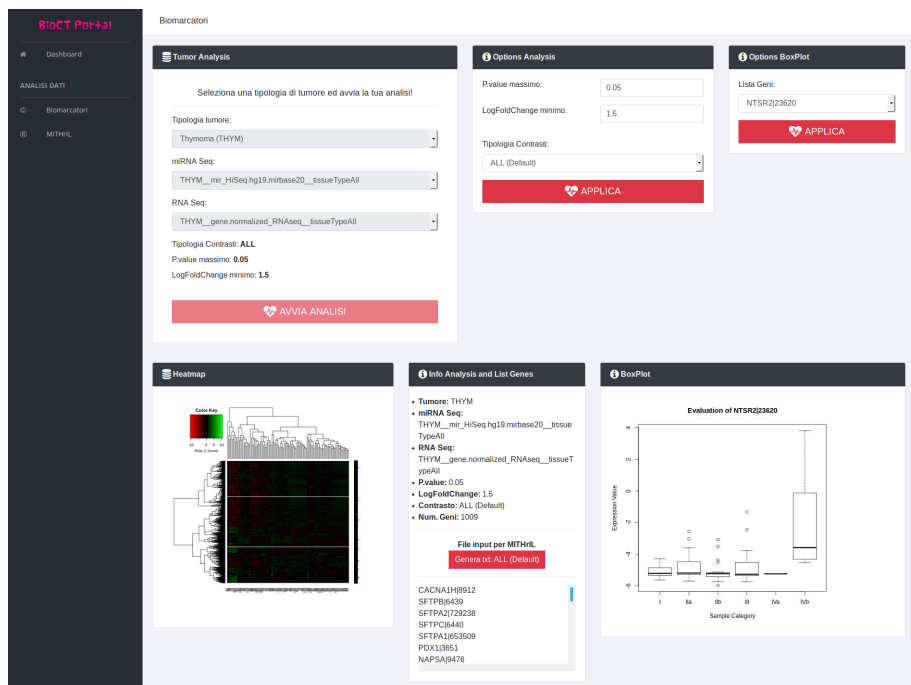


## 2.2.2 Biomarcatori



The screenshot shows the 'Biomarcatori' section of the BioCT Portal. On the left is a dark sidebar with the 'BioCT Portal' logo and navigation links: 'Dashboard', 'ANALISI DATI', 'Biomarcatori', and 'MITHIL'. The main content area is titled 'Biomarcatori' and contains a 'Tumor Analysis' form. The form has a header 'Seleziona una tipologia di tumore ed avvia la tua analisi!'. It includes three dropdown menus: 'Tipologia tumore:' (set to 'Thymoma (THYM)'), 'miRNA Seq:' (set to 'THYM\_miR\_hg19.mirbase20\_\_issueTypeAll'), and 'RNA Seq:' (set to 'THYM\_gene.normalized\_RNAseq\_\_issueTypeAll'). Below these, it shows 'Tipologia Contrasti: ALL', 'Pvalue massimo: 0.05', and 'LogFoldChange minimo: 1.5'. A red button with a heart icon and the text 'AVVIA ANALISI' is at the bottom.

In questa sezione l'utente può avviare l'analisi dei biomarcatori, selezionando la tipologia di tumore, scegliendo la sequenza RNA e la sequenza miRNA. L'analisi verrà avviata dopo che l'utente avrà cliccato sul bottone "AVVIA ANALISI", il quale attraverso una chiamata POST verrà eseguito un script in python che, dopo aver effettuato alcuni controlli sui parametri della chiamata POST, eseguirà lo script R associato al tumore selezionato dall'utente. Ultimata l'analisi iniziale dei Biomarcatori, all'utente verranno mostrati i risultati di quest'ultima così come nella figura sottostante.



In questa fase all'utente sono mostrati i dettagli dell'analisi da lui eseguita. Come mostrato in figura l'utente attraverso il box **“Options Analysis”** posto al centro in alto potrà ripetere l'analisi del tumore e dei dataset relativi scelti in precedenza cambiando il valore del pvalue massimo, del LogFoldChange minimo e inoltre potrà scegliere la tipologia di contrasti<sup>1</sup> su cui concentrarsi.

Attraverso il box **“Options BoxPlot”** in alto a destra potrà selezionare uno dei geni differenzialmente espressi estratti dall'analisi, ed automaticamente verrà generato il boxplot relativo.

Nel box centrale, **“Info Analysis and List Genes”** l'utente avrà informazioni sull'analisi appena effettuata (Tumore selezionato, sequenze selezionate, pvalue, LogFoldChange, la tipologia di contrasto e il numero di geni che sono rientrati nell'analisi). All'interno dello stesso box, all'utente è data la possibilità di generare, relativamente al tumore e al tipo di contrasto selezionato, i file di input per MITHrIL.

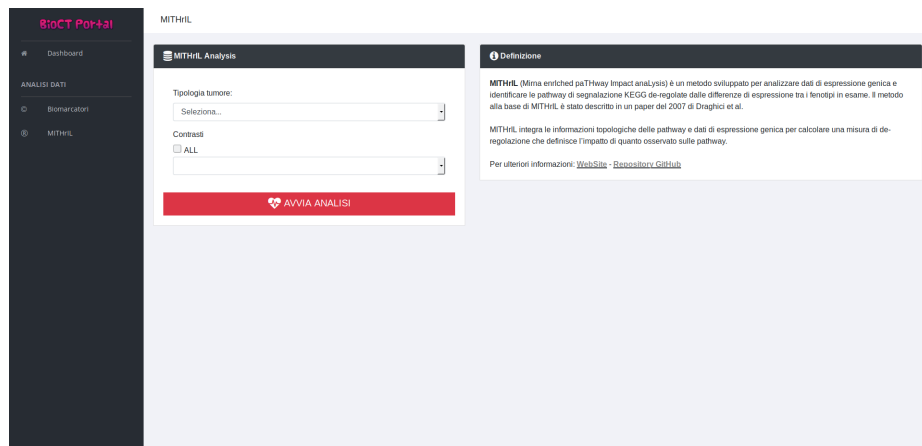
---

<sup>1</sup>Per contrasto si intende i vari stadi tumorali del tumore selezionato messe in relazione, secondo questa regola: Stadio 1 vs Normale, Stadio 2 vs Stadio 1, Stadio 3 vs Stadio 2, Stadio 4 vs Stadio 3 etc.

### 2.2.3 MITHrIL

MITHrIL è un metodo sviluppato per analizzare dati di espressione genica e identificare le pathway di segnalazione KEGG de-regolate dalle differenze di espressione tra i fenotipi in esame. Il metodo alla base di MITHrIL è stato descritto in un paper del 2007 di Draghici et al.

MITHrIL integra le informazioni topologiche delle pathway e dati di espressione genica per calcolare una misura di de-regolazione che definisce l'impatto di quanto osservato sulle pathway.



Generati i file di input per MITHrIL, l'utente potrà avviare l'analisi selezionando la tipologia di tumore e il contrasto che si vuole analizzare con MITHrIL. Il sistema andrà alla ricerca del file di input selezionato dall'utente e attraverso la seguente istruzione in python :

```
sp.call(['java', '-jar', 'MITHrIL2.jar', 'merged-mithril',
        '-verbose', '-i', input.txt, '-o', output.txt, '-p',
        perturbations.tx])}
```

verrà avviato MITHrIL.

Se la checkbox “ALL” viene selezionata dall’utente, esso dovrà selezionare anche il tipo di contrasto, in modo tale che automaticamente la piattaforma sceglierà la colonna relativa al contrasto selezionato dall’utente dal file contenente tutti i contrasti del relativo tumore.

Nel momento in cui l’utente cliccherà sul bottone “Avvia Analisi”, verrà effettuata una chiamata POST eseguendo il codice python che permetterà di eseguire MITHrIL.

Terminata l’analisi MITHrIL genererà due file:

1. **out\_mithril\_TumorName\_0\_TypeContrast.txt**
2. **out\_mithril\_pertubations\_TumorName\_0\_TypeContrast.txt**

dove:

- TumorName: è il nome del tumore selezionato dall’utente
- 0: indica se l’utente ha selezionato la checkbox “ALL”
- TypeContrast: indica la tipologia di contrasto selezionata dall’utente.

Il primo file di output conterrà tutte le statistiche delle pathway computate da MITHrIL, il secondo file è l’output delle perturbazioni, che conterrà tutte le statistiche dei nodi computati da MITHrIL.

L’utente potrà visualizzare in una tabella l’output contenente tutte le statistiche delle pathway che corrispondono al primo file output di MITHrIL (così come mostrato nella prima figura sottostante). Mentre nella seconda figura potrà visualizzare in una tabella l’output contenente tutte le statistiche delle pathway che corrispondono al file delle output delle perturbazioni.

Output Mithril								
Show: 10		Search:						
Pathway Id	Pathway Name	Raw Accumulator	Impact Factor	Probability Pi	Total Perturbation	Corrected Accumulator	pValue	Adjusted pValue
path:hsa00190	Oxidative phosphorylation - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa00072	Synthesis and degradation of ketone bodies - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa01040	Biosynthesis of unsaturated fatty acids - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa04672	Intestinal immune network for IgA production - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa03460	Fanconi anemia pathway - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa04310	Wnt signaling pathway - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa00071	Fatty acid degradation - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa04670	Leukocyte transendothelial migration - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa04550	Signaling pathways regulating pluripotency of stem cells - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
path:hsa00630	Glyoxylate and dicarboxylate metabolism - Enriched	0.0	0.0	1.0	0.0	0.0	1.0	1.0
Showing 1 to 10 of 279 records								
Pages: Previous 1 2 3 ... 28 Next								

Output Perturbation Mithril						
Show: 10 ▾		Search: <input type="text"/>				
Pathway Id	Pathway Name	Gene Id	Gene Name	Perturbation	Accumulator	pValue
path:hsa00190	Oxidative phosphorylation - Enriched	64077	LHPP, HDHD2B	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	5464	PPA1, HEL-S-66p, IOPPP, PP, PP1, SID6-8061	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	hsa-miR-101-3p	hsa-miR-101-3p	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	245972	ATP6V0D2, ATP6D2, VMA6	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	245973	ATP6V1C2, ATP6C2, VMA5	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	479	ATP12A, ATP1AL1	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	513	ATP5D	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	514	ATP5E, ATP5E, MC5DN3	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	515	ATP5F1, PIG47	0.0	0.0	1.0
path:hsa00190	Oxidative phosphorylation - Enriched	516	ATP5G1, ATP5A, ATP5G	0.0	0.0	1.0
Showing 1 to 10 of 42188 records				Pages: Previous <b>1</b> 2 3 ... 4219 Next		

## Capitolo 3

# Guida all'utilizzo

1. Importare il progetto (sviluppato su PyCharm)
2. Installare le opportune librerie richieste da R e Python
3. Aprire il file “main.py” ed avviarlo con PyCharm o IDE alternativo, oppure da un terminale eseguire il seguente comando:

```
python3.6 main.py
```

- (a) Il web server sarà completamente avviato quando apparirà a video la seguente schermata:

```
/usr/bin/python3.6 /media/giuseppe/DATA/WorkspacePycharm/BioInformatica/main.py
* Serving Flask app "main" (lazy loading)
* Environment: production
  WARNING: Do not use the development server in a production environment.
  Use a production WSGI server instead.
* Debug mode: on
* Running on http://127.0.0.1:2892/ (Press CTRL+C to quit)
* Restarting with stat
* Debugger is active!
* Debugger PIN: 292-334-178
```

- (b) Fare click sull'url localhost, in blu nell'immagine sovrastante, oppure inserirlo manualmente nel proprio browser. Si aprirà automaticamente il browser con la Dashboard iniziale della piattaforma BioCT.