Università di Modena e Reggio Emilia
Facoltà di Scienze Fisiche, Informatiche e Matematiche

Corso di Laurea in
Informatica

Progettazione, realizzazione ed accessibilità di un database biomolecolare sulle *sequenze ultraconservate* del genoma umano

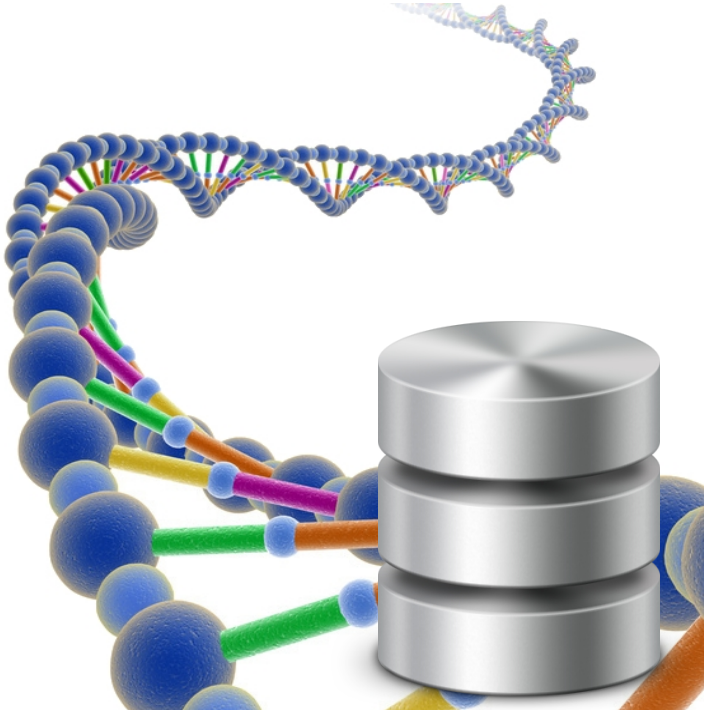
---

Tesi di Laurea di
Vincenzo Lomonaco

Relatore
Prof. Riccardo Martoglia

Correlatori
Prof. Federica Mandreoli
Dott. Cristian Taccioli

- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**



- **Progettazione e realizzazione** di un **database biotecnologico** sulle *sequenze nucleotidiche ultraconservate* e sui meta-dati relativi ad esse.
- **Progettazione e realizzazione** di un **portale web** che ne funga da interfaccia e che soddisfi le moderne esigenze di interrogazione efficente ed efficace dei dati.

Per..

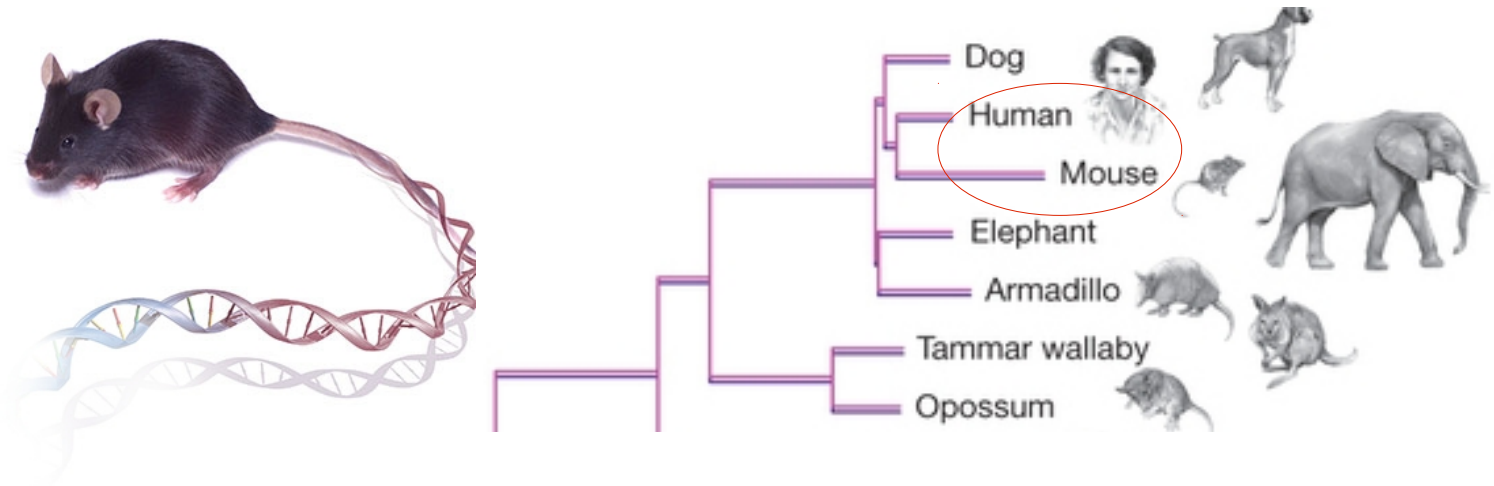
- **Soddisfare le moderne esigenze di interoperabilità** delle banche dati biotecnologiche e **fornire un centro nevralgico di studio** sulle *sequenze nucleotidiche ultraconservate* indirizzato a ricercatori, medici e biologi.
- **Rispondere alle domande chiave sul ruolo ancora incerto e sull'importanza** delle *sequenze ultraconservate*

Cosa sono le sequenze ultraconservate?

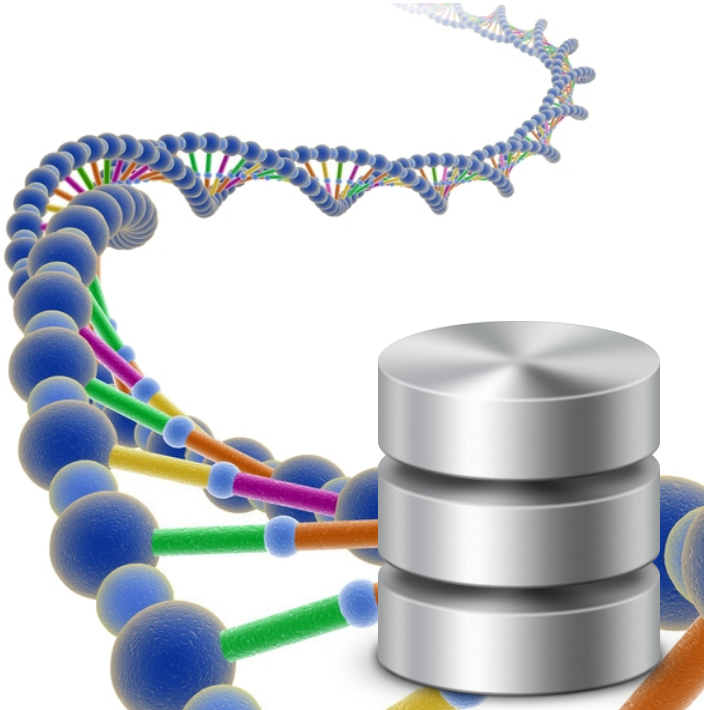
"An ultra-conserved element (UCE) is a region of DNA that is identical in at least two different species.

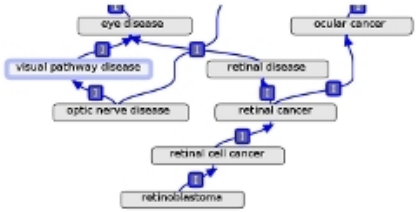
Perchè questo interesse nei confronti delle sequenze UC?

"A small number of those which are transcribed have been connected with human carcinomas and leukemias. A study comparing ultra-conserved elements between humans and Takifugu rubripes proposed an importance in vertebrate development. Several ultra-conserved elements are located near transcriptional regulators or developmental genes. Other functions include enhancing and splicing regulation."



- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**

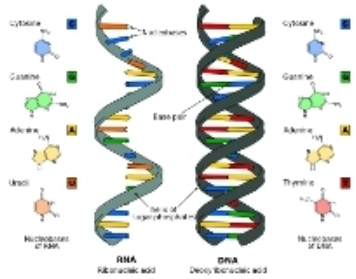




Human disease Hontology

bio mart

Web Service

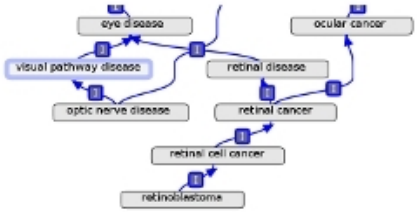


Dati Sequenze UC

UC Server

Database applicazione

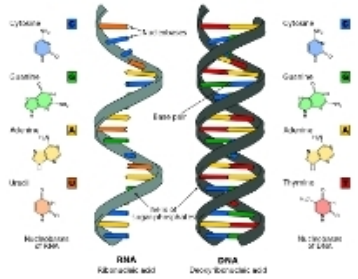
User Browser



Human disease Hontology

Web Service

bio mart

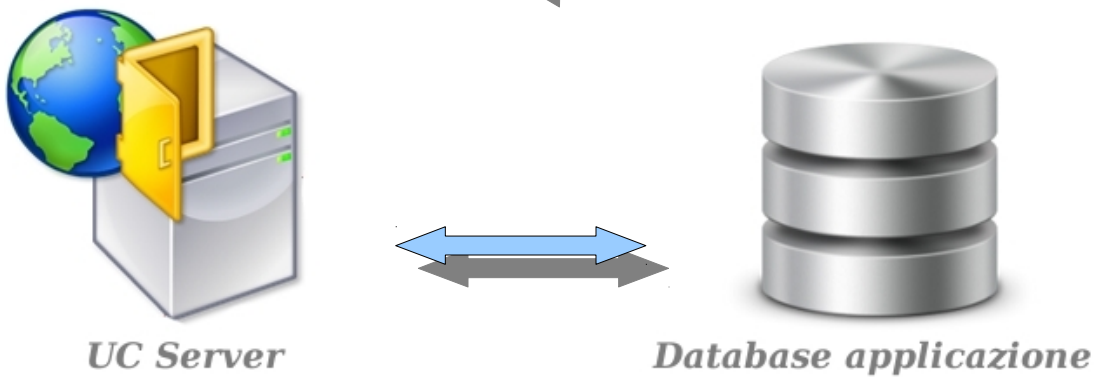


Dati Sequenze UC

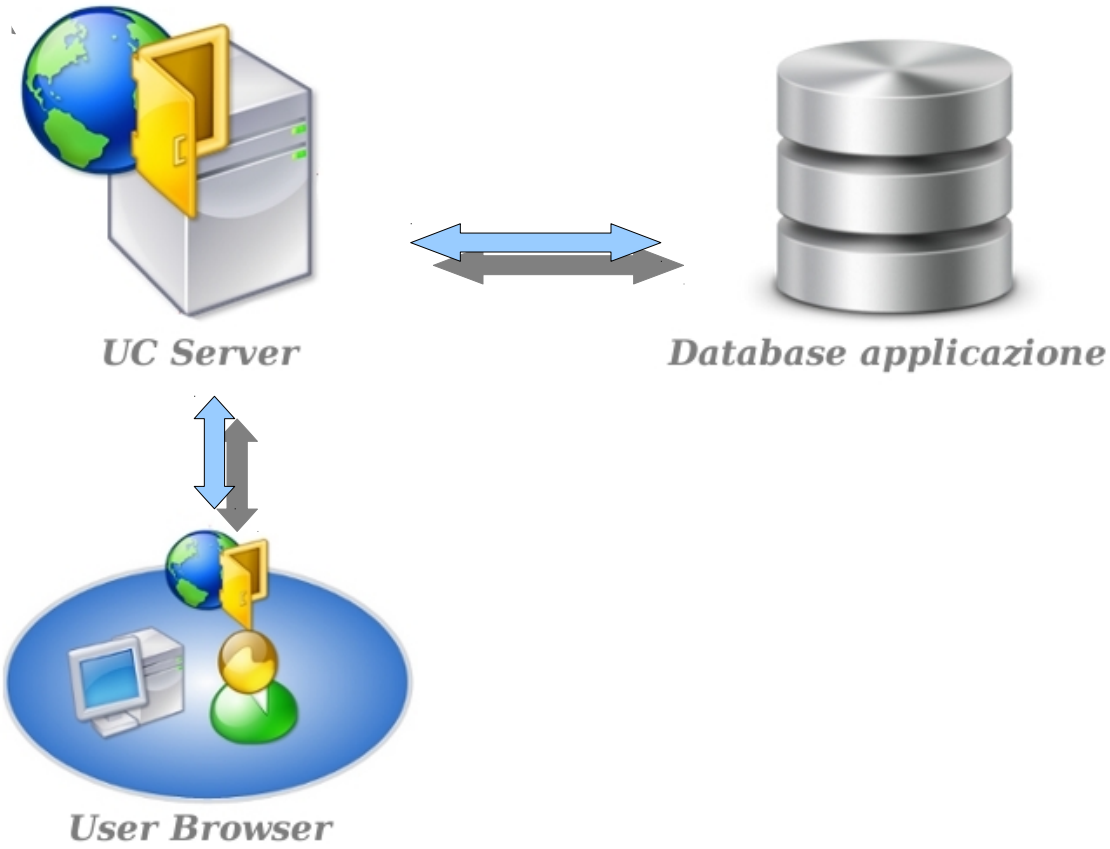
UC Server

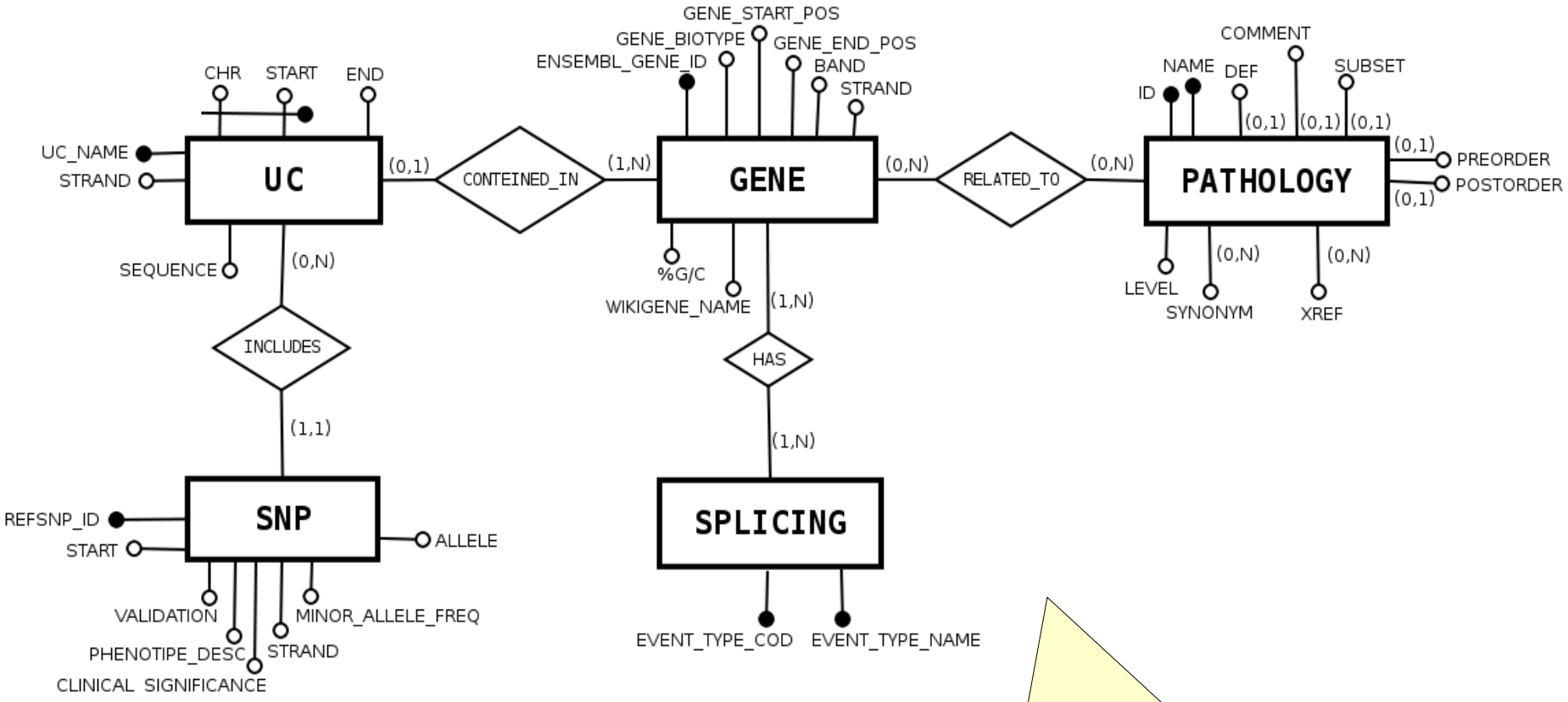
Recupero dei dati

Realizzazione e riempimento del database

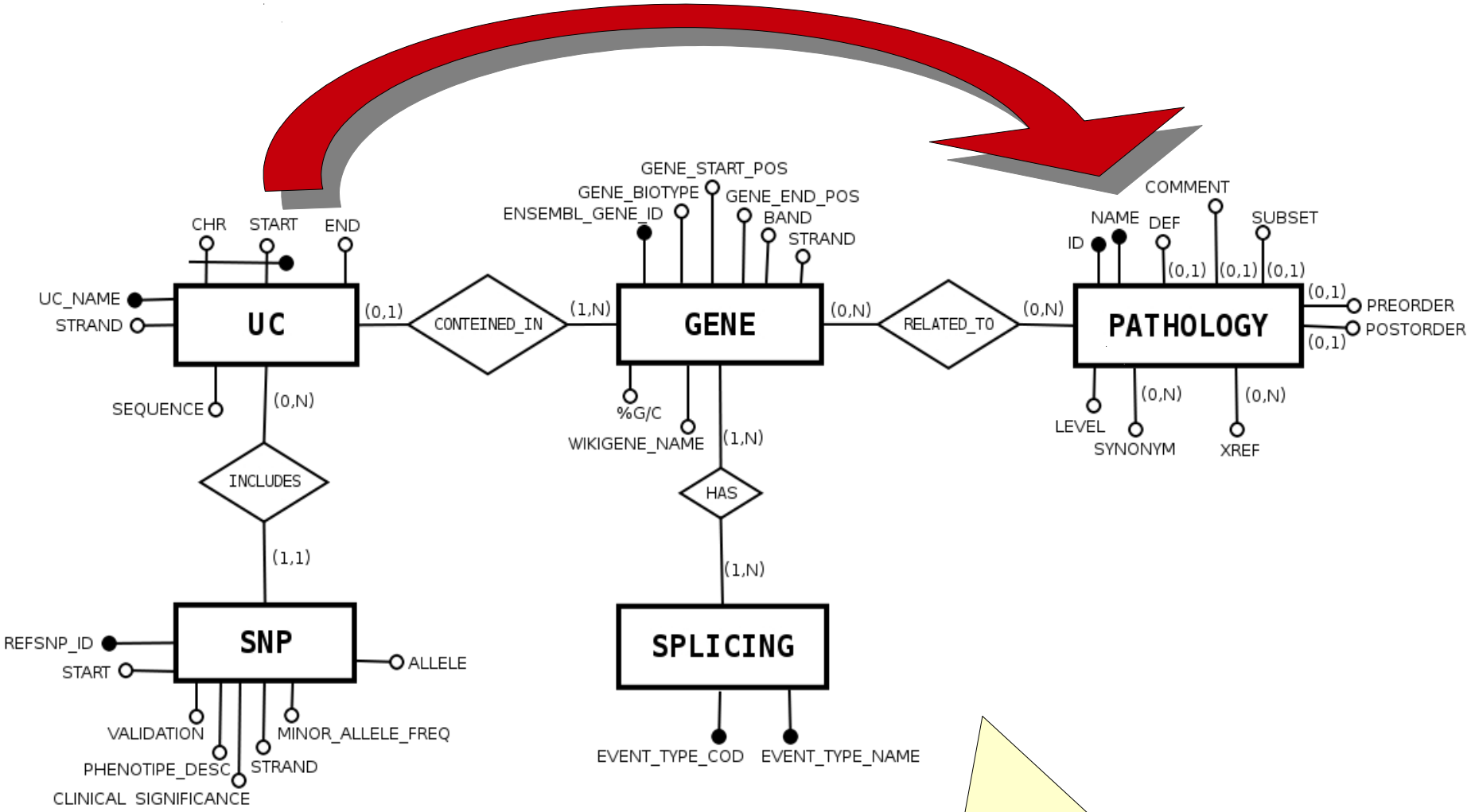


Interrogazione attraverso l'interfaccia web

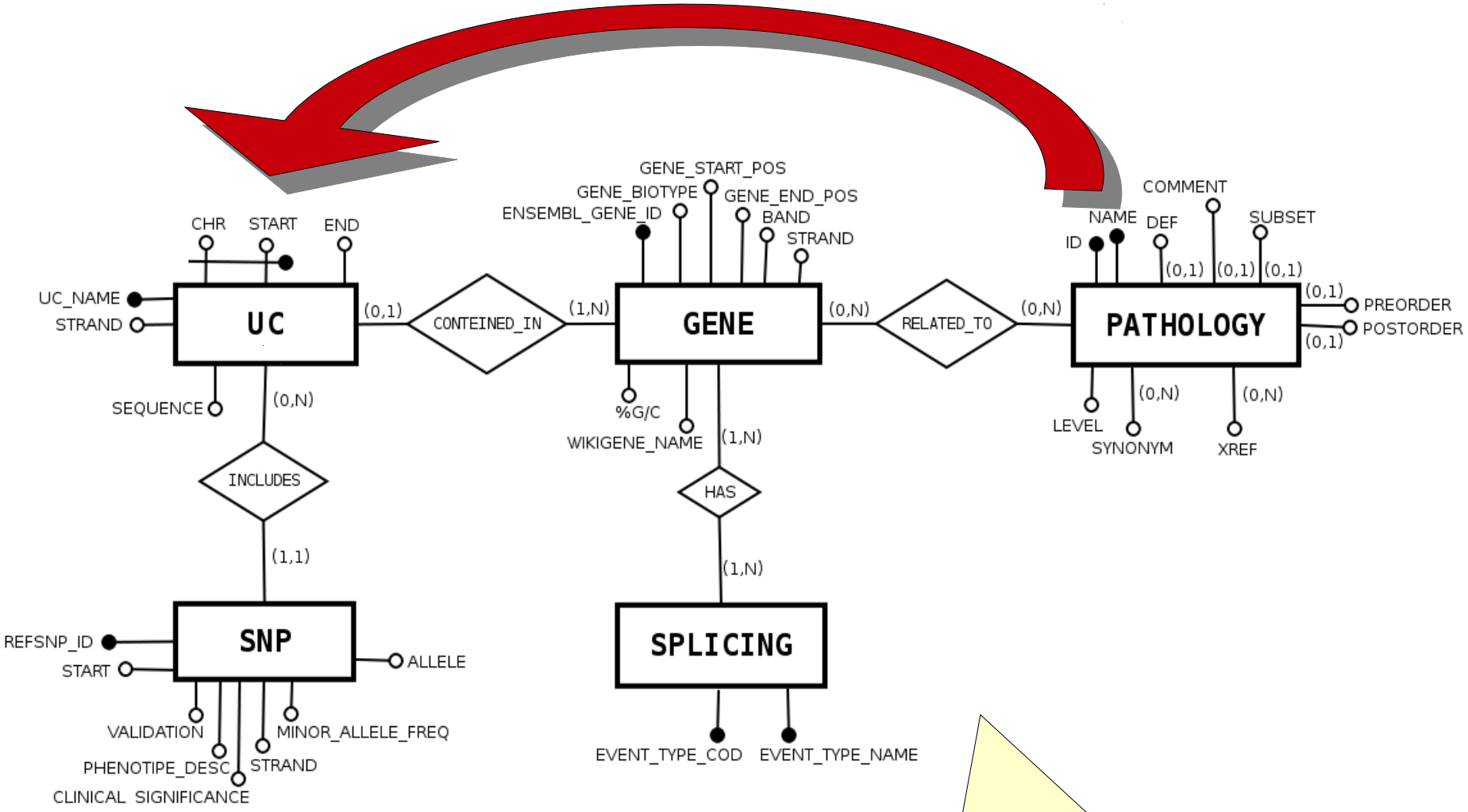




Progetto del database

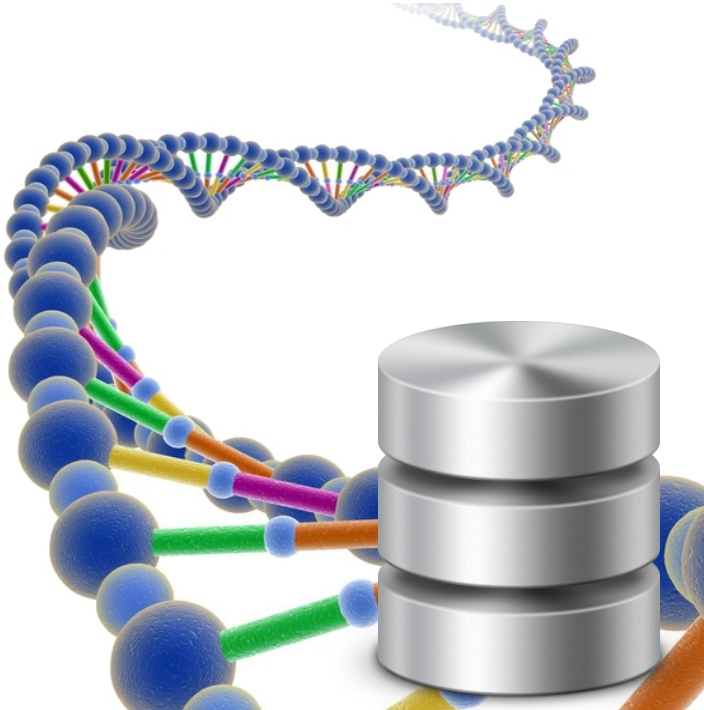


Progetto del database



Progetto del database

- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**



Nuove modalità d'indagine e possibilità esplorative

- Per gli utenti meno esperti una serie di **query prestrutturate.**
- Altrimenti un'accesso completo al database mediante la possibilità di somministrare **query libere** al sistema.

Query prestrutturate offerte. Dalle più semplici...

- Possibilità di conoscere **tutte le informazioni** relative una **sequenza uc** specificata.
- Possibilità di conoscere **tutte le informazioni** relative un **gene** specificato.
- Possibilità di conoscere **tutte sequenze uc rilevate su di uno specifico cromosoma** e comprese tra diverse bp specificate.

...Alle più complesse

- Possibilità di conoscere **tutte le sequenze uc correlate ad una patologia o a tutti i suoi sottotipi** (Si notino le potenzialità introdotte dall'inserimento delle informazioni gerarchiche dell'ontologia).

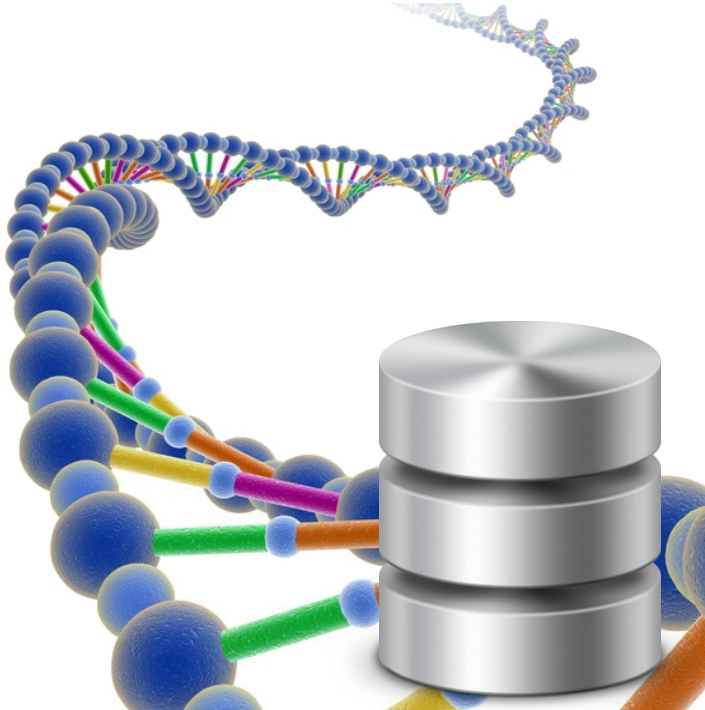
- Possibilità di confrontare secondo **matching approssimato** una *sequenza nucleotidica* con quelle delle uc presenti nel database ed ottenere una lista di possibili risultati elencati per scores.

- Possibilità di confrontare secondo **matching approssimato** una *sequenza nucleotidica* con quelle delle uc presenti nel database ed ottenere una lista di possibili risultati elencati per *scores* e filtrati secondo una patologia (ossia che siano necessariamente correlati a quella specifica patologia o ad un suo sottotipo).

Diversi layout di visualizzazione

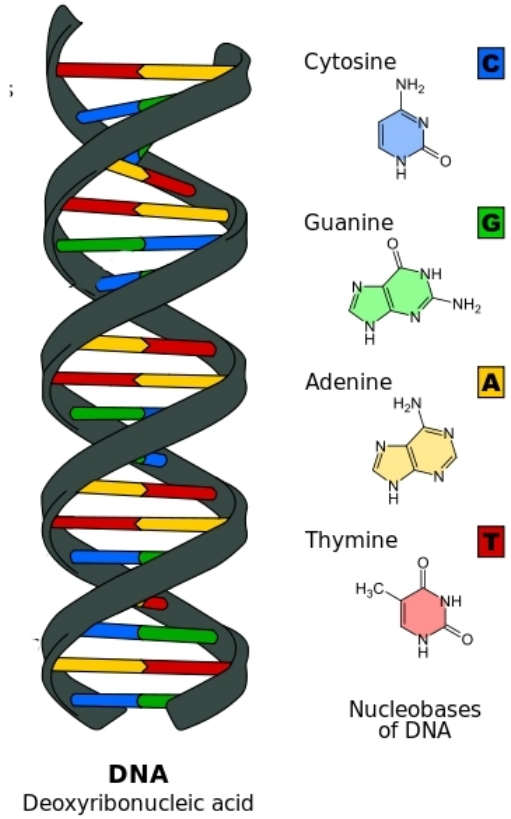
- Classico **formato tabellare** per rispondere ad una query generica come quelle provenienti dal form delle query libere oppure da alcune di quelle prestrutturate.
- Formato un po' più da **“schedario”** per le informazioni circa i campi di un singolo record.
- **Formato adatto alle query prestrutturate con matching approssimato** per la visualizzazione del ranking nell'ordinamento.

- Obiettivi della tesi
- Progettazione e logica dell'applicazione
- Studio e progettazione dell'accessibilità
- **Recupero dei dati**
- Implementazione
- Rapida demo di utilizzo
- Conclusioni e sviluppi futuri



I dati specifici sulle sequenze ultraconservate

I dati specifici sulle sequenze nucleotidiche ultraconservate consistenti nelle vere e proprie sequenze di nucleotidi e sulle coordinate all'interno di un cromosoma specifico e su di un determinato filamento.
Forniti dal refente **Dott. Cristian Taccioli** del Lab. Di ricerca degli Istituti Biologici coordinato dal **Prof. Bicciato**.



>uc.1
TCCACCGACAATGACCAGTTAGTCC...
>uc.2
GCCCGCCCCCCTCCCCGGGCCCAAT...
>uc.3
TTTTTTTTTATTAGCATTTGTGGAAT...

"**Id**","**Chromosome**","**Start**","**End**","**Strand**"

"**uc.1**","1",10597697,10597903,1

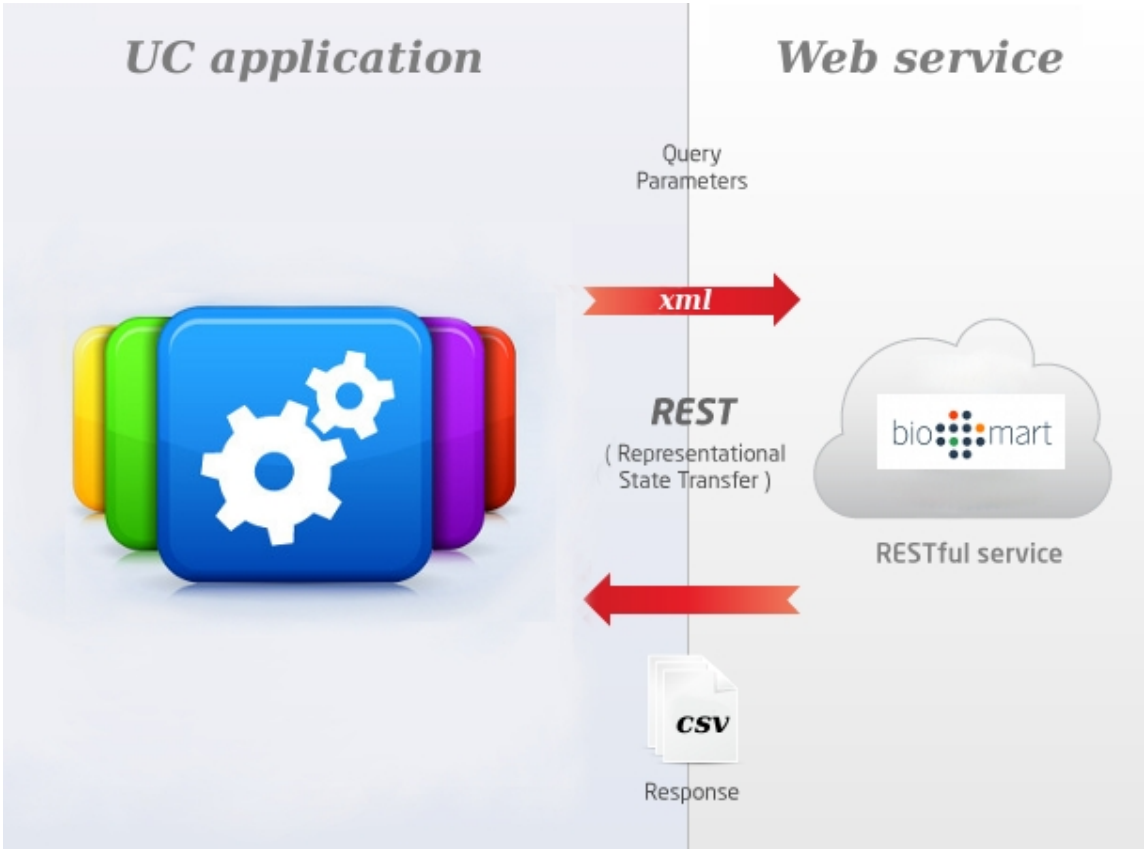
"**uc.2**","1",10732543,10732749,1

"**uc.3**","1",10751165,10751389,1

"**uc.4**","1",10758249,10758607,1

MartService: il web Service di BioMart

“The BioMart project provides free software and data services to the international scientific community in order to foster collaboration and facilitate thescientific discovery process.” “The BioMart Central Portal has introduced an innovative alternative to the large data stores maintained by specialized organizations such as The European Bioinformatics Institute (EBI) or The National Center for Biotechnology Information (NCBI)”

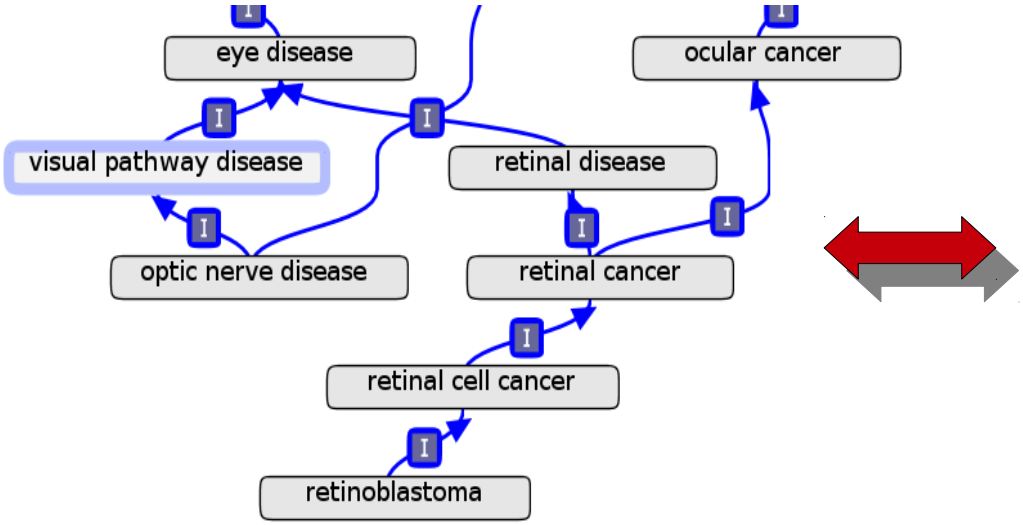


```
String myxml = "<Query  virtualSchemaName = \"default\"
formatter = \"CSV\" header = \"1\" "
+"uniqueRows = \"0\" count = \"\" datasetConfigVersion = \"0.6\" >"
+"<Dataset name = \"hsapiens_gene_ensembl\" interface = \"default\" >"
+"<Filter name = \"chromosome_name\" value= \""
+chr[lineNumber]+"\"/>"
+"<Filter name = \"start\" value = \""+bp_start[lineNumber]+"\"/>"
+"<Filter name = \"end\"value = \""+bp_end[lineNumber]+"\"/>"
+"<Attribute name = \"wikigene_name\" />"
+"<Attribute name = \"ensembl_gene_id\" />"
....
+"</Dataset>"
+"</Query>";

String encoded = URLEncoder.encode(myxml, "utf-8");
URL url = new URL("http://www.biomart.org/biomart/martservice?
query="+encoded);
InputStream response = url.openStream();
BufferedReader reader = new BufferedReader(new
InputStreamReader(response));
```

Obo-foundry e la “Human Disease Onthology”

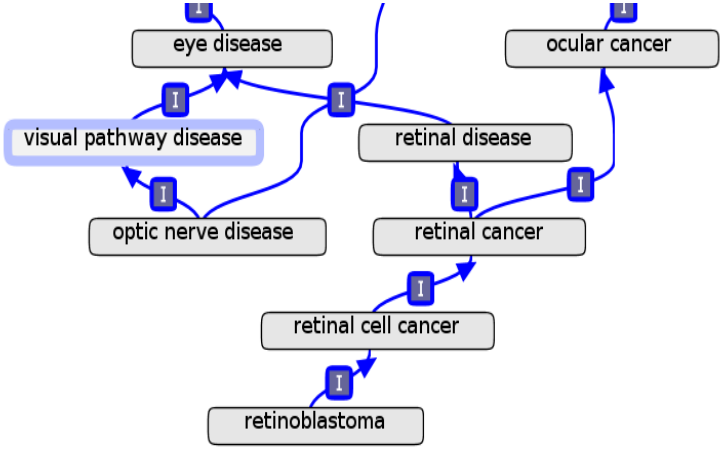
“The **OBO Foundry is a collaborative experiment** involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The mission the **Disease Ontology** (DO) is to provide an **open source ontology** for the integration of biomedical data that is associated with human disease.”



Obo file format

[Term]
id: DOID:162
name: cancer
def: "A disease of cellular proliferation that is malignant and primary, characterized by uncontrolled cellular proliferation, local cell invasion and metastasis."
synonym: "malignant neoplasm" EXACT []
synonym: "malignant tumor " EXACT []
synonym: "primary cancer" EXACT []
xref: ICD9CM:239.4
xref: SNOMEDCT_2010_1_31:189535002
xref: UMLS_CUI:C0027639
is_a: DOID:14566 ! disease of cellular proliferation

Estrazione dei termini ed inserimento nel database



Estrazione dei termini di rilievo

Trasformazione in Albero

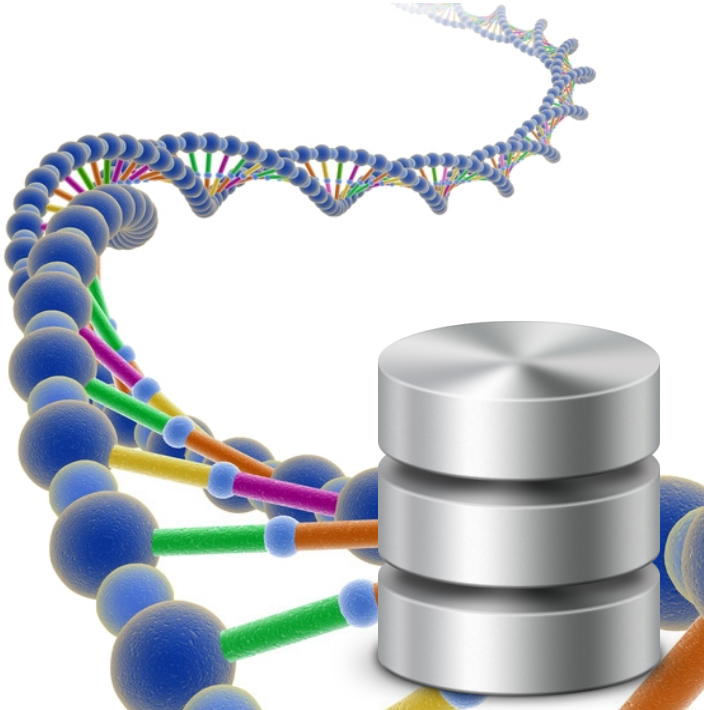
Calcolo attributi preordine e postordine

Sfruttare attributi di preordine e postordine

"Given a node v [..] with pre(v) and post(v) ranks, the following properties are important towards our objectives:
all nodes x with pre(x) < pre(v) are the ancestors or preceding nodes of v;
all nodes x with pre(x) > pre(v) are the descendants or following nodes of v;
all nodes x with post(x) < post(v) are the descendants or preceding nodes of v;
all nodes x with post(x) > post(v) are the ancestors or following nodes of v;
for any v [..], we have pre(v) - post(v) + size(v) = level(v);
if pre(v) = 1, v is the root, if pre(v) = n, v is a leaf. [..] "

```
SELECT UC_NAME,NAME,MIN(LEVEL)AS LEVEL FROM UC INNER
JOIN (SELECT ENSAMBL_GENE_ID,NAME,LEVEL FROM RELATED_TO
INNER JOIN (SELECT ID,NAME,LEVEL FROM PATHOLOGY WHERE
PREORDER >= (SELECT PREORDER FROM PATHOLOGY WHERE
NAME='disease') AND POSTORDER <=(SELECT POSTORDER FROM
PATHOLOGY WHERE NAME='disease') ORDER BY LEVEL)AS B WHERE
RELATED_TO.ID = B.ID)AS C WHERE
UC.ENSAMBL_GENE_ID = C.ENSAMBL_GENE_ID group by
uc_name, name ORDER BY LEVEL
```

- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**



Implementazione in java dell'ambito gestionale

Il progetto è stato interamente sviluppato con **java in ambiente Eclipse** come soluzione efficace e portabile per:

- Il recupero **automatico, la gestione ed il filtraggio** dei dati dal Web Service di BioMart e dei dati sul disco dell *sequenze uc.*
- La **gestione dell'albero delle patologie**, dei controlli, delle numerazione e visite in preordine e postordine, dell estrazioni dei termini.
- Della creazione degli **Script SQL** validi per la creazione del database e l'inserimento dei dati nello stesso

Implementazione dell'interfaccia web

Per l'interfaccia web sono stati utilizzati i linguaggi più in voga **PHP, CSS, HTML** con il database **MySQL**, per la **gestione dinamica dei contenuti**, la fruizione di **suggerminti durante l'immissione** ed il **collegamento efficente al database**.

Numeri

File html, php, java: **30+**
Linee di codice per file in media: **~100**

Implementazione delle Query

Matching esatto

Per quanto riguarda le query prestrutturate con **marching esatto** non si è utilizzato niente di più del linguaggio **SQL** per implementare delle query parametriche I cui parametri sono specificati mediante delle text box

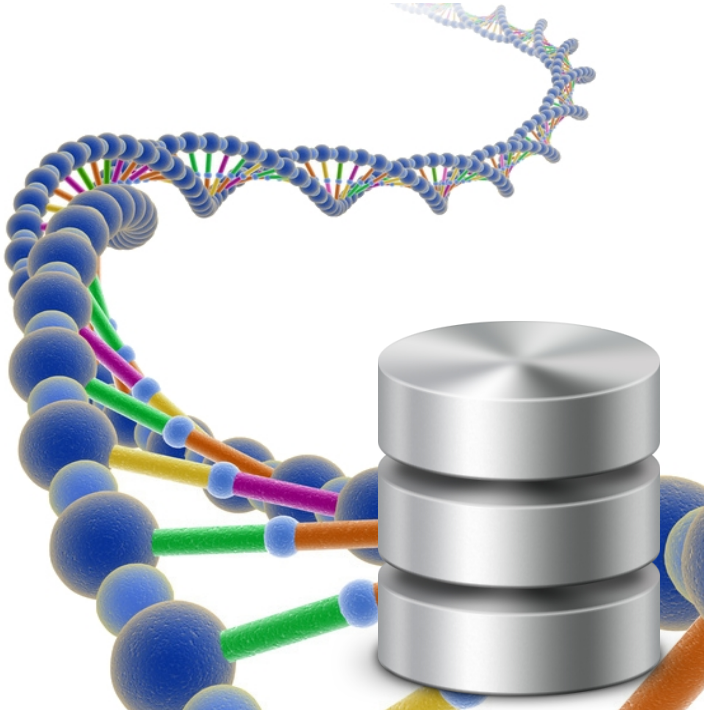
Matching approssimato

Per il **matching approssimato**, invece, non si è creato un algoritmo di confronto e di ranking a sè stante o in SQL ma si è usato il collaudato e rinomato **BLAST** (*Basic Local Alignment Search Tool, ovvero strumento di ricerca di allineamento locale*), un software scritto in **C++** , patrocinato dell'**NCBI,** usato per comparare le informazioni contenute nelle strutture biologiche primarie. Nello specifico la sua versione specializzata per il confronto di sequenze nucleotidiche Blastn

Esempio di utilizzo:

```
./blastn -db UCdb -query query.fasta
 -task blastn-short -out results.out
```

- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**



- **Obiettivi della tesi**
- **Progettazione e logica dell'applicazione**
- **Studio e progettazione dell'accessibilità**
- **Recupero dei dati**
- **Implementazione**
- **Rapida demo di utilizzo**
- **Conclusioni e sviluppi futuri**

Conclusioni

È stato prodotto uno **strumento utile alla comunità scientifica**, **liberamente accessibile a tutti mediante il web** e disponibile all'attività di ricerca di medici e biologi interessati a apprendere o approfondire, apliare il tema delle *sequenze nucleotidiche ultraconservate* **per rispondere alle domande chiave che ancora ruotano attorno al loro significato ed alla loro importanza**.

Sviluppi futuri

- **Ampliamento strutturale** del database per le classi di sequenze utraconservate del ratto e del topo
- **Introduzione di nuove modalità di interrogazione e visualizzazione dei dati**
- **Ottimizzazione e miglioria degli script** per l'**aggiornamento automatico** dei dati scaricati dal Web Service

GRAZIE PER L'ATTENZIONE