UNIVERSITA' DEGLI STUDI DI NAPOLI

"Federico II"

Facoltà di Medicina Veterinaria

DOTTORATO DI RICERCA IN

Produzione e Sanità degli alimenti di origine animale

Indirizzo: Scienze dell'allevamento animale

- CICLO XX -

TITOLO TESI

Un approccio informatico per lo studio della specie bufalina: dal genoma all'organizzazione dei dati.

Coordinatore

Ch.ma Prof.ssa M.L. Cortesi

Tutor :

Ch.mo Prof. Giuseppe Campanile

Dottorando

Dott.Guido Fusco

Novembre 2007

Indice

1. Introduzione

L'allevamento del bufalo ha radici antiche nel territorio e nella cultura campana.

In Campania si alleva circa il 75% del patrimonio bufalino nazionale con un coinvolgimento, tenendo conto dell'intero indotto, di oltre 15000 operatori. La Campania, così come il basso Lazio e la Capitanata di Foggia che pure ospitano allevamenti bufalini, è stata caratterizzata fino al secolo scorso dalla presenza di paludi che rendevano impensabile qualunque altro tipo di allevamento e/o attività agricola. L'allevamento del bufalo in Italia ha rappresentato quindi, la prima forma di utilizzazione di questi territori marginali a scopi economici. La diffusione e la successiva permanenza del bufalo in questi terreni paludosi, caratterizzati da produzioni foraggiere grossolane e in cui predominava la malaria sono state possibili per la elevata capacità di adattamento dimostrata da questo animale. Infatti il bufalo, presenta una spiccata resistenza agli agenti patogeni endemici, ecto ed endoparassiti, e in condizioni di carenza foraggiera mostra una spiccata capacità di migliorare l'efficienza di utilizzazione degli alimenti. Queste caratteristiche, grazie anche allo sviluppo di una serie di tecniche di allevamento, consentono il mantenimento di livelli costanti di produzione di latte destinato alla produzione di mozzarella. E' ormai automatica l'associazione tra la bufala campana e la mozzarella. Il 91% della produzione di mozzarella DOP è campano.

Il potenziale produttivo ed economico dell'allevamento bufalino però non si esaurisce nella sola produzione di latte, sebbene molto rinomata e redditizia. L'allevamento bufalino può avere un ulteriore importante indotto rappresentato dalla produzione di carne. La maggiore valorizzazione della carne bufalina non solo può aprire a nuovi mercati con notevoli incrementi di fatturato ma può contribuire all'eliminazione di molte diseconomie aziendali.

Oltre all'aspetto economico, sono note molte caratteristiche biochimiche della carne di bufalo che la rendono particolarmente interessante anche sul piano nutrizionale. Il ridotto contenuto in colesterolo e trigliceridi, l'alto contenuto in proteine, il contenuto in ferro, l'apporto calorico sono solo alcuni dei parametri che rendono questa carne adatta ad un consumo generalizzato ma anche di categorie con particolari esigenze nutrizionali quali anziani ed atleti. Dal punto di vista organolettico la carne di bufala ha una sua intrinseca gradevolezza che nel tempo, è stata messa in discussione forse a causa di un cattivo consumo. Per lo più, infatti la carne destinata al consumo proveniva da animali male alimentati, cresciuti allo stato semi selvatico, anziani o malati per cui il sapore ne risultava compromesso. In quest'ottica, un ulteriore miglioramento e valorizzazione della carne potrebbe basarsi sull'acquisizione di strumenti scientifici e tecnologici che consentano metodi di allevamento più mirati allo scopo; ovvero metodi che, mediante l'acquisizione di tecniche specifiche di analisi e monitoraggio, conducano alla ulteriore esaltazione degli aspetti nutrizionali ed organolettici della carne.

Così come sta accadendo per altri aspetti delle scienze naturali e mediche da un lato e della gestione aziendale dall'altro, un notevole contributo di innovazione e miglioramento di efficienza potrebbe derivare dall'applicazione di nuove tecnologie in approcci di tipo multidisciplinare. Tali approcci consentono infatti di trarre insegnamento dalle esperienze maturate in campi diversi e di rendere nuove acquisizioni efficacemente fruibili in situazioni apparentemente distanti. L'informatica con le sue diverse specializzazioni ed applicazioni è sicuramente una delle "nuove" scienze che hanno rivoluzionato alla base il modo di affrontare problematiche scientifiche ed economiche. Gli strumenti informatici consentono di trarre il massimo contenuto informativo dai dati provenienti dalle diverse fonti scientifiche ed applicative. In questa tesi si tratterà in particolare di due utili applicazioni all'allevamento bufalino.

Da un lato l'uso degli strumenti di bioinformatica che sono alla base della genomica comparata per la individuazione di marcatori molecolari utili al monitoraggio di vari aspetti biologici e produttivi in risposta a particolari approcci nutrizionali, di allevamento etc. Dall'altro la produzione di un "database" dedicato che consenta di catalogare ed organizzare le più disparate informazioni sul bufalo, comprese quelle prodotte con l'approccio genomico ora menzionato. Un tale strumento che consenta di ottenere e collegare velocemente informazioni di vario genere (molecolari, fisiologiche, epidemiologiche, commerciali etc.) potrebbe costituire uno strumento versatile per operatori in campi diversi.

Allo scopo di chiarire le basi degli approcci scelti, qui di seguito sono riportati alcuni nozioni relative allo sviluppo della bioinformatica, più in particolare della biologia computazionale, e delle basi di dati.

Preliminarmente vediamo una brevissima trattazione degli aspetti biologici dello sviluppo muscolare.

Le caratteristiche strutturali ed organolettiche della carne di bufalo sono il frutto di una stretta interazione tra la componente genetica (il genoma bufalino e la sua complessa regolazione) e l'ambiente.

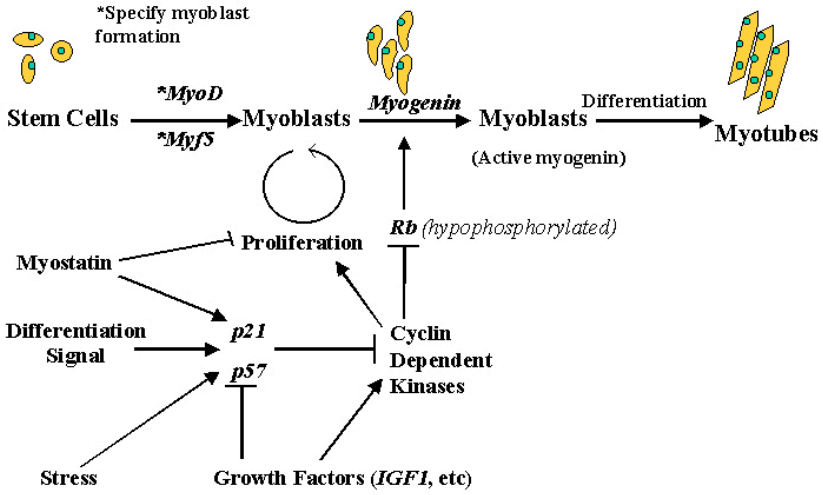
La parola ambiente sintetizza l'insieme degli stimoli esterni provenienti a diverse livelli e che comprendono sia l'attività intra-cellulare che le particolari condizioni nutrizionali e di allevamento. Considerata la stretta interazione che esiste tra il genoma e l'ambiente nel quale esso si esprime, una profonda conoscenza di entrambe le componenti consente di sviluppare strumenti mirati ed efficaci per la conservazione del fenotipo o per il suo miglioramento genetico.

Per la comprensione a livello molecolare dei fattori che condizionano la produzione di carne bovina, e bufalina in particolare, un approccio utile è partire dall'analisi dei geni che determinano la formazione del tessuto muscolare ed il suo sviluppo. Storicamente lo sviluppo muscolare è stato uno dei primi meccanismi dello sviluppo dei mammiferi per cui è stato possibile individuare la sequenza degli eventi molecolari che portano alla determinazione del tessuto muscolare.

L'individuazione e caratterizzazione del gene MyoD1 **[1]**, considerato il "master gene" dello sviluppo muscolare, ha aperto strade

importanti per la comprensione delle fasi della determinazione cellulare verso tipi cellulari specifici. MyoD1 è una proteina con funzione di regolazione della trascrizione e che lega specifiche sequenze di DNA poste nella regione regolatrice di geni la cui espressione viene così modulata [2]. Si innesca così una catena di eventi trascrizionali regolativi che portando all'attivazione o silenziamento di specifici geni determina il peculiare trascrittoma che caratterizza le cellule muscolari . Inoltre la proteina MyoD1 interagisce direttamente con altre proteine che a loro volta contribuiscono alla formazione del muscolo. La ricostruzione delle caratteristiche trascrizionali dei geni coinvolti in questi eventi costituisce un primo passo per caratterizzare fasi fondamentali nello sviluppo del muscolo e quindi anche caratteristiche fenotipiche importanti nella produzione.

In Figura, sono schematizzate le fasi fondamentali nella determinazione e, di seguito, nel differenziamento muscolare.



In particolare è visualizzata la relazione tra l'espressione di alcuni geni determinanti e le modificazioni fenotipiche che portano dalla cellula staminale mesodermica alla fibra muscolare matura. La cascata molecolare

è innescata dal gene MyoD1 che insieme a Myf5 [3] regola la miogenina la cui espressione determina il passaggio fondamentale per il differenziamento terminale a miotubo e quindi a fibra scheletrica matura. In particolare tale cascata di eventi porta all'attivazione trascrizionale di quei geni che codificano per le proteine caratteristiche del muscolo quali la miosina, la mioglobina etc. Un ruolo particolarmente interessante è anche quello svolto dal regolatore trascrizionale Miostatina [4]. La proteina codificata da tale gene determina infatti una regolazione negativa nella proliferazione dei mioblasti, ovvero "modera" la quantità finale di mioblasti che concorrernno alla formazione dei miotubi. Esempi di alterata funzionalità di questo gene sono le razze bovine Belgian Blu e Piemontese nelle quali la inattivazione di miostatina, dovuta ad una sola sostituzione nucleotidica, porta al peculiare fenotipo detto "double muscle". Tale fenotipo è molto interessante anche perché evidenzia come un'opportuna conoscenza dei passaggi molecolari che determinano un certo fenomeno biologico possano poi essere utilizzabili ai fini di una mirata selezione e/o manipolazione.

1.1 Genomica comparata

Alla fine degli anni 80 è stato avviato il Progetto Genoma Umano i cui risultati sono stati resi noti nel 2000 dall'azienda privata Celera Genomics e nel 2001 dal consorzio pubblico internazionale che lo aveva lanciato [5,6]. Il progetto si proponeva di identificare la sequenza nucleotidica dei geni che caratterizzano la specie umana e di indicarne,

approssimativamente il numero. La sequenza fu pubblicata nel 2001 in un articolo su Nature [6], che combinava i risultati di entrambi i progetti. Si trattava di una bozza pari al 90% delle sequenze uniche e ancora con notevoli probabilità di errori. Una sequenza accurata al 99,99% è stata pubblicata nel 2003. Il primo risultato evidente è che sebbene il nostro organismo sia verosimilmente costituito da ben più di 100.000 proteine, esse quali risultano però essere sintetizzate a partire da "solo" 25.000 geni mediante vari meccanismi regolativi quali ad esempio lo splicing.

Sia pure nell'incompletezza del lavoro (le sequenze allora immesse in banca dati, rappresentano solo una piccola percentuale del totale del genoma) il progetto genoma ha totalmente rivoluzionato l'approccio agli studi biologici.

È nata così la branca scientifica nota come genomica [7] e tutta una serie di nuovi approcci culturali e tecnologici che partendo dal massimo sfruttamento del contenuto informativo dei dati provenienti dalla sequenza, stanno portando a numerose ricadute che non riguardano solo l'uomo ma, più o meno direttamente, tutti gli esseri viventi. La genomica prende le mosse dall'allestimento di complete mappe genetiche del DNA degli organismi viventi, proseguendo con il completo sequenziamento. La sequenza del DNA viene poi annotata, ovvero vengono identificati e segnalati tutti i geni e le sequenze significative dal punto di vista strutturale e funzionale, insieme a tutte le informazioni conosciute su tali geni. In questo modo è possibile ritrovare in maniera organizzata ed efficace le informazioni in appositi database, normalmente accessibili via Internet

gratuitamente. I risultati del progetto genoma umano hanno spinto all'avvio di nuovi progetti genoma per organismi d'interesse biologico e/o commerciale quali topo, batteri, lieviti etc. che grazie all'esperienza maturata sull'uomo, procedono in maniera più veloce e mirata.

L'accumularsi di dati di sequenza più o meno ordinati su organismi differenti ha portato alla nascita della cosiddetta genomica comparata.

La genomica comparata si basa sul confronto **[8]** tra i genomi di diversi organismi, nella loro organizzazione e sequenza ed anche delle varianti nell'ambito della stessa specie. La comparazione tra genomi diversi anche molto distanti evolutivamente consente di ricostruire i meccanismi attraverso cui ha agito la selezione nel corso dell'evoluzione ed anche la definizione di una filogenesi molecolare. L'analisi di un genoma è importantissima per la comprensione della biologia di un organismo, ma ancora più importante è l'analisi comparata di più genomi. Infatti, ci può aiutare ad identificare sia le regioni codificanti ed assegnare loro una funzione, che le regioni non tradotte coinvolte nella regolazione genica, e può permettere di dare nuove risposte a vecchie domande. Oggi la genomica comparata ha evidenziato non tanto la diversità dei geni in organismi diversi quanto la conservazione in questi d'intere famiglie geniche; quindi ha messo in luce che la diversità degli organismi è determinata non tanto dalla varietà dei geni quanto dalla loro diversa regolazione nell'espressione che dipende dalla fisiologia ed evoluzione della specie a cui l'organismo appartiene **[9]**.

Si tratta di un campo relativamente nuovo che si giova notevolmente dell'uso di strumenti informatici e che sta già dando molte ricadute pratiche.

Una di queste riguarda la possibilità di utilizzare informazioni genomiche ottenute in specie per le quali il sequenziamento del genoma sia già molto avanzato, per ottenere la sequenza di geni in specie per le quali esistono poche informazioni [10].

L'analisi comparativa rappresenta certamente l'approccio bioinformatico più rilevante per la caratterizzazione funzionale delle sequenze nucleotidiche e proteiche. I siti funzionalmente più rilevanti mostreranno infatti un elevato grado di conservazione o risulteranno invariati in tutte le sequenze considerate.

Al contrario i siti corrispondenti a regioni funzionalmente meno importanti mostreranno una maggiore variabilità.

In modo analogo si può predire la funzione di un gene o di una proteina sulla base dell'osservazione di somiglianza o similarità significativa con altri geni o proteine a funzione nota.

In questo contesto, risulta indispensabile lo sviluppo di adeguati strumenti provenienti dalla matematica, informatica, fisica, statistica per la soluzione di problemi derivanti dall'analisi di sequenze biologiche.

1.1.2 Algoritmi di allineamento

Tutte le problematiche computazionali inerenti alla progettazione, l'implementazione e l'applicazione di matematico-statistici rivolti alla

caratterizzazione funzionale delle bio-sequenze, vengono spesso fatti rientrare nel settore della Biologia Computazionale**.** **[11]**

Il tema centrale della Biologia Computazionale è la computazione su sequenze molecolari (stringhe); questo costituisce un importante punto di contatto tra la Biologia e l'Informatica in quanto la computazione su stringhe è materia di ricerca di notevole interesse in diversi settori del calcolo computazionale. Una branca fondamentale in quest'ambito, è lo studio degli algoritmi di allineamento di stringhe di natura biologica.

In generale, date due stringhe di sequenze, un allineamento a coppie, permette di mettere in evidenza come una sequenza può essere ottenuta da un'altra tramite una serie di operazioni di sostituzione, cancellazione e inserimento. Nel caso in cui le sequenze in questione rappresentano macromolecole biologiche, come filamenti di DNA o proteine, se questa serie di operazioni è fisicamente probabile significa che le due sequenze sono vicine dal punto di vista biologico o evolutivo; in tal caso le sequenze possono rappresentare proteine con struttura e funzioni simili

Tipicamente, tuttavia, le sequenze sono raggruppabili in famiglie: di solito si assume che le differenze all'interno di una famiglia siano conseguenza di mutazioni da un antenato comune avvenute nel corso dell'evoluzione.

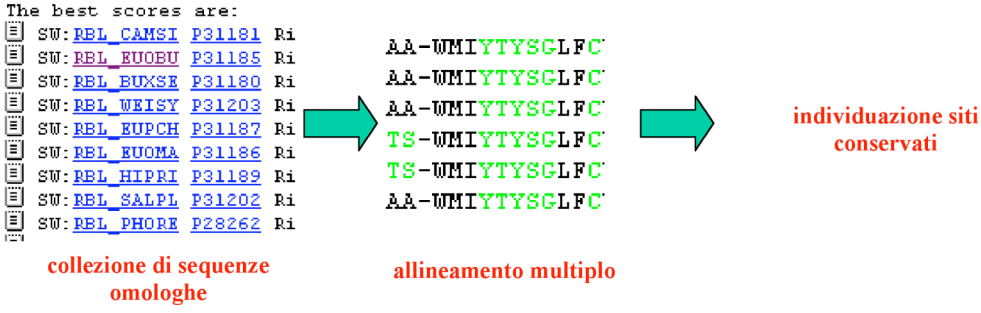
Sequenze appartenenti ad una famiglia solitamente sono simili dal punto di vista funzionale, anche se possono essere distanti dal punto di vista strutturale. **[12]**

Sono desiderabili pertanto metodi di analisi che permettano di

stabilire la relazione tra una data sequenza e una famiglia di sequenze. Ciò consente spesso di inferire alcune caratteristiche funzionali della sequenza in esame. A tale scopo il confronto a coppie non è adeguato, in quanto non consente di cogliere aspetti complessivi di una famiglia statisticamente rilevanti. Ad esempio, certe parti delle sequenze possono essere maggiormente conservate rispetto ad altre: è opportuno allora verificare fino a che grado queste parti siano presenti in una nuova sequenza di cui si vuole testare la relazione con la famiglia. Per poter eseguire analisi comparative è dunque necessario collezionare un certo numero di sequenze omologhe ed organizzarle in un allineamento multiplo. [13]

Per costruire la collezione di geni omologhi da cui iniziare le analisi è sufficiente eseguire un'interrogazione testuale in banca dati (ad es. fornendo il nome del gene) oppure effettuare una ricerca di similarità di sequenza mediante algoritmi di database searching usando come sonda (query) una sequenza nota del gene. [12]

L'allineamento multiplo è rappresentato sotto forma di una tabella costituita da righe, corrispondenti alle sequenze omologhe considerate, e da colonne, corrispondenti a ciascun sito dell'allineamento.



In generale dunque, per poter eseguire analisi comparative tra stringhe rappresentanti molecole di natura biologica è necessario sviluppare alcuni metodi (algoritmi) di confronto di sequenze, facendo riferimento sia agli allineamenti a coppie sia a quelli multipli.

Per entrambe le metodologie esistono due tipi di tecniche d'allineamento, basate su ***algoritmi esatti*** e ***algoritmi euristici o sub-ottimi*** **[12]**.

Per quanto concerne gli algoritmi esatti di coppie di sequenze di aminoacidici, esistono due tipi di algoritmi, algoritmo di *Wunsch-Needleman* **[14]**, e quello di *Smith-Waterman* **[15]**, che sono in grado di determinare il migliore allineamento possibile tra due sequenze in base ad una determinata matrice di sostituzione **[16]**.

Gli algoritmi esatti sono ideali per ottenere il miglior allineamento, ma sono troppo lenti perchè effettuino ricerche di similarità in banche dati. Infatti, data una sequenza sonda, essa dovrà essere allineata con ciascuna sequenza presente in banca dati, vale a dire milioni di sequenze. Il che equivale a dire che il tempo di esecuzione diventerebbe di decine di migliaia di secondi, equivalenti a molte ore.

Ecco che si sono sviluppati programmi in grado di portare a termine velocemente ricerche di similarità, grazie a soluzioni euristiche che sono basate su assunzioni non certe, ma estremamente probabili. Due sono i software più popolari: *Blast* **[17]** e *Fasta* **[18]**.

Per quanto riguarda gli allineamenti multipli, si potrebbero applicare degli algoritmi esatti di due sequenze a singole coppie dell'allineamento,

però il tempo di esecuzione, anche usando il più potente computer, sarebbe decisamente lungo [19].

Infatti, se le sequenze da allineare sono pari ad un numero N, applicando un algoritmo esatto, avremmo una rappresentazione dell'allineamento come percorso in una matrice N-dimensionale.

Se la lunghezza delle sequenze da allineare è L ed n sono le sequenze, la complessità dell'algoritmo è dell'ordine di L^n (per due sequenze è $L \times L$). Di fatto è impraticabile allineare più di una decina di sequenze con questi metodi esatti [20].

Quindi si usano algoritmi euristici meno precisi ma più facili da computare.

Vediamo in dettaglio i passi salienti di due algortimi euristici, Blast e ClustalW [21].

1.1.2 Blast

In realtà Blast e Fasta rientrano nella categoria dei programmi di ricerca di similarità di una sequenza (query) all'interno di un database ma producono un output in forma di allineamento della sequenza query con le sequenze del database con cui hanno mostrato una similarità significativa.

BLAST che è l'acronimo di *Basic Local Alignment Search Tool*, è un programma euristico per la ricerca di omologie locali di sequenza basato sulla dimostrazione data da Karlin & Altschul ricercatori del NCBI(1990) ed è in realtà costituito da un insieme di **5** programmi:

BLASTP paragona una sequenza aminoacidica ad un database di sequenze proteiche.

BLASTN paragona una sequenza nucleotidica ad un database di sequenze nucleotidiche

BLASTX paragona una sequenza nucleotidica (traducendola in tutti 6 possibili frame di lettura) ad un database di proteine. Di tutti i programmi che fanno parte di **BLAST** è il più usato.

TBLASTN paragona una sequenza aminoacidica ad un database di acidi nucleici tradotto dinamicamente nelle 6 possibili sequenze di aminoacidi che possono derivarne.

TBLASTX paragona una sequenza nucleotidica letta secondo tutti i 6 possibili frame di lettura con un database di acidi nucleici anch'esso letto secondo tutti i 6 possibili frame di lettura. Poiché ne derivano 36 combinazioni, questo programma viene utilizzato solo per ricerche su database di tipo EST.

Analogamente a FASTA, BLAST ricerca il migliore allineamento fra l'*intera sequenza* sottoposta ad indagine e il database di sequenze usato come riferimento.

BLAST usa inoltre una *scoring matrix* durante tutte le fasi della ricerca (scansione ed estensione), a differenza di FASTA che usa una *scoring matrix* solo durante la fase di estensione del confronto **[22]**.

Inoltre, mentre FASTA esamina gli aminoacidi a coppie (*ktup=2*) o singolarmente presi (*ktup=1*), BLAST utilizza per il confronto gruppi di 3-4 aminoacidi (**words**) il che consente una *velocizzazione* del processo. Per far fronte alla riduzione di specificità derivante dall'uso di questi gruppi piuttosto "*ampi*", BLAST prende in considerazione solo quei gruppi di 3-4 aminoacidi il cui punteggio è superiore ad un *valore-soglia* **T (CUTOFF)**, in modo che l'eventuale omologia identificata possa considerarsi probabile (su base statistica) già *a priori*.

Così come prevede l'algoritmo che governa le prime fasi di FASTA, anche BLAST *non* ammette la presenza di *gap* all'interno di ciascun segmento di sequenza preso in considerazione.

A differenza di FASTA che nell'ultima fase prende in considerazione eventuali inserzioni e delezioni nei segmenti allineati, BLAST *non* contempla tale possibilità in nessuna fase.

L'esecuzione del programma è suddiviso in 3 fasi.

1ª Fase: **Compilazione di una lista di words a punteggio superiore ad un valore-soglia T**

2ª Fase: **Scansione del database per ricercare le corrispondenze**

3ª Fase: **Estensione della ricerca delle zone di corrispondenza**

1ª Fase: Compilazione di una lista di words a punteggio superiore ad un valore-soglia T

La prima fase consiste nella creazione di un elenco di parole creato leggendo a una a una tutte le parole di lunghezza *w* della sequenza query (generalmente per le proteine *w=3* e per i nucleotidi *w=12)*

Il numero totale di *words* presenti in una sequenza da sottoporre a confronto, risulta essere:

$$n = l - w + 1$$

ove *w* è il numero degli aminoacidi che compongono una *word* ed *l* è la lunghezza della sequenza in esame.

Esempio:

Sia data la sequenza di lunghezza *l=10*: `KLTWASNGTD`. Sia $w = 3$. Allora il numero totale di *words* presenti è

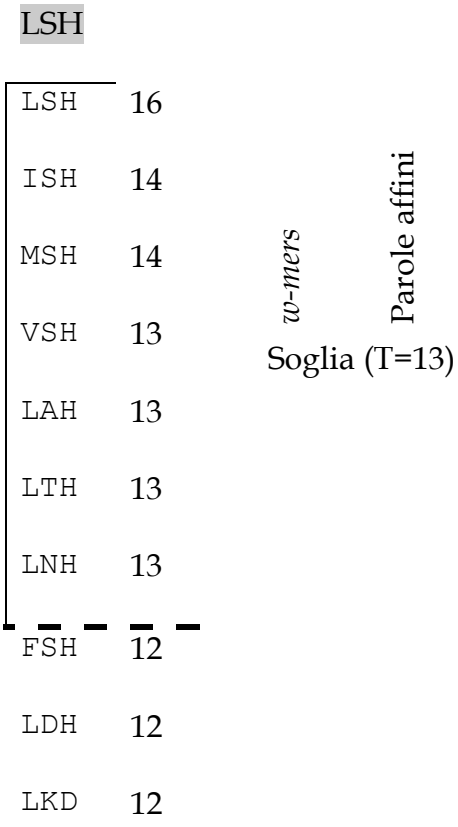
$$n = l - w + 1 = 10 - 3 + 1 = 8$$

LSHWASNGTD	LSHWASNGTD	LSHWASNGTD	LSHWASNGTD
1°word	2°word	3°word	4°word
LSHWASNGTD	LSHWASNGTD	LSHWASNGTD	LSHWASNGTD
5°word	6°word	7°word	8°word

Per ogni *word* della sequenza da esaminare viene costruita una lista di possibili *"parole affini"* chiamate *w-mers*, se confrontate con la sequenza in questione, abbiano un punteggio superiore ad un *valore-soglia* **T** (compreso fra 11 e 15) calcolato di volta in volta in base alla composizione e

alla lunghezza della sequenza in esame e in base alla matrice di sostituzione **[3]** utilizzata (normalmente PAM 120 o BLOSUM 62).

Tutti questi *w-mers* che presentano uno score sopra la soglia **T** sono inseriti nell'elenco. Nel nostro esempio costruiamo una creazione di parole affini relative al primo *word*



Nel caso dei nucleotidi (**BLASTN**) il punteggio è di più semplice valutazione: viene assegnato un punteggio di **+5** ad una identità di residui e di **-4** per una mancata identità.

Dati questi presupposti, si è visto che la combinazione che è il miglior compromesso fra sensibilità, specificità del metodo e velocità di esecuzione del confronto fra le sequenze, è quella con $w = 3$ e T**=11-15**.

Utilizzando questi valori, si ottengono delle liste di circa *50 w-mers di confronto* denominate ***neighbors*** per ogni *word* della sequenza da testare, cioè circa 12.500 *w-mers* nel caso di una sequenza di 250 aminoacidi.

Questo dato è ben diverso dalle 20^3 combinazioni possibili (per w = 3) per ciascuna *word* della sequenza da testare, che sarebbero necessarie se non venisse effettuata questa preselezione.

2ª Fase: Scansione del database per ricercare le corrispondenze

In questa fase ciascuna delle *w-mers* della lista compilata (12.500 circa nel caso di una sequenza di 250 aminoacidi), viene confrontata con il database delle sequenze.

Ogni corrispondenza trovata (hit) potrebbe rappresentare una porzione di un possibile allineamento più esteso e viene pertanto considerata come tale.

3ª Fase: Estensione della ricerca delle zone di corrispondenza

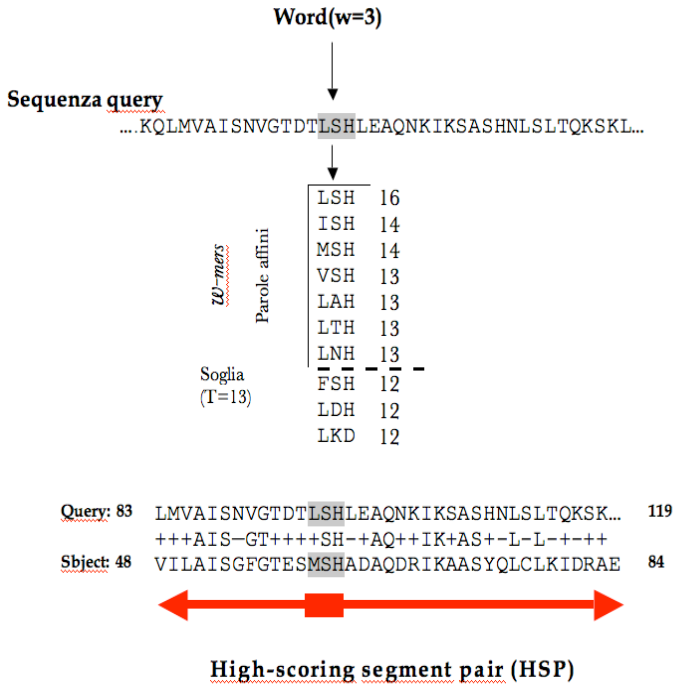
Quando viene riscontrata una corrispondenza ***(hit)***, essa viene estesa a monte e a valle (senza la possibilità di inserire GAP) per vedere se è possibile definire un tratto di sequenza in grado di raggiungere un punteggio superiore ad un *valore-soglia* detto ***S***.

Tale valore ***S***, è funzione di un altro valore, detto ***E***, che è il numero atteso (*Expected*) di tratti di sequenza *casualmente* omologhi, aventi punteggio superiore a ***S***.

Come detto, c'è una relazione tra ***E*** ed ***S***: tanto più elevato è ***E***, tanto minore diventa ***S*** (nel senso di essere più restrittivi e pertanto abbassare il valore di soglia ***S***), per cui aumenta la sensibilità del risultato, ma si riduce del pari la specificità del metodo.

I tratti di sequenza omologhi aventi un punteggio (*score*) superiore al *valore-soglia* ***S***, vengono denominati **HSP** (*High Score Segment Pair*). Essi possono essere anche più di uno all'interno di una medesima sequenza e definiscono una *zona locale di omologia*.

Nell'esempio illustrato **[12]**



dopo completata la lista, BLAST cerca corrispondenze esatte (hits) dei *w-mers* nelle sequenze della banca dati. Ogni volta che un hit viene

identificato il programma verifica quanto sia possibile estenderlo e se il risultante HSP sia superiore alla soglia ***S***.

Sempre nell'esempio illustrato sopra, immediatamente a destra dell'allineamento `LSH-MHS` si trova un appaiamento `L-A` caratterizzato da uno score negativo, (deducibile da un segno "-" sulla line centrale dell'allineamento); successivamente si noti che il programma non abbia immediatamente rinunciato a estendere ulteriormente l'allineamento, nonostante la presenza di uno score negativo. Sul lato sinistro dell'allineamento si può osservare la presenza di due posizioni consecutive con score negativi (`NV-GF`); anche in questo caso il programma ha comunque provato a estendere l'allineamento riuscendo infatti ad aumentare lo score totale del segmento.

A questo proposito viene introdotto un altro parametro ***X*** (misurato in termini di perdita di score) che stabilisce quanto il programma debba insistere nel cercare di estendere gli HSP quando vengono incontrati termini negativi.

In definitiva i principali parametri usati dall'algoritmo BLAST sono dunque quattro:

w, T, E, X

Particolarmente importanti sono ***w, T*** perché determinano la grandezza della lista dei *w-mers*. Per un dato valore di ***w***, più è basso il valore di ***T*** e più risulterà estesa la lista dei *w-mers*, con il corrispondente aumento del tempo di esecuzione del programma. D'altra parte, valori alti

di ***T*** portano a un aumento del rischio di non identificare alcun HSP; ci potrebbero infatti essere HSP che superano lo score pur non avendo *w-mers* maggiori o uguali a ***T***.

Anche il parametro ***X*** influenza le prestazioni del programma. Aumentando il valore di ***X*** aumenta il tempo esecuzione perché l'intorno di ogni hit è esplorato a maggiore profondità; anche in questo caso bisogno dunque trovare un compromesso tra la velocità di esecuzione del programma e sensibilità dell'analisi.

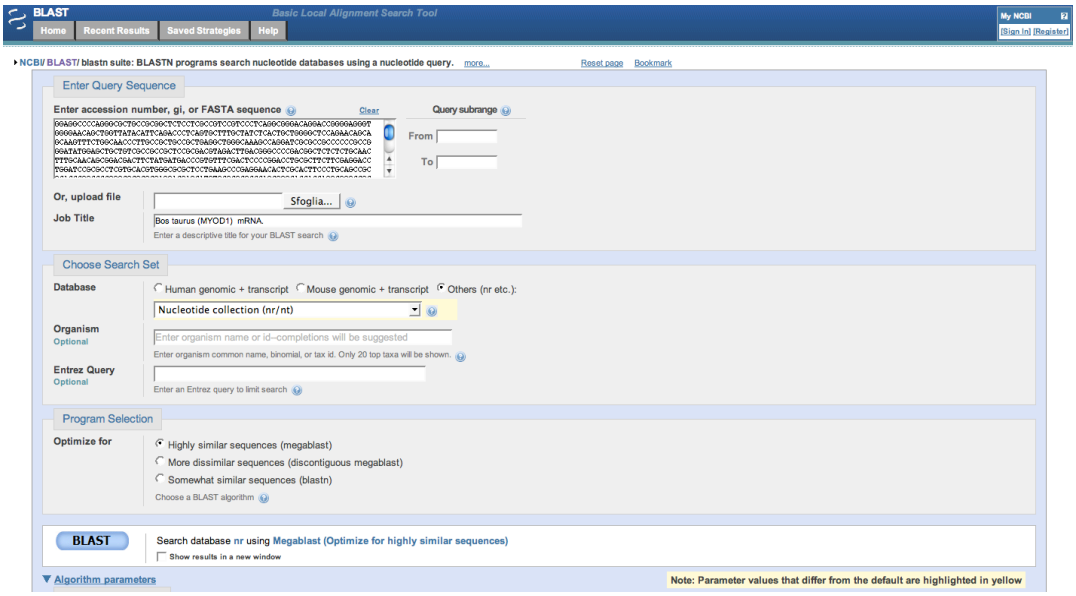
In considerazione dal fatto che è più facile pensare in termini di ***E*** piuttosto che di ***S***, il programma calcola automaticamente il valore di ***S*** impostato dall'utente.

Se il valore di ***E*** non viene definito, il programma lo imposta automaticamente a 10; viceversa, impostando per esempio ***E***=0,1, il valore di soglia sarà maggiore e conseguentemente non saranno considerati HSP con scores bassi ma che raggiungono tale soglia.

Molto spesso si preferisce impostare un valore di ***E*** piuttosto piccolo per evitare che il programma restituisca allineamenti con scorse bassi, generalmente privi di significato.

L'implementazione più popolare dell'algoritmo BLAST si trova all'interno del sito dell'NCBI.

http://www.ncbi.nlm.nih.gov/blast



Pagina web sul sito del NCBI per eseguire BLAST su banche dati di nucleotidi. La pagina è divisa in tre settori.

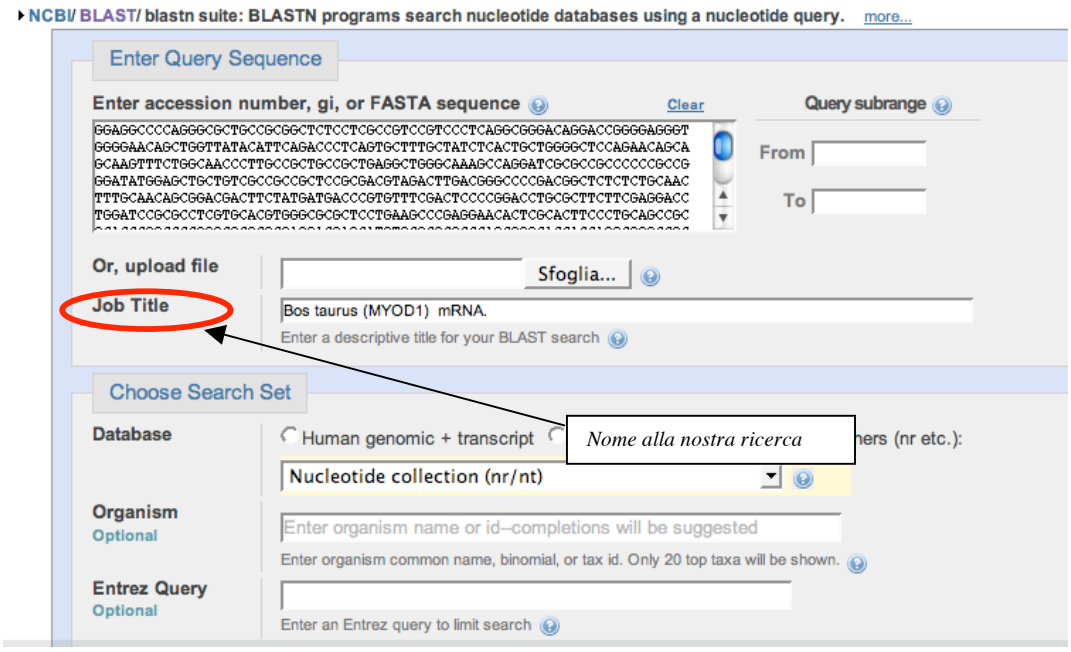
1) In quello superiore si deve "*incollare*" la sequenza query oppure recuperarla da un file esterno. E' inoltre possibile dar un nome al lavoro di ricerca che si sta effettuando;
2) Nella sezione centrale si deve selezionare la banca dati su cui effettuare la ricerca; dove esiste anche la banca di sequenze non ridondanti <<nr>>, sviluppata al NCBI.
3) Nella sezione inferiore si possono selezionare gli algortimi da utilizzare (megablas. PSI-Blast).

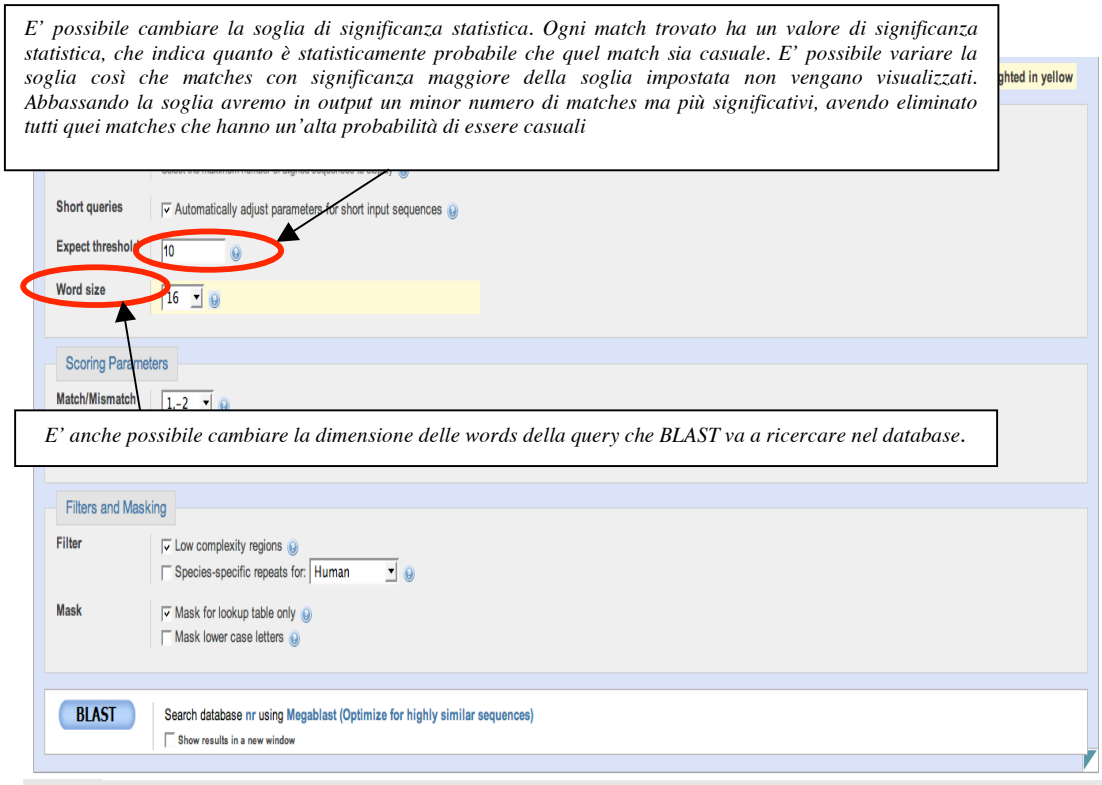
Inoltre vi è una sezione aggiuntiva (*Algorithm parameters*) dove possono essere definiti diversi parametri tra cui ***E*** (*Expect*), ***w*** (*Word Size*), la matrice di sostituzione e le penalità per ogni GAP. E' inoltre possibile impostare i filtri per mascherare regioni di sequenza, per esempio quelle a bassa complessità. Al seguente link:

http://www.ncbi.nih.gov/Education/BLASTinfo/information3.html

è possibile avere delle informazioni dettagliate del software BLAST. Vediamo un esempio utilizzando BLASTN.

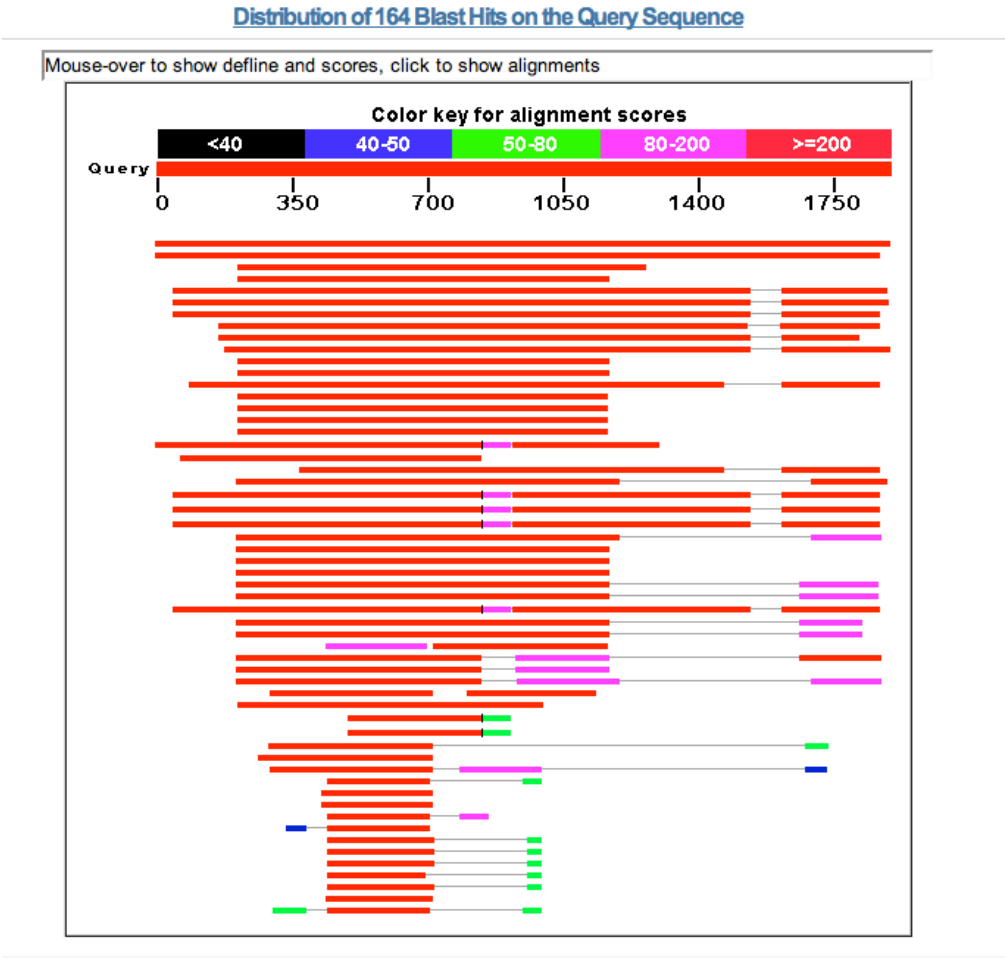
Dopo aver eseguito una ricerca in banca dati della sequenza di *mRNA* del gene *Myod* di *Bos taurus*, incolliamo la sequenza (formato FASTA) nel box di Blast:





Una volta definiti i parametri, cliccando prima su BLAST si ottiene il risultato della ricerca:

BLAST fornisce in output la distribuzione dei matches trovati, assegnando a colori diversi i diversi scores: ovviamente uno score maggiore indica un match più significativo.



BLAST fornisce ovviamente anche l'elenco delle sequenze trovate, dove all'interno è possibile individuare l'*accession number,* la descrizione degli allineamenti e soprattutto il valore in percentuale di identità. Di sotto una parte di questo elenco:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_001040478.2	Bos taurus myogenic differentiation 1 (MYOD1), mRNA >gb|BC120454.1| Bos taurus myogenic differentiation 1, mRNA (cDNA clone IMAGE:8287333)	3504	3504	100%	0.0	100%	U G
NM_001009390.1	Ovis aries myogenic differentiation 1 (MYOD1), mRNA >emb|X62102.1|OAMYOD1 O.aries mRNA for MyoD1	3072	3072	98%	0.0	96%	U G
AB110599.1	Bos taurus mRNA for MyoD, complete cds	1943	1943	55%	0.0	99%	U G
NM_001002824.1	Sus scrofa myogenic differentiation 1 (MYOD1), mRNA	1399	1399	50%	0.0	93%	U G
NM_002478.4	Homo sapiens myogenic differentiation 1 (MYOD1), mRNA	1363	1715	92%	0.0	90%	U E G
BC000353.1	Homo sapiens cDNA clone IMAGE:2961494, **** WARNING: chimeric clone ****	1358	1710	93%	0.0	89%	E
XM_508311.2	PREDICTED: Pan troglodytes myogenic differentiation 1 (MYOD1), mRNA	1352	1679	92%	0.0	90%	G
XM_001088038.1	PREDICTED: Macaca mulatta similar to myogenic differentiation 1 (LOC696222), mRNA	1327	1649	85%	0.0	89%	G
CR612159.1	full-length cDNA clone CS0DI057YK10 of Placenta Cot 25-normalized of Homo sapiens (human)	1327	1575	82%	0.0	89%	U G
BC064493.1	Homo sapiens myogenic differentiation 1, mRNA (cDNA clone MGC:71135 IMAGE:5022419), complete cds	1319	1686	86%	0.0	90%	U G
XM_001504953.1	PREDICTED: Equus caballus similar to MyoD (LOC100071803), mRNA	1315	1315	50%	0.0	91%	G
XM_849663.1	PREDICTED: Canis familiaris similar to myogenic factor 3 (LOC611940), mRNA	1304	1304	50%	0.0	91%	G
X56677.1	Human MyoD mRNA	1271	1568	86%	0.0	88%	U E G
BT007461.1	Synthetic construct Homo sapiens myogenic factor 3 mRNA, partial cds	1155	1155	50%	0.0	88%	
BT007157.1	Homo sapiens myogenic factor 3 mRNA, complete cds	1155	1155	50%	0.0	88%	U G
AY891398.1	Synthetic construct Homo sapiens clone FLH022867.01L myogenic factor 3 (MYOD1) mRNA, partial cds	1155	1155	50%	0.0	88%	
AY888744.1	Synthetic construct Homo sapiens clone FLH022871.01X myogenic factor 3 (MYOD1) mRNA, complete cds	1155	1155	50%	0.0	88%	
U12574.1	Sus scrofa myogenic regulatory factor MyoD (myoD) gene, complete cds	1127	1669	68%	0.0	91%	G
AY646094.1	Sus scrofa eukaryotic myogenic factor MYF-3 (MYF-3) gene, exon 1 and partial cds	1088	1088	41%	0.0	92%	
X17650.1	Human Myf-3 mRNA for myogenic determining factor 3'-fragment	983	1280	71%	0.0	88%	U G

cliccando sullo *max score*, si ottiene l'allineamento corrispondente:

In questo frammento di allineamento abbiamo il 96% di corrispondenze e si tratta di un *Ovies aries*

```
>ref|NM_001009390.1|  Ovis aries myogenic differentiation 1 (MYOD1), mRNA
 emb|X62102.1|OAMYOD1  O.aries mRNA for MyoD1
Length=2022

 Score = 3072 bits (1663),  Expect = 0.0
 Identities = 1822/1895 (96%), Gaps = 26/1895 (1%)
 Strand=Plus/Plus

Query  1     GGAGGCCCCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGGGACAGGAC  60
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  129   GGAGGCCTCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGGGACAGGAC  188

Query  61    CGGGGAGGGTGGGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCTATCTCACT  120
             |||||||||| ||||||||||||||||||||||||||||||||||||| ||||| |
Sbjct  189   CGGGGAGGGTGAGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCTCTCTCATT  248

Query  121   GCTGGGGCTCCAGAACAGCAGCAAGTTTCTGGCAACCCTTGCCGCTGCCGCTGAGGCTGG  180
             || |||||||||||||| |||| |||||||||||||| ||||||||||||||||||||||
Sbjct  249   GCCGGGGCTCCAGAACCGCAGTAAGTTTCTGGCAACCCCTGCCGCTGCCGCTGAGGCTGG  308

Query  181   GCAAAGCCAGGATCGCGCCGCCCCCCGCCGGGATATGGAGCTGCTGTCGCCGCCGCTCCG  240
             |||||||||||| |||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  309   GCAAAGCCAGGACCGCGCCGCCCCACGCCGGGATATGGAGCTGCTGTCGCCGCCGCTCCG  368
```

In questo frammento di allineamento (*Sus scrofa*) abbiamo il 90% di corrispondenze

```
>ref|NM_001002824.1|  Sus scrofa myogenic differentiation 1 (MYOD1), mRNA
Length=960

 Score = 1399 bits (757),  Expect = 0.0
 Identities = 895/962 (93%), Gaps = 7/962 (0%)
 Strand=Plus/Plus

Query  215   ATGGAGCTGCTGTCGCCGCCGCTCCGCGACGTAGACTTGACGGGCCCCGACGGCTCTCTC  274
             |||||||||||||||| |||||||||||||||| |||||||||||||||||||||||||
Sbjct  1     ATGGAGCTGCTGTCGCCACCGCTCCGCGACGTAGATTTGACGGGCCCCGACGGCTCTCTC  60

Query  275   TGCAACTTTGCAACAGCGGACGACTTCTATGATGACCCGTGTTTCGACTCCCCGGACCTG  334
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    TGCAACTTTGCAACAGCGGACGACTTCTATGATGACCCGTGTTTCGACTCCCCGGACCTG  120

Query  335   CGCTTCTTCGAGGACCTGGATCCGCGCCTCGTGCACGTGGGCGCGCTCCTGAAGCCCGAG  394
             |||||||||||||||||| |||||||||||||||||||||||||| |||||||||
Sbjct  121   CGCTTCTTCGAGGACCTGGACCCGCGCCTCGTGCACGTGGGCGCGCTCCTAAAGCCCGAG  180
```

1.1.3 ClustalW

Il metodo più comune di eseguire un allineamento multiplo euristico è il cosiddetto allineamento progressivo (*progressive alignment*), basato sulla costruzione di una successione di allineamenti a coppie **[20]**.

Si scelgono due sequenze e si allineano: da quest'allineamento si ottiene una *sequenza consenso*, ossia una sequenza che presenta solo i residui più conservati per ogni posizione **[21]**.

Una sequenza consenso:

1. Riassume un multiallineamento;
2. Si possono definire dei simboli che la definiscano e che indichino anche conservazioni non perfette in una posizione;
3. E' possibile utilizzare una formattazione precisa che permetta di capire anche le variazioni in una posizione, non solo le conservazioni.

Dopodichè si sceglie una terza sequenza e si allinea al precedente allineamento, e cosi via. Questo approccio è euristico e non garantisce di trovare l'allineamento ottimo: per contro, è efficiente e spesso da risultati ragionevoli.

L'euristica più importante utilizzata negli algoritmi d'allineamento progressivo prevede che le coppie di sequenze con maggiore grado di somiglianza o, equivalentemente, la cui "distanza genetica" sia minore, siano allineate per prime. Questo modo di procedere è giustificato dal fatto che coppie di sequenze maggiormente somiglianti hanno maggiore

probabilità di essere derivate più recentemente da un antenato comune, e quindi il loro allineamento fornisce l'informazione più "affidabile" che è possibile ricavare dalle sequenze. In particolare, le posizioni dei gap in sequenze maggiormente correlate sono tipicamente più accurate rispetto a quelle relative a sequenze meno simili.

Ciò porta a formulare la regola euristica per cui i gap degli allineamenti iniziali vadano preservati quando si allineano nuove sequenze (*once a gap, always a gap*). Molti algoritmi di questo tipo utilizzano i cosiddetti alberi guida (*guide tree*), alberi binari le cui foglie sono etichettate con sequenze e i cui nodi interni rappresentano gruppi (*cluster*) di sequenze. Gli alberi guida sono simili agli alberi filogenetici, ma poiché il loro scopo è solo quello di determinare l'ordine in cui effettuare un allineamento progressivo, la loro costruzione è meno accurata e l'informazione che forniscono è meno precisa rispetto ad un vero e proprio albero filogenetico.

L'algoritmo di *Feng-Doolittle* **[22]** è uno dei primi algoritmi per l'allineamento progressivo, in cui trovano applicazione le idee appena descritte.

Durante un allineamento progressivo è però vantaggioso usare l'informazione, dipendente dalla posizione del residuo, che si acquisisce quando si è allineato un gruppo di sequenze.

A tale scopo vengono implementati algoritmi che si basano sui ***profili*** ossia schemi di valutazione dipendente dalla posizione. Un noto algoritmo che si basa su questa tecnica è quello di *Thompson-Higgins-Gibson* **[22]**.

I passi dell'algoritmo di *Thompson-Higgins-Gibson*, sono a grandi linee i seguenti:

1. Il punto di partenza di un allineamento multiplo progressivo è l'allineamento tra tutte le possibili coppie di sequenze. Date k sequenze, si dovranno effettuare.

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$$

allineamenti a coppie.

2. I punteggi di similarità saranno calcolati mediante le matrici di distanze filogenetiche;

3. I punteggi ottenuti vengono utilizzati poi per la costruzione di alberi o dendogrammi mediante il metodo di clustering UPGMA (*Unweighted Pair Group Method with Arithmetic mean*), ossia metodi di raggruppamento a coppie non pesate che utilizza medie aritmetiche;

4. Allineamento progressivo delle sequenze seguendo l'albero guida. L'idea di base consiste nel utilizzare una serie di allineamenti, usando appunto allineamenti *sequenza-sequenza*, *sequenza-profilo*,

profilo-profilo, seguendo l'ordine suggerito dall'albero guida, procedendo dalle "foglie" verso la "radice"

Ogni passo dell'algoritmo consiste dunque nel raggruppare due nodi in modo da creare un nuovo nodo seguendo la guida dell'albero, utilizzato un'algoritmo di programmazione dinamica per coppie di allineamenti oppure considerando gli allineamenti sequenza-profilo oppure profilo-profilo, mantenendo fissi i gap dei due allineamenti iniziali, che vengono considerati semplicemente come nuovi simboli dell'alfabeto, e aggiungendo nuovi gap indipendenti dai precedenti.

Il calcolo dei gap al passo k è indipendente da quello effettuato al passo k+1 nel senso che l'inserimento di un nuovo gap in un gap gia esistente è considerato di nuovo come un *gap opening*.

Una nota implementazione dell'allineamento progressivo mediante profili è il programma CLUSTALW **[22]**.

CLUSTALW fa uso inoltre di molte regole ad hoc. Ad esempio, le sequenze di una famiglia hanno associato un peso (che serve per compensare un eventuale sbilanciamento nella distribuzione statistica delle sequenze); si usano matrici di sostituzione diverse a seconda del grado di similarità fra le sequenze da confrontare; i punteggi di gap variano in relazione alla frequenza dei residui allineati con i gap.

1.2 Basi di dati

Le basi di dati e la tecnologia delle basi di dati stanno esercitando un'influenza fondamentale nell'uso sempre più esteso del calcolatore.

Oggigiorno le basi di dati giocano un ruolo fondamentale in quasi tutti i campi in cui i calcolatori sono utilizzati, dal mondo degli affari, dell'ingegneria, della medicina, della biologia.

Negli ultimi anni l'evoluzione delle tecnologie ha portato a nuove e suggestive applicazioni di sistemi di basi di dati, in particolare le tecniche di ricerche proprie delle basi di dati che sono applicate al World Wide Web per migliorare la ricerca dell'informazione desiderata dagli utenti che navigano in internet.

1.2.1 Sistemi informativi, informazioni e dati

Nello svolgimento di ogni attività, sia a livello individuale sia in organizzazioni di ogni dimensione, sono essenziali la disponibilità di informazioni e la capacità di gestirle in modo efficace; ogni organizzazione è dotata di un *sistema informativo*, che organizza e gestisce le informazioni necessarie per perseguire gli scopi dell'organizzazione stessa. **[24]**

L'esistenza del sistema informativo è in parte indipendente dalla sua automatizzazione. A sostegno di quest'affermazione possiamo ricordare che i sistemi informativi esistono da molto prima dell'invenzione e della diffusione dei calcolatori elettronici; per esempio, gli archivi delle banche o dei servizi anagrafici sono istituiti da vari secoli. Per indicare la porzione automatizzata del sistema informativo viene di solito utilizzato il termine di

sistema informatico. La diffusione capillare dell'informatica a quasi tutte le attività umane, fa sì che gran parte dei sistemi informativi sia, anche, in buona misura, sistemi informatici.

Nei sistemi informatici, le informazioni sono rappresentate per mezzo dei dati i quali necessitano di una interpretazione per fornire *informazioni*. Sebbene molti ritengono questi due termini intercambiabili la loro differenza è abissale. I *dati* sono solo una serie di fatti mentre le *informazioni* sono conoscenza. Le informazioni sono costituite da dati organizzati e presenti in modo tale da risultare utili nel processo decisionale.

1.2.2 Basi di dati: definizione

Dal concetto di *dato*, deriva direttamente quello di *base di dati*. Infatti possiamo definire una ***base di dati*** [24] una collezione di dati correlati aventi le seguenti proprietà implicite:

- Rappresenta un certo aspetto del mondo reale, talvolta detto ***mini-mondo*** o ***universo del discorso*** dove un cambiamento del *mini- mondo* si riflettono sulla base di dati;
- I dati sono logicamente coerenti con un significato intrinseco; un assortimento casuale di dati non può essere correttamente considerato una base di dati;
- E' progettata, costruita, popolata con dati per uno scopo specifico.

In altre parole, una base di dati ha una sorgente dalla quale i dati sono derivati, un certo grado di interazione con gli eventi del mondo reale, e un pubblico che è attivamente interessato al suo contenuto.

1.2.3 Sistemi di gestione di basi di dati

Una base di dati può essere generata e mantenuta manualmente oppure essere computerizzata. Una base di dati computerizzata può essere prodotta e mantenuta da un gruppo di programmi applicativi scritti specificamente allo scopo o da un sistema di gestione di basi dati.

Un sistema di gestione di base dati (DBMS, *Database Management System*) è un sistema software in grado di gestire collezioni di dati che siano *grandi, condivise e persistenti* assicurando la loro *affidabilità e privatezza*. Inoltre, in quanto prodotto informatico, deve essere *efficiente* e *efficace*. Una base di dati è una collezione di dati gestita da un DBMS **[25]**.

Riassumendo, le caratteristiche dei DBMS e delle basi di dati sono:

1) Le basi di dati sono *grandi*: nel senso che possono avere anche dimensioni enormi (terabyte e oltre) e quindi oltre le capacità della memoria centrale di un elaboratore. Di conseguenza un DBMS deve essere in grado di gestire memorie secondarie.

2) Le basi di dati sono *condivise:* perché un DBMS deve permettere a più utenti di accedere contemporaneamente ai dati comuni. In tal modo viene anche ridotta la *ridondanza* e *inconsistenza* dei dati, dato che esiste una sola copia dei dati. Per controllare l'accesso condiviso di più utenti il DBMS dispone di un meccanismo apposito, detto controllo di concorrenza .

3) Le basi di dati sono *persistenti*, cioè hanno un tempo di vita che non è limitato a quello delle singole esecuzioni dei programmi che lo utilizzano;

4) I DBMS garantiscono *l'affidabilità* cioè la capacità del sistema di conservare intatto il contenuto della base di dati, in caso di malfunzionamento. I DBMS forniscono, per tali scopi, procedure di salvataggio e ripristino della base di dati (*backup* e *recovery*).

5) I DBMS garantiscono la *privatezza* dei dati cioè ciascun utente viene abilitato a svolgere solo determinate azioni sui dati, attraverso meccanismi di autorizzazione.

6) I DBMS devono garantire *l'efficienza* e *l'efficacia* ossia la capacità di svolgere le operazioni utilizzando un insieme di risorse (tempo e spazio) accetabile dagli utenti, e di rendere produttive, le attività dei suoi utenti

Affinché i dati di interesse siano correttamente organizzati e la loro struttura sia descritta in un modo comprensibile per l'elaboratore, si ricorre ad un insieme di concetti detto "modello dei dati". Esistono due categorie principali di tali modelli:

- *Modelli logici*: utilizzati nei DBMS esistenti per l'organizzazione dei dati, ad essi fanno riferimento i programmi. I modelli logici utilizzano strutture che, pur astratte, riflettono una particolare organizzazione logica dei dati (Relazionale, Gerarchico, Reticolare,

ad Oggetti) ed è per questo motivo che la conoscenza del modello logico è necessaria per l'utilizzo di una base di dati;

- *Modelli concettuali*: permettono di rappresentare i dati in maniera completamente indipendente dal modello logico e sono utilizzati nelle fasi preliminari di progettazione. Il termine "concettuale" deriva dal fatto che tali modelli tendono a descrivere concetti del mondo reale, piuttosto che i dati utili per rappresentarli. Il più noto è *Entity-Relationship* (ER) [25] .

1.2.4 Modelli di dati

Un approccio fondamentale dell'approccio con basi di dati è che fornisce un certo di livello d'astrazione dei dati nascondendo quei dettagli sulla memorizzazione degli stessi che non sono necessari alla maggior parte degli utenti.

In generale il concetto di *astrazione dei dati* fa riferimento all'eliminazione dei dettagli relativi all'organizzazione e memorizzazione dei dati, mettendo in risalto le loro caratteristiche essenziali, al fine di migliorarne la comprensione.

Un ***modello di dati*** - un insieme di concetti che possono essere usati per descrivere la struttura di una base di dati - fornisce i mezzi necessari per raggiungere questa astrazione.

Per ***struttura di una base*** di dati si intendono i tipi di dati, le associazioni e i vincoli che dovrebbero valere su di essi.

Sono stati proposti molti modelli di dati, classificabili in base ai tipi di concetti da essi utilizzati per descrivere la struttura di una base di dati:

1. ***Modelli dei dati di alto livello***: Questi modelli usano concetti come *entità, attributi* e *associazioni*. Un **entità** rappresenta un oggetto del mondo reale, come Impegato ,Progetto, che è descritto nella base di dati. Un **attributo** rappresenta una qualche proprietà di interesse che descrive più a fondo un'entità, come il nome, numero di matricola, data di nascita di un impiegato. Un'**associazione** tra due o più entità rappresenta un legame tra le entità, per esempio un'associazione tra un impiegato e un progetto su cui lavorare. Questi tipi di modelli, forniscono concetti che sono vicini al modo in cui molti utenti percepiscono i dati;

2. ***Modelli dei dati di basso livello o fisici***: foniscono concetti che descrivono i dettagli sul modo in cui i dati sono memorizzati nel calcolatore. Questi modelli sono generalmente destinati a specialisti dei sistemi di elaborazione, non a utenti finali tipici;

3. ***Modelli dei dati implementabili:*** forniscono concetti che possono essere compresi dagli utenti finali ma che non sono troppo lontani dal modo in cui i dati sono organizzati

all'interno del calcolatore: essi nascondono alcuni dettagli di memorizzazione dei dati ma si possono implementare direttamente sul calcolatore.

I modelli dei dati implementabili sono i modelli usati più frequentemente nei tradizionali DBMS commerciali, e includono il diffuso modelli dei dati Relazionale.

Il **modello relazionale** [26] dei dati (modello su cui si concentra l'attenzione di questa tesi) permette di definire tipi per mezzo del costruttore di *relazione*, che consente di organizzare i dati in insiemi di record a struttura fissa. Una relazione viene spesso rappresentata mediante una tabella in cui le righe rappresentano i specifici record e le colonne corrispondono ai campi dei record.

1.2.5 Basi di dati relazionali

Rappresenta il modello su cui si basa la maggior parte dei sistemi di basi di dati oggi sul mercato. Tale modello fu proposto in una pubblicazione scientifica nel 1970 al fine di superare le limitazioni logiche dei modelli allora utilizzati, che non permettevano di realizzare efficacemente la proprietà di indipendenza dei dati, già riconosciuta come fondamentale. Sebbene i primi prototipi di database basati sul modello relazionale risalgano ai primi anni settanta bisognerà aspettare la metà degli anni ottanta perché tale modello acquisisca una frazione significativa di mercato. La lentezza di affermazione del modello relazionale deriva

principalmente dal suo alto livello di astrazione: non è stato immediato per gli operatori del settore imparare ad individuare relazioni efficienti.

1.2.6 Modelli logici nei sistemi di basi di dati: Il modello relazionale

Il **modello relazionale** rappresenta la base di dati come una collezione di *relazioni*. Formalmente ogni relazione assomiglia ad una tabella di valori. Quando si pensa a una relazione come a una tabella di valori, ogni riga della tabella rappresenta una collezioni di dati collegati.

Di seguito un esempio di base di dati UNIVERSITA' con le relative relazioni (tabelle) [25].

IMPIEGATO

NOME_BATT	INIZ_INT	COGNOME	SSN	DATA_N	INDIRIZZO	SESSO	STIPENDIO	SUPER_SSN	N_D
John	B	Smith	123456789	1965-01-09	731 Fondren, Houston, TX	M	30000	333445555	5
Franklin	T	Wong	333445555	1955-12-08	638 Voss, Houston, TX	M	40000	888665555	5
Alicia	J	Zelaya	999887777	1968-07-19	3321 Castle, Spring, TX	F	25000	987654321	4
Jennifer	S	Wallace	987654321	1941-06-20	291 Berry, Bellaire, TX	F	43000	888665555	4
Ramesh	K	Narayan	666884444	1962-09-15	975 Fire Oak, Humble, TX	M	38000	333445555	5
Joyce	A	English	453453453	1972-07-31	5631 Rice, Houston, TX	F	25000	333445555	5
Ahmad	V	Jabbar	987987987	1969-03-29	980 Dallas, Houston, TX	M	25000	987654321	4
James	E	Borg	888665555	1937-11-10	450 Stone, Houston, TX	M	55000	NULL	1

DIPARTIMENTO

NOME_D	NUMERO_D	SSN_DIR	DATA_INIZIO_DIR
Ricerca	5	333445555	1988-05-22
Amministrazione	4	987654321	1995-01-01
Sede centrale	1	888665555	1981-06-19

SEDI_DIP

NUMERO_D	SEDE_D
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston

LAVORA_SU

SSN_I	N_P	ORE
123456789	1	32,5
123456789	2	7,5
666884444	3	40,0
453453453	1	20,0
453453453	2	20,0
333445555	2	10,0
333445555	3	10,0
333445555	10	10,0
333445555	20	10,0
999887777	30	30,0
999887777	10	10,0
987987987	10	35,0
987987987	30	5,0
987654321	30	20,0
987654321	20	15,0
888665555	20	NULL

PROGETTO

NOME_P	NUMERO_P	SEDE_P	NUM_D
ProdottoX	1	Bellaire	5
ProdottoY	2	Sugarland	5
ProdottoZ	3	Houston	5
Informatizzazione	10	Stafford	4
Riorganizzazione	20	Houston	1
Nuove opportunità	30	Stafford	4

PERSONA_A_CARICO

SSN_I	NOME_PERSONA_A_CARICO	SESSO	DATA_N	PARENTELA
333445555	Alice	F	1986-04-05	FIGLIA
333445555	Theodore	M	1983-10-25	FIGLIO
333445555	Joy	F	1958-05-03	CONIUGE
987654321	Abner	M	1942-02-28	CONIUGE
123456789	Michael	M	1988-01-04	FIGLIO
123456789	Alice	F	1988-12-30	FIGLIA
123456789	Elizabeth	F	1967-05-05	CONIUGE

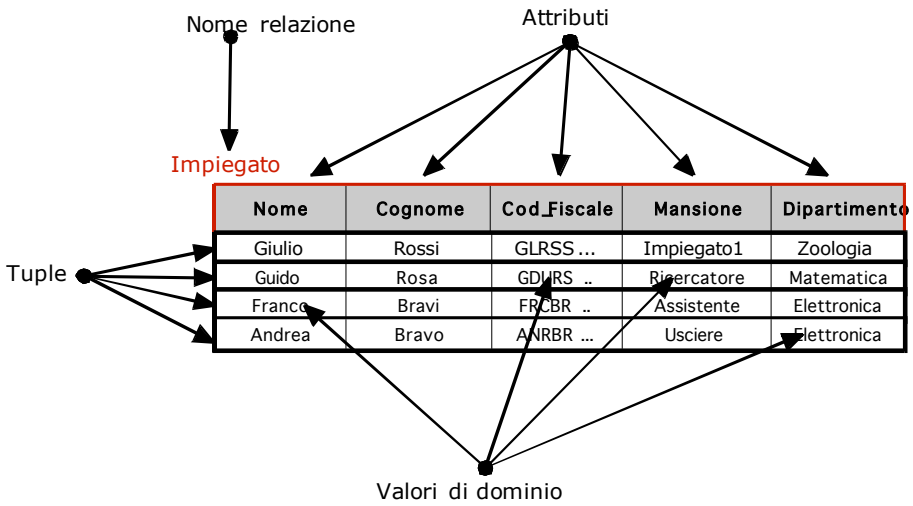
Precedentemente abbiamo visto i tipi di entità e di associazioni come concetti o modelli concettuali per modellare i dati del mondo reale. Nel

modello relazionale, ogni riga della tabella rappresenta un fatto che generalmente corrisponde a un'entità o un associazione del mondo reale.

Il nome della tabella e i nomi delle colonne sono usati per aiutare ad interpretare il significato dei valori precisi in ogni riga.

Ad esempio la prima tabella della figura di sopra è chiamata IMPIEGATO perché ogni riga rappresenta fatti relativi a una specifica entità impiegato. I nomi delle colonne – NOME_BATT, INIZ_INT, COGNOME, SSN, DATA_N, INDIRIZZO, SESSO, STIPENDIO, SUPER_SSN, N_D specificano come interpretare i valori presenti in ogni riga, sulla base della colonna in cui si trova ogni valore. Tutti i valori presenti in una colonna sono dello stesso tipo di dati.

Nella terminologia formale del modello relazionale, una riga è detta *Tupla*, un intestazione di colonna è detta *Attributo* e la tabella è detta *Relazione*. Il tipo di dati che descrive i tipi di valori che possono apparire in ogni colonna è rappresentata da un *Dominio* di possibili valori.



1.2.7 Metodologia pratica di progettazione di basi di dati

L'attività generale di progettazione di basi di dati deve seguire un processo sistematico chiamato **metodologia di progettazione**, indipendente dal fatto che la base di dati sia gestita da una RDBMS relazionale, da un sistema di base dati orientata ad oggetti (ODBMS *object database management system*) oppure da un sistema di gestione di base dati relazionale ad oggetti (ORDBMS).

In genere la progettazione di piccole basi dati, per esempio da venti utenti, non è un'attività molto complessa. Per le basi di dati di medie o grandi dimensioni che devono essere utilizzate da numerosi gruppi di utenza differenti, ciascuno con decine o centinaia di utenti, diventa invece necessario adottare un approccio sistematico all'attività di progettazione. La dimensione vera e propria di una base di dati popolata non riflette la complessità della progettazione, per la quale il fattore più importante è rappresentato dallo schema. Una base di dati con uno schema in cui vi siano più di 30 o 40 tipi di entità e un analogo numero di tipi di associazioni richiede l'uso di una metodologia di progettazione.

1.2.8 Processo di progettazione e implementazione delle basi di dati.

In questo paragrafo ci concentreremo sulla progettazione e implementazione della base di dati. Il problema di progettare una base di dati può essere formulato come segue:

Progettare la struttura logica e fisica di una o più base di dati al fine di gestire le esigenze informative degli utenti per un insieme ben definita di applicazioni.

Gli obiettivi della progettazione di base di dati sono molteplici:

- Soddisfare i requisiti sui dati degli utenti e delle applicazioni previste;
- Fornire una strutturazione delle informazioni naturale e facile da comprendere;
- Soddisfare i requisiti di elaborazione e i parametri connessi con le prestazioni, come il tempo di risposta, il tempo di elaborazione e lo spazio di memoria.

Il risultato dell'attività di progettazione è uno schema rigido, che non può essere modificato facilmente una volta che la base di dati viene implementata.

Nel processo di progettazione e implementazione di una base di dati è possibile individuare sei fasi principali:

1. Raccolta e analisi dei requisiti;
2. Progettazione concettuale della base di dati;
3. Scelta di un DBMS;

4. Mapping di modello dei dati *(detta anche progettazione logica della base di dati);*
5. Progettazione fisica della base di dati;
6. Implementazione e ottimizzazione (*tuning*) del sistema della base di dati.

Analizziamo in dettaglio le sei fasi di progettazione delle basi di dati.

1.2.9 Raccolta e analisi dei requisiti

Prima di progettare una base di dati è necessario conoscere e analizzare nel modo più dettagliato possibile le aspettative degli utenti e l'utilizzo della base di dati previsto. Questo processo viene chiamato ***raccolta e analisi dei requisiti***. Per specificare i requisiti bisogna innanzitutto identificare le parti del sistema informativo che interagiscono con la base di dati. Queste comprendono le applicazioni e gli utenti, sia nuovi sia già esistenti.

Talvolta vengono raccolte e analizzate le risposte scritte a una serie di domande poste ai potenziali utenti o gruppi di utenti della base di dati. Tali domande intendono rilevare le priorità degli utenti e l'importanza che essi attribuiscono alle varie applicazioni. Per agevolare la valutazione della validità delle informazioni e la definizione delle priorità si intervistano soltanto gli utenti principali.

L'analisi dei requisiti viene eseguita da un gruppo di analisti o esperti di requisiti . E' probabile che i requisiti iniziali siano informali,

incompleti, incoerenti e parzialmente errati. Pertanto c'è molto da fare per trasformare questi requisiti iniziali in una specifica applicativa.

Il coinvolgimento dell'utente nel processo di sviluppo aumenta senz'altro il suo livello di soddisfazione nei confronti del sistema che userà.

La fase di raccolta e analisi dei requisiti può richiedere abbastanza tempo , ma gioca un ruolo fondamentale nel successo del sistema informativo.

Un errore a livello dei requisiti è molto più costoso che correggere un errore a livello di implementazione, perché gli effetti di un errore nei requisiti sono estesi e generalizzati: interessando quanto già prodotto nelle fasi successive, tali errori danno luogo infatti alla necessità di procedere ad una nuova implementazione.

1.2.10 Progettazione concettuale della base di dati

La seconda fase della progettazione della base di dati comprende l'attività dei ***progettazione dello schema concettuale***, ossia si analizzano i requisiti risultanti dalla fase 1 e produce lo schema concettuale della base di dati.

Lo *schema concettuale* è una descrizione grafica coincisa dei requisiti sui dati degli utenti e comprende descrizioni dettagliate dei tipi di entità, delle associazioni e dei vincoli; questi sono espressi usando i concetti forniti dal modello dei dati di alto livello. Poiché non comprendono dettagli implementativi, essi sono generalmente più semplici da comprendere e possono essere usati per comunicare con tecnici e non tecnici. Infatti, la

descrizione diagrammatica dello schema concettuale può fungere da eccellente veicolo di comunicazione tra gli utenti della base di dati, i progettisti e gli analisti. Poiché i modelli dei dati ad alto livello si basano di regola su concetti più comprensibili rispetto a quelli utilizzati nei modelli dei dati di livello più basso dei DBMS, o rispetto alle definizioni sintattiche dei dati, qualsiasi comunicazione relativa alla progettazione dello schema diventa più esatta e più semplice.

Il modello più frequentemente usato nella progettazione concettuale è il **modello Entità-Associazione** (ER, *entity-relation-ship*). o in modo più generale il **modello Entità-Associazione-Estesa** (EER, *entity-enhanced-relation-ship*) che comprende tutti i concetti propri del modello ER, cui si aggiungono i concetti di *sottoclasse* e *superclasse* e i concetti collegati di *specializzazione* e *generalizzazione*.

1.2.11 Scelta del DBMS

La scelta del DBMS è dettata da molti fattori tecnici, altri economici, e altri ancora connessi con le politiche di organizzazione. I fattori tecnici attengono all'adeguamento del DBMS in relazione al compito da svolgere. I problemi da considerare sono in questo caso il tipo di DBMS (relazionale, oggetti, o altri), le strutture di memorizzazione e i percorsi di accesso che il DBMS supporta, le interfacce disponibili per gli utenti e i programmatori, i tipi di linguaggi di interrogazione di alto livello, la disponibilità di strumenti di sviluppo, la possibilità di interfacciarlo.

1.2.12 Mapping del modello dei dati - progettazione logica della base dei dati –

La fase successiva della progettazione della base di dati, consiste nella traduzione degli schemi prodotti nella fase di progettazione della fase concettuale.

In questa fase la traduzione non considera alcuna caratteristica specifica e neppure i casi speciali che si applicano all'implementazione del modello dei dati del DBMS.

Un esempio tipico è la traduzione di uno schema ER in uno schema relazionale, e quelli di uno schema EER in uno relazionale.

1.2.13 Progettazione fisica della base di dati

La progettazione fisica della base di dati riguarda il processo di scelta delle strutture di memorizzazione e di accesso ai file della base di dati al fine di garantire buone prestazioni in relazione alle varie applicazazioni.

Per la scelta delle opzioni di progettazione fisica della base di dati si seguono i seguenti criteri:

1. *Tempo di risposta:* E' il tempo che intercorre tra la richiesta di esecuzione di una transazione sulla base di dati e la ricezione di una risposta;
2. *Utilizzo dello spazio:* E' la quantità di spazio di memorizzazione usata dai file della base di dati e dale loro strutture di accesso sui dischi.

1.2.14 Implementazione e ottimizazzione del sistema di basi di dati.

Completate le attività di progettazione logica e fisica, si può passare all'implementazione della base di dati. Di solito questo compito tocca ai programmatori di basi di dati.

La basi di dati può essere poi riempita (popolata) con i file vuoti della base di dati. Se si devono convertire i dati da un sistema software precedente, può essere necessario avvalersi di procedure di conversione per riformattare I dati prima di caricarli nella nuova base di dati.

2. Scopo della tesi

Il lavoro alla base di questa tesi costituisce parte di un approccio multidisciplinare che si svolge nell'ambito di un ampio progetto di valorizzazione e miglioramento della carne di bufalo. In particolare il lavoro di questa tesi è stato volto all'utilizzo di diversi strumenti informatici che consentono da un lato l'ottenimento di nuove informazioni genomiche e biologiche sul tessuto muscolare bufalino e dall'altro una versatile catalogazione ed organizzazione del "know how" scientifico e tecnologico riguardante il bufalo al fine di migliorarne l'utilizzo e sfruttare al massimo il contenuto informazionale di ciascun dato conosciuto.

Questa tesi consta di due parti:

1) la prima parte si basa sulla individuazione "in silico", grazie agli strumenti della biologia computazionale e della bioinformatica, di nuovi marcatori genomici del muscolo bufalino che costituiscono un validissimo strumento per lo sviluppo di nuovi sistemi di analisi e monitoraggio;

2) la seconda parte è volta alla progettazione di un database relazionale dedicato alla raccolta, catalogazione e collegamento di informazioni relative al bufalo dal punto di vista biologico e della produzione. Una organizzazione razionale e snella di informazioni anche molto

distanti tra loro consente di estrarre nuove informazioni e nuovi spunti applicativi.

3. Risultati

3.1 Individuazione di marcatori genomici del muscolo bufalino

La caratterizzazione molecolare del tessuto muscolare bufalino rappresenta una valida strategia per ottenere strumenti per formulare nuovi approcci tecnologici utili sia nel management di allevamento che per avviare programmi di miglioramento genetico.

La caratterizzazione genomica può realizzarsi seguendo approcci su diversi livelli: un approccio su larga scala che parte dal sequenziamento e caratterizzazione dei cloni di una library a cDNA tessuto specifica per la caratterizzazione di tutti i geni tessuto specifici ed un approccio su singolo gene volto al clonaggio di geni specifici scelti sulla base del loro interesse nella struttura e funzione del gene di interesse.

In questa parte dei risultati della tesi viene riportato il lavoro bioinformatico svolto per ottenere cloni di specifici trascritti muscolo specifici. In particolare si è giunti alla identificazione di regioni di sequenza nucleotidica su cui disegnare coppie di oligonucleotidi (un primer forward ed uno reverse) da utilizzare come primer d'innesco nella reazione di PCR.

Tali reazioni si svolgono poi usando come stampo o pool di cDNA prodotti a partire da RNA muscolare oppure DNA genomico.

Come base di partenza per tale lavoro si è proceduto nella scelta dei geni di cui interessarsi. Volendo arrivare ad una caratterizzazione funzionale e strutturale del muscolo per ottenere parametri validi per

avviare e monitorare successivi approcci sperimentali e tecnologici, si è cercato di individuare geni chiave di diversi aspetti della biologia del muscolo.

In particolare sono stati scelti geni coinvolti nello sviluppo del muscolo, nella struttura del sarcomero e nel metabolismo.

Gene	Funzione	Tessuto
MyoD1	differenziamento	Muscolo specifico
Myostatin	differenziamento	Muscolo specifico
Myogenin	differenziamento	Muscolo specifico
Myf5	differenziamento	Muscolo specifico
TropomiosinaTMP3	Struttura sarcomero	Muscolo specifico
Miosina catena pesante	Struttura sarcomero	Non Muscolo specifico
Miosina catena leggera	Struttura sarcomero	Non Muscolo specifico
Troponina T	Struttura sarcomero	Muscolo specifico
MLCK	metabolismo	ubiquitaria
Mioglobina	metabolismo	Muscolo specifico

Poiché le informazioni genomiche su bufalo sono ancora molto limitate, un secondo livello di scelta ha riguardato la individuazione delle specie da usare come riferimento nella selezione delle sequenze da utilizzare per mettere in atto una strategia di genomica comparata **[9,10]**; ovvero per ottenere le sequenze da usare come base per la comparazione e quindi per la definizione di una sequenza consenso su cui andare a disegnare i primer per le amplificazioni.

Considerando la filogenesi dei mammiferi e tenendo conto della disponibilità di materiale genomico esistente nelle banche dati per le diverse specie, si è scelto di prendere in esame principalmente le

informazioni genomiche esistenti per il bovino e poi di confrontarle con le altre note prendendo poi prioritariamente in considerazione quelle di uomo, topo, pecora, suino.

Per ottenere le sequenze dei geni di interesse è stato utilizzato il database pubblico NCBI (National Center for Biotechnology Information, (http://www.ncbi.nlm.nih.gov/), in particolare mediante i *tool* Entrez database e Nucleotide database. Interrogando i database mediante tali strumenti si può innanzitutto verificare l'esistenza di una sequenza già nota per il gene di interesse nella specie bufalina e una volta verificatane l'assenza si cerca di ottenere la sequenza relativa a quel gene in una specie vicina, i.e. bovino. Una volta ottenuta tale sequenza è possibile rapidamente confrontarla con le altre esistenti in rete grazie all'algoritmo BLAST (*Basic Local Alignment Search Tool*). Grazie a questo strumento è possibile andare a vedere se la sequenza di interesse esiste anche in altre specie, ed eventualmente con quale grado di similarità e a quanta parte della sequenza si estende tale similarità. Il numero dei geni per i quali abbiamo potuto attuare questo tipo di approccio è rimasto abbastanza limitato.

Durante la stesura della tesi si è registrato un notevole aumento delle immissioni per molte specie.

Una volta ottenute le sequenze relative alle diverse specie, sono state allineate mediante ClustalW, che è un algoritmo dell'allineamento progressivo mediante profili.

In questo modo grazie all'allineamento è possibile estrarre una sequenza consenso ovvero un'unica sequenza finale che tenga conto del

contributo di tutte le altre. Su tale sequenza è possibile individuare delle zone di alta conservazione, ovvero delle zone in cui la sequenza è identica, o molto poco variabile, in tutte le specie considerate. In linea di principio le sequenze che forniscono il massimo livello di similarità sono le sequenze codificanti, in quanto sono quelle che subiscono la maggiore pressione di conservazione in quanto la loro alterazione porta ad alterazione nella sequenza della proteina codificata. Sulla sequenza consenso si avvia l'analisi per il disegno dei primer d'innesco per la reazione di PCR. In assenza di regioni conservate o in presenza di regioni troppo piccole si può adottare la strategia dei primer degenerati che spiegheremo operativamente in dettaglio nella descrizione del lavoro svolto per i singoli geni.

Mediamente la lunghezza dei primer viene fissata tra 18 e 22 basi in quanto in questo intervallo si ottiene un buon compromesso tra specificità ed efficienza di reazione **[27]**.

Per disegnare i primer ci si può avvalere di vari software, tra cui alcuni disponibili liberamente in rete. Per questo lavoro ci siamo avvalsi del programma Primer3 (http://primer3.sourceforge.net/). Tale programma consente di individuare coppie di primer, compatibili tra loro, con lunghezza, dimensioni del prodotto, temperature di *annealing* e caratteristiche termodinamiche e steriche opportune. Nel nostro caso essendo i vincoli di scelta delle regioni da amplificare molto stretti (solo le zone più "conservate") il programma è stato utilizzato soprattutto per verificare a posteriori la qualità dei primer scelti piuttosto che per determinarne la scelta.

Per ogni gene, laddove possibile, si cerca di disegnare più coppie di primer sia per aumentare la probabilità di riuscita del clonaggio che per coprire tutta la lunghezza conosciuta del gene.

I primer disegnati vengono ulteriormente verificati mediante l'algoritmo Blast. Con tale strumento infatti, anche se solo parzialmente, si può verificare se queste brevi sequenze di DNA possono riconoscere altre regioni genomiche e quindi creare interferenze nella reazione di PCR.

Questo tipo di analisi è particolarmente importante per evitare primer che possono amplificare regioni altamente o mediamente ripetute del genoma. Una volta ottenuto il frammento desiderato mediante PCR su DNA genomico o cDNA, questo viene clonato e sequenziato. La sequenza ottenuta viene "ripulita" dalle parti di sequenza proprie del vettore di clonaggio e viene risottoposta ad analisi con Blast per verificare che corrisponda alla sequenza richiesta e per valutare il grado di similarità con le sequenze dello stesso gene e delle altre specie presenti in banca dati.

Sulla stessa sequenza così ottenuta e che è certamente bufalo specifica si possono ora produrre nuovi primer per iniziare il cosiddetto *primer-walking* per ottenere l'intero trascritto o l'intero locus genomico.

Qui di seguito riportiamo in dettaglio il lavoro svolto per il gene MyoD1.

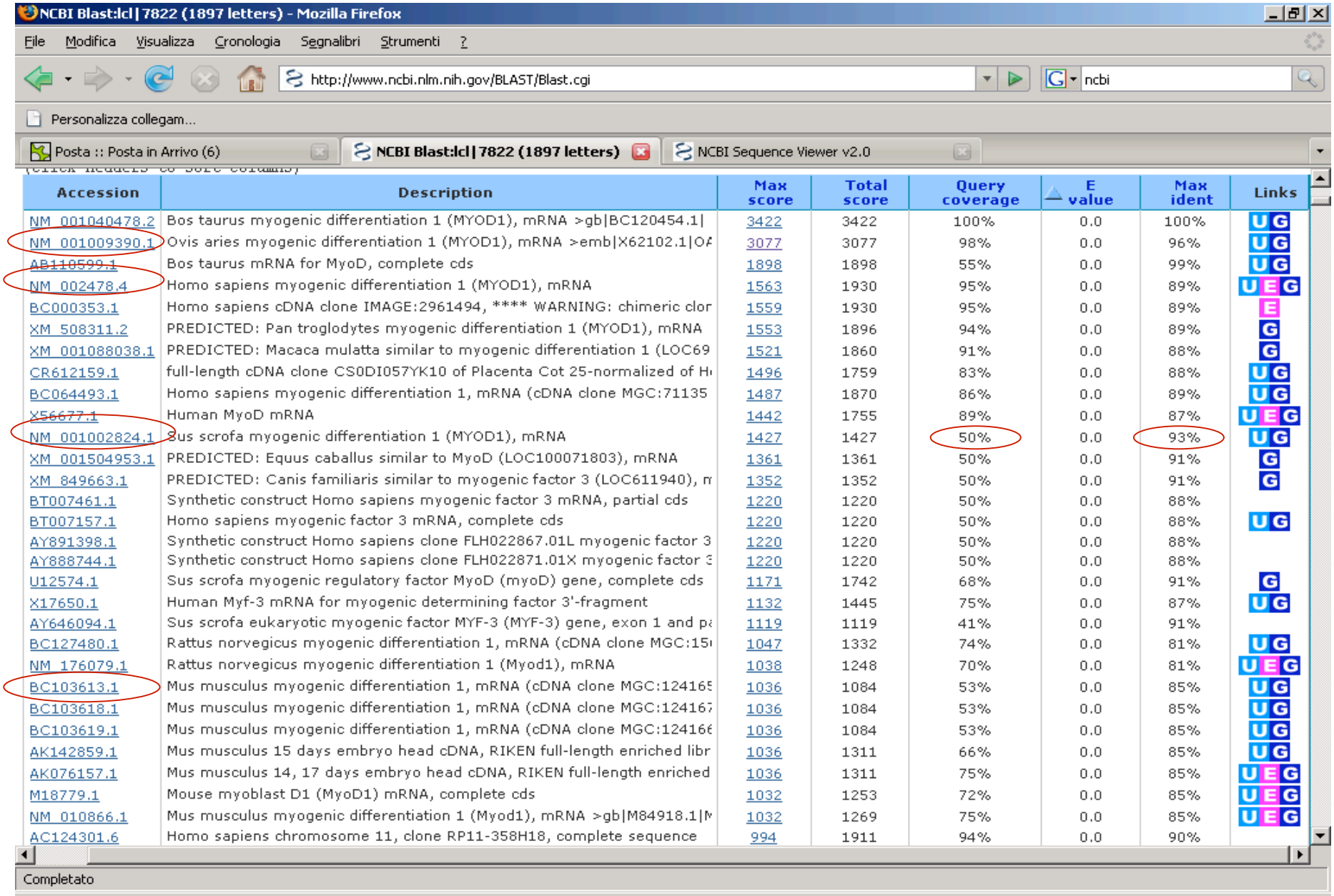
MyoD1

Il gene MyoD è considerato il master gene dello sviluppo muscolare **[2]**. La proteina codificata da tale gene (318 aa) è un regolatore

trascrizionale che appartiene alla famiglia di proteine basic helix-loop-helix (bHLH). Tale proteina, regolando la trascrizione di altri geni codificanti per regolatori della trascrizione muscolare e geni muscolo-specifici, innesca il processo di differenziamento cellulare.

La sequenza di Bos taurus utilizzata come riferimento è la BC120454 di 1897 bp mRNA.

Tale sequenza è stata immessa in Blast, ottenendo il seguente schema relativo alle migliori percentuali di similarità.



Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_001040478.2	Bos taurus myogenic differentiation 1 (MYOD1), mRNA >gb|BC120454.1|	3422	3422	100%	0.0	100%	U G
NM_001009390.1	Ovis aries myogenic differentiation 1 (MYOD1), mRNA >emb|X62102.1|OA	3077	3077	98%	0.0	96%	U G
AB110599.1	Bos taurus mRNA for MyoD, complete cds	1898	1898	55%	0.0	99%	U G
NM_002478.4	Homo sapiens myogenic differentiation 1 (MYOD1), mRNA	1563	1930	95%	0.0	89%	U E G
BC000353.1	Homo sapiens cDNA clone IMAGE:2961494, **** WARNING: chimeric clor	1559	1930	95%	0.0	89%	E
XM_508311.2	PREDICTED: Pan troglodytes myogenic differentiation 1 (MYOD1), mRNA	1553	1896	94%	0.0	89%	G
XM_001088038.1	PREDICTED: Macaca mulatta similar to myogenic differentiation 1 (LOC69	1521	1860	91%	0.0	88%	G
CR612159.1	full-length cDNA clone CS0DI057YK10 of Placenta Cot 25-normalized of H	1496	1759	83%	0.0	88%	U G
BC064493.1	Homo sapiens myogenic differentiation 1, mRNA (cDNA clone MGC:71135	1487	1870	86%	0.0	89%	U G
X56677.1	Human MyoD mRNA	1442	1755	89%	0.0	87%	U E G
NM_001002824.1	Sus scrofa myogenic differentiation 1 (MYOD1), mRNA	1427	1427	50%	0.0	93%	U G
XM_001504953.1	PREDICTED: Equus caballus similar to MyoD (LOC100071803), mRNA	1361	1361	50%	0.0	91%	G
XM_849663.1	PREDICTED: Canis familiaris similar to myogenic factor 3 (LOC611940), m	1352	1352	50%	0.0	91%	G
BT007461.1	Synthetic construct Homo sapiens myogenic factor 3 mRNA, partial cds	1220	1220	50%	0.0	88%	
BT007157.1	Homo sapiens myogenic factor 3 mRNA, complete cds	1220	1220	50%	0.0	88%	U G
AY891398.1	Synthetic construct Homo sapiens clone FLH022867.01L myogenic factor 3	1220	1220	50%	0.0	88%	
AY888744.1	Synthetic construct Homo sapiens clone FLH022871.01X myogenic factor 3	1220	1220	50%	0.0	88%	
U12574.1	Sus scrofa myogenic regulatory factor MyoD (myoD) gene, complete cds	1171	1742	68%	0.0	91%	G
X17650.1	Human Myf-3 mRNA for myogenic determining factor 3'-fragment	1132	1445	75%	0.0	87%	U G
AY646094.1	Sus scrofa eukaryotic myogenic factor MYF-3 (MYF-3) gene, exon 1 and pa	1119	1119	41%	0.0	91%	
BC127480.1	Rattus norvegicus myogenic differentiation 1, mRNA (cDNA clone MGC:15	1047	1332	74%	0.0	81%	U G
NM_176079.1	Rattus norvegicus myogenic differentiation 1 (Myod1), mRNA	1038	1248	70%	0.0	81%	U E G
BC103613.1	Mus musculus myogenic differentiation 1, mRNA (cDNA clone MGC:12416	1036	1084	53%	0.0	85%	U G
BC103618.1	Mus musculus myogenic differentiation 1, mRNA (cDNA clone MGC:12416	1036	1084	53%	0.0	85%	U G
BC103619.1	Mus musculus myogenic differentiation 1, mRNA (cDNA clone MGC:12416	1036	1084	53%	0.0	85%	U G
AK142859.1	Mus musculus 15 days embryo head cDNA, RIKEN full-length enriched libr	1036	1311	66%	0.0	85%	U G
AK076157.1	Mus musculus 14, 17 days embryo head cDNA, RIKEN full-length enriched	1036	1311	75%	0.0	85%	U E G
M18779.1	Mouse myoblast D1 (MyoD1) mRNA, complete cds	1032	1253	72%	0.0	85%	U E G
NM_010866.1	Mus musculus myogenic differentiation 1 (Myod1), mRNA >gb|M84918.1|M	1032	1269	75%	0.0	85%	U E G
AC124301.6	Homo sapiens chromosome 11, clone RP11-358H18, complete sequence	994	1911	94%	0.0	90%	

Sulla base di alcuni parametri principali (max-score, query coverage; max identity), le sequenze più informative sono risultate essere: O.aries (REF:NM_001009390.1), Homo sapiens (REF:NM_002478.4), Mus musculus (gb|BC103613.1), Sus Scrofa (ref|NM_001002824.1). Queste ultime pur avendo un basso query coverage, mostrano comunque un alta percentuale di similarità (max identity). Qui di seguito riportiamo per brevità il risultato di Blast per la sola sequenza O.aries (REF:NM_001009390.1):

```
REF:NM_001009390.1
GENE ID: 443405 MYOD1 | myogenic differentiation 1 [Ovis aries]
Score = 3077 bits (3412),  Expect = 0.0
Identities = 1819/1893 (96%), Gaps = 22/1893 (1%)
Strand=Plus/Plus

Query  1     GGAGGCCCCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGGGACAGGAC  60
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  129   GGAGGCCTCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGGGACAGGAC  188

Query  61    CGGGGAGGGTGGGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCTATCTCACT  120
             |||||||||| |||||||||||||||||||||||||||||||||||| ||||| |
Sbjct  189   CGGGGAGGGTGAGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCTCTCTCATT  248

Query  121   GCTGGGGCTCCAGAACAGCAGCAAGTTTCTGGCAACCCTTGCCGCTGCCGCTGAGGCTGG  180
             || |||||||||||| |||| |||||||||||||| ||||||||||||||||||||||
Sbjct  249   GCCGGGGCTCCAGAACCGCAGTAAGTTTCTGGCAACCCCTGCCGCTGCCGCTGAGGCTGG  308

Query  181   GCAAAGCCAGGATCGCGCCGCCCCCCGCCGGGATATGGAGCTGCTGTCGCCGCCGCTCCG  240
             |||||||||||| |||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  309   GCAAAGCCAGGACCGCGCCGCCCCACGCCGGGATATGGAGCTGCTGTCGCCGCCGCTCCG  368

Query  241   CGACGTAGACTTGACGGGCCCCGACGGCTCTCTCTGCAACTTTGCAACAGCGGACGACTT  300
             |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  369   CGACGTAGACTTGACGGGCCCCGACGGCTCTCTCTGCAACTTTGCAACAGCGGATGACTT  428

Query  301   CTATGATGACCCGTGTTTCGACTCCCCGGACCTGCGCTTCTTCGAGGACCTGGATCCGCG  360
             |||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  429   CTATGATGACCCGTGTTTCGACTCCCCGGATCTGCGCTTCTTCGAGGACCTGGATCCGCG  488

Query  361   CCTCGTGCACGTGGGCGCGCTCCTGAAGCCCGAGGAACACTCGCACTTCCCTGCAGCCGC  420
             |||||||||||||||||||||||||||| |||||||||||||||||||||||||||| ||
Sbjct  489   CCTCGTGCACGTGGGCGCGCTCCTGAAGCCGGAGGAACACTCGCACTTCCCTGCAGCGGC  548

Query  421   GCACCCGGCCCCGGGCGCGCGCGAGGACGAGCATGTGCGCGCGCCCAGCGGGCACCACCA  480
             |||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||
Sbjct  549   GCACCCGGCCCCGGGCGCGCGCGAGGACGAACATGTGCGCGCGCCCAGCGGGCACCACCA  608

Query  481   GGCGGGCCGCTGTTTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGACTAACGCCGA  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  609   GGCGGGCCGCTGTTTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGACTAACGCCGA  668

Query  541   CCGCCGCAAGGCTGCTACCATGCGCGAGCGGCGCCGCCTGAGCAAAGTCAACGAGGCCTT  600
             |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||
Sbjct  669   CCGCCGCAAGGCTGCTACCATGCGCGAGCGGCGCCGCTTGAGCAAAGTCAACGAGGCCTT  728

Query  601   CGAGACGCTCAAACGCTGCACGTCTAGCAACCCAAACCAGCGGCTGCCCAAGGTGGAGAT  660
```

```
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  729   CGAGACGCTCAAACGCTGCACGTCTAGCAACCCAAACCAGCGGCTGCCCAAGGTGGAGAT  788

Query  661   CCTGCGCAACGCCATCCGCTATATCGAAGGCCTGCAGGCGCTACTTCGCGACCAGGACGC  720
              ||||||||||||| |||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  789   CCTGCGCAACGCAATCCGCTATATCGAAGGCCTGCAGGCGCTGCTTCGCGACCAGGACGC  848

Query  721   CGCGCCTCCCGGCGCTGCCGCTGCCTTTTACGCGCCTGGCCCGTTGCCCCCCGGCCGCAG  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  849   CGCGCCTCCCGGCGCTGCCGCTGCCTTTTACGCGCCTGGCCCGTTGCCCCCCGGCCGCAG  908

Query  781   CGGCGAACACTACAGCGGCGACTCGGACGCTTCCAGTCCGCGCTCCAACTGTTCCGACGG  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  909   CGGCGAACACTACAGCGGCGACTCGGACGCTTCCAGTCCGCGCTCCAACTGTTCCGACGG  968

Query  841   CATGATGGACTACAGCGGCCCCCCGAGTGGTGCCCGGCGGCGGAACTGCTACGACCGCAC  900
              |||||||||||||||||||||||||||||||||||||| |||||||||||||||||| |
Sbjct  969   CATGATGGACTACAGCGGCCCCCCGAGTGGTGCCCGGCGACGGAACTGCTACGACCGCGC  1028

Query  901   TTACTACAGCGAGGCGCCCAACGAACCCCGGCCCGGGAAGAGCGCTGCGGTGTCGAGCCT  960
              |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  1029  TTACTACAGCGAGGCGCCCAATGAACCCCGGCCCGGGAAGAGCGCTGCGGTGTCGAGCCT  1088

Query  961   CGACTGCCTGTCCAGCATCGTGGAGCGCATCTCCACCGAGAGCCCCGCCGCGCCCGCGCT  1020
              |||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  1089  CGACTGCCTGTCCAGCATCGTGGAGCGCATCTCCACCGAGAGCCCCGCAGCGCCCGCGCT  1148

Query  1021  TCTGCTAGCCGACGCGCCGCCGGAGTCGTCTCCTGGCCCGCAGGA---GGCCGCCGGGAG  1077
              ||| || |||||||||||||||||||||||| ||||||||||   ||||||||||||
Sbjct  1149  TCTACTGGCCGACGCGCCGCCGGAGTCGTCTCCCGGCCCGCAGGAGGCGGCCGCCGGGAG  1208

Query  1078  CGAGGTGGAGCGCGGCACCCCCGCTCCTTCCCCCGACACTGCCCCTCAGGGCCTCGCGGG  1137
              |||||||||| |||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  1209  CGAGGTGGAGTGCGGCACCCCCGCCCCTTCCCCCGACACTGCCCCTCAGGGCCTCGCGGG  1268

Query  1138  CGCGAACCCCAACCCGATTTACCAGGTGCTCTGAGGGGTTGTGCGGCCTCATCGGGGGGC  1197
              |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  1269  CGCGAACCCCAACCCGATTTACCAGGTGCTCTGAGGGGTTGGGCGGCCTCATCGGGGGGC  1328

Query  1198  GCCGCTGCCCACGGGCCCGAGGGATGGCGCCCTCAGGGTCCCTCGCGCCC-AAAGATTGC  1256
              |||||||||||| |||||||||||||| ||||  |||||||||||||||| |||||||||
Sbjct  1329  GCCGCTGCCCACAGGCCCGAGGGATGGTGCCCCTAGGGTCCCTCGCGCCCAAAAGATTGC  1388

Query  1257  GCTTAAGTGCCAACCACTCTCCTCCCAACAGCGCTTTAAAAGCGACCCTCCCGAGGTAGG  1316
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1389  GCTTAAGTGCCAACCACTCTCCTCCCAACAGCGCTTTAAAAGCGACCCTCCCGAGGTAGG  1448

Query  1317  AGAGGCGAAGGAACTGTTGTGTTTCCGCCTCCCGCACCCCAGGGCAAGGACATGGTCCtt  1376
              |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1449  AGAGGCGAGGGAACTGTTGTGTTTCCGCCTCCCGCACCCCAGGGCAAGGACATGGTCCTT  1508

Query  1377  --------ttttt CCCG---------- CCTACTCAGCGCTCTTCCCTGAGACCCGCTGTGG  1419
                       ||||         || ||||||||||||||||||||||||||||||||
Sbjct  1509  CCCCCCCACCACCCCCGCCCCCCCCTCCCCACTCAGCGCTCTTCCCTGAGACCCGCTGTGG  1568

Query  1420  CGGCCGTTTGATGTTACTCCGTGGGCCAGAGCTGACCCTTGAGGAGCCAGGCCCTTCCTC  1479
               ||||  |||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  1569  TGGCCACTTGATGTTACTCCGTGGGCCAGAGCTGACCCTTGAAGAGCCAGGCCCTTCCTC  1628

Query  1480  TTCCTCGCCCCCTCCGCCATGGGGGTGTGTGACCCACGCAGGCCTAAGCCCCGCCCCCAA  1539
              ||||||||||||||| |||||| || ||| | ||||||||||||||||||||||||||||||
Sbjct  1629  TTCCTCGCCCCCTCCCCCATGGCGGGGTGGGGCCCACGCAGGCCTAAGCCCCGCCCCCAA  1688

Query  1540  GACCCCCGACCGCTTAAAGGGGCGTGCCCCTCCTTTCCGGAGCCCCCTGGGGACTTCAGC  1599
              |||||||||||||||| |||||||||||||| |||||||||||||||| |||||||||
Sbjct  1689  GACCCCCGACCGCTTACAGGGGCGTGCCCCTCCGTTCCGGAGCCCCCTGGTGACTTCAGC  1748

Query  1600  TGTTTCCCGCCGCTCCCTTCCCAAGTG-TAACAGGTGTAATGGTAACCACTcccacccccc  1658
              |||||||||||||||||||||||||| |||||||||||| ||||||||||||||||||||
Sbjct  1749  TGTTTCCCGCCGCTCCCTTCCCAAGTGTTAACAGGTGTAACGGTAACCACTCCCACCCCC  1808
```

```
Query  1659  caaccccacccccGGTTCAGGACCACTTTTTGTAATACTTTTGTAATCTATTCCTGTAA  1718
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1809  CAACCCCCACCCCCGGTTCAGGACCACTTTTTGTAATACTTTTGTAATCTATTCCTGTAA  1868

Query  1719  ATAAGAGTTGCTTTGCCAGAGCAGGAGCCCCTCGGGCTGTATTTATCTCTGAGGCATGGT  1778
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1869  ATAAGAGTTGCTTTGCCAGAGCAGGAGCCCCTCGGGCTGTATTTATCTCTGAGGCATGGT  1928

Query  1779  GTGCGGTGCGACAGGGACTTTGTATGTTTATACGGCAGGCAGGCGAGCCGCGGGCGCTCG  1838
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1929  GTGCGGTGCGACAGGGACTTTGTATGTTTATACGGCAGGCAGGCGAGCCGCGGGCGCTCG  1988

Query  1839  CTCAGGTGTTCGAAATAAAGACGCTAATTTATA  1871
             |||||||||||||||||||||||||||||||||
Sbjct  1989  CTCAGGTGTTCGAAATAAAGACGCTAATTTATA  2021
```

Le sequenze così identificate sono state poi allineate mediante ClustalW per individuare le zone di maggiore conservazione.

```
CLUSTAL W (1.83) multiple sequence alignment


Bos             ------------------------------------------------------------
Ovis            AAGCTCCTCCCTGCTCTGTTCCTATTGGCCTCGGGCGCCCCCGCCCCTAGCCGCTAGCTG 60
Homo            ------------------------------------------------------------
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------


Bos             ------------------------------------------------------------
Ovis            GGGCTCGGGGGCCCTTAGGCTACTACGGGATAAATAGCCCTGGGAGCCTGGTGTGAAGGT 120
Homo            ----------------------------------------------------GAGAAGCT 8
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------


Bos             --------GGAGGCCCCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGG 52
Ovis            AGGGGTGGGGAGGCCTCAGGGCGCTGCCGCGGCTCTCCTCGCCGTCCGTCCCTCAGGCGG 180
Homo            AGGGGTGAGGAAGCCCTGGGGCGCTGCCGCCGCTTTCCTTAAC---CACAAATCAGGCCG 65
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------


Bos             GACAGGACCGGGGAGGGTGGGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCT 112
Ovis            GACAGGACCGGGGAGGGTGAGGAACAGCTGGTTATACATTCAGACCCTCAGTGCTTTGCT 240
Homo            GACAGGAGAGGGAGGGGTGGGGGACAG-TGGGTGGGCATTCAGACTGCCAGCACTTTGCT 124
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------


Bos             ATCTCACTGCTGGGGCTCCAGAACAGCAGCAAGTTTCTGGCAACCCTTGCCGCTGCCGCT 172
Ovis            CTCTCATTGCCGGGGCTCCAGAACCGCAGTAAGTTTCTGGCAACCCCTGCCGCTGCCGCT 300
Homo            ATCT-ACAGCCGGGGCTCCCGAGCGGCAGAAAGTTCC-GGCCACTCT-----CTGCCGCT 177
Sus             ------------------------------------------------------------
Mus             --------------------GAGTGGCAGAAAGTTAA-GACGACTCTCACGGCTTGGGTT 39


Bos             GAGGCTGGGCAAAGCCAGGATCGCGCCGCCCCCC-GCCGGGATATGGAGCTGCTGTCGCC 231
Ovis            GAGGCTGGGCAAAGCCAGGACCGCGCCGCCCCAC-GCCGGGATATGGAGCTGCTGTCGCC 359
Homo            TGGGTTGGGCGAAGCCAGGACCGTGCCGCGCCACCGCCAGGATATGGAGCTACTGTCGCC 237
Sus             -------------------------------------------ATGGAGCTGCTGTCGCC 17
Mus             GAGGCTGGAC----CCAGGA---------------ACTGGGATATGGAGCTTCTATCGCC 80
                                                    ******** ** *****
```

```
Bos             GCCGCTCCGCGACGTAGACTTGACGGGCCCCGACGGCTCTCTCTGCAACTTTGCAACAGC 291
Ovis            GCCGCTCCGCGACGTAGACTTGACGGGCCCCGACGGCTCTCTCTGCAACTTTGCAACAGC 419
Homo            ACCGCTCCGCGACGTAGACCTGACGGCCCCCGACGGCTCTCTCTGCTCCTTTGCCACAAC 297
Sus             ACCGCTCCGCGACGTAGATTTGACGGGCCCCGACGGCTCTCTCTGCAACTTTGCAACAGC 77
Mus             GCCACTCCGGGACATAGACTTGACAGGCCCCGACGGCTCTCTCTGCTCCTTTGAGACAGC 140
                 ** ***** *** ****  **** * ******************  *****  *** *

Bos             GGACGACTTCTATGATGACCCGTGTTTCGACTCCCCGGACCTGCGCTTCTTCGAGGACCT 351
Ovis            GGATGACTTCTATGATGACCCGTGTTTCGACTCCCCGGATCTGCGCTTCTTCGAGGACCT 479
Homo            GGACGACTTCTATGATGACCCGTGTTTCGACTCCCCGGACCTGCGCTTCTTCGAAGACCT 357
Sus             GGACGACTTCTATGATGACCCGTGTTTCGACTCCCCGGACCTGCGCTTCTTCGAGGACCT 137
Mus             AGACGACTTCTATGATGACCCGTGTTTCGACTCACCAGACCTGCGCTTTTTTGAGGACCT 200
                 ** ****************************** ** ** ******** ** ** *****

Bos             GGATCCGCGCCTCGTGCACGTGGGCGCGCTCCTGAAGCCCGAGGAACACTCGCACTTCCC 411
Ovis            GGATCCGCGCCTCGTGCACGTGGGCGCGCTCCTGAAGCCGGAGGAACACTCGCACTTCCC 539
Homo            GGACCCGCGCCTGATGCACGTGGGCGCGCTCCTGAAACCCGAAGAGCACTCGCACTTCCC 417
Sus             GGACCCGCGCCTCGTGCACGTGGGCGCGCTCCTAAAGCCCGAGGAACACTCGCACTTCCC 197
Mus             GGACCCGCGCCTGGTGCACATGGGAGCCCTCCTGAAACCGGAGGAGCACGCACACTTCCC 260
                *** ********  ***** **** ** ***** ** ** ** ** *** * ********

Bos             TGCAGCCGCGCACCCGGCCCCGGGCGCGCGCGAGGACGAGCATGTGCGCGCGCCCAGCGG 471
Ovis            TGCAGCGGCGCACCCGGCCCCGGGCGCGCGCGAGGACGAACATGTGCGCGCGCCCAGCGG 599
Homo            CGCGGCGGTGCACCCGGCCCCGGGCGCACGTGAGGACGAGCATGTGCGCGCGCCCAGCGG 477
Sus             TGCCGCAGCGCACCCGGCCCCGGGAGCTCGTGAGGACGAGCATGTGCGCGCGCCCAGCGG 257
Mus             TACTGCGGTGCACCCAGGCCCAGGCGCTCGTGAGGATGAGCATGTGCGCGCGCCCAGCGG 320
                  * ** * ****** * *** ** ** ** ***** ** ********************

Bos             GCACCACCAGGCGGGCCGCTGTTTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGAC 531
Ovis            GCACCACCAGGCGGGCCGCTGTTTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGAC 659
Homo            GCACCACCAGGCGGGCCGCTGCCTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGAC 537
Sus             GCACCACCAGGCGGGCCGCTGTCTACTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGAC 317
Mus             GCACCACCAGGCGGGTCGCTGCTTGCTGTGGGCCTGCAAGGCGTGCAAACGCAAGACGAC 380
                *************** *****  * ***********************************

Bos             TAACGCCGACCGCCGCAAGGCTGCTACCATGCGCGAGCGGCGCCGCCTGAGCAAAGTCAA 591
Ovis            TAACGCCGACCGCCGCAAGGCTGCTACCATGCGCGAGCGGCGCCGCTTGAGCAAAGTCAA 719
Homo            TAACGCCGACCGCCGCAAGGCCGCCACCATGCGCGAGCGGCGCCGCCTGAGCAAAGTAAA 597
Sus             TAACGCCGACCGCCGCAAGGCCGCCACCATGCGCGAGCGGCGCCGCTTGAGCAAAGTCAA 377
Mus             TAACGCTGATCGCCGCAAGGCCGCCACCATGCGCGAGCGCCGCCGCCTGAGCAAAGTGAA 440
                ****** ** *********** ** ************** ****** ********** **

Bos             CGAGGCCTTCGAGACGCTCAAACGCTGCACGTCTAGCAACCCAAACCAGCGGCTGCCCAA 651
Ovis            CGAGGCCTTCGAGACGCTCAAACGCTGCACGTCTAGCAACCCAAACCAGCGGCTGCCCAA 779
Homo            TGAGGCCTTTGAGACACTCAAGCGCTGCACGTCGAGCAATCCAAACCAGCGGTTGCCCAA 657
Sus             CGAGGCCTTCGAGACTCTCAAGCGCTGCACGTCTAGCAACCCGAATCAGCGGCTGCCCAA 437
Mus             TGAGGCCTTCGAGACGCTCAAGCGCTGCACGTCCAGCAACCCGAACCAGCGGCTACCCAA 500
                 ******** ***** ***** *********** ***** ** ** ****** * *****

Bos             GGTGGAGATCCTGCGCAACGCCATCCGCTATATCGAAGGCCTGCAGGCGCTACTTCGCGA 711
Ovis            GGTGGAGATCCTGCGCAACGCAATCCGCTATATCGAAGGCCTGCAGGCGCTGCTTCGCGA 839
Homo            GGTGGAGATCCTGCGCAACGCCATCCGCTATATCGAGGGCCTGCAGGCTCTGCTGCGCGA 717
Sus             GGTGGAAATCCTGCGCAACGCCATCCGCTATATCGAAGGCCTGCAGGCGCTGCTTCGCGA 497
Mus             GGTGGAGATCCTGCGCAACGCCATCCGCTACATCGAAGGTCTGCAGGCTCTGCTGCGCGA 560
                ****** ************** ******** ***** ** ******** ** ** *****

Bos             CCAGGACGCCGCGCCTCCCGGCGCTGCCGCTGCCTTTTACGCGCCTGGCCCGTTGCCCCC 771
Ovis            CCAGGACGCCGCGCCTCCCGGCGCTGCCGCTGCCTTTTACGCGCCTGGCCCGTTGCCCCC 899
Homo            CCAGGACGCCGCGCCCCCTGGCGCCGCAGCCGCCTTCTATGCGCCGGGCCCGCTGCCCCC 777
Sus             CCAGGACGCCGCCCCCCCCTGGCGCTGCAGCGGCCTTTTACGCGCCTGGCCCGCTGCCCCC 557
Mus             CCAGGACGCCGCGCCCCCTGGCGCCG---CTGCCTTCTACGCACCTGGACCGCTGCCCCC 617
                ************ ** ** ***** *   * ***** ** ** ** ** *** *******

Bos             CGGCCGCAGCGGCGAACACTACAGCGGCGACTCGGACGCTTCCAGTCCGCGCTCCAACTG 831
Ovis            CGGCCGCAGCGGCGAACACTACAGCGGCGACTCGGACGCTTCCAGTCCGCGCTCCAACTG 959
Homo            GGGCCGCGGCGGCGAGCACTACAGCGGCGACTCCGACGCGTCCAGCCCGCGCTCCAACTG 837
Sus             GGGCCGAGGCGGAGAGCACTACAGCGGTGACTCAGACGCATCCAGCCCGCGCTCCAACTG 617
Mus             AGGCCGTGGCAGCGAGCACTACAGTGGCGACTCAGATGCATCCAGCCCGCGCTCCAACTG 677
                 *****  ** * ** ******** ** ***** ** ** ***** **************

Bos             TTCCGACGGCATGATGGACTACAGCGGCCCCCCGAGTGGTGCCCGGCGGCGGAACTGCTA 891
```

```
Ovis            TTCCGACGGCATGATGGACTACAGCGGCCCCCCGAGTGGTGCCCGGCGACGGAACTGCTA 1019
Homo            CTCCGACGGCATGATGGACTACAGCGGCCCCCCGAGCGGCGCCCGGCGGCGGAACTGCTA 897
Sus             TTCCGACGGCATGATGGATTATAGCGGCCCCCCGAGCGGTGCCCGGCGGCGGAACTGCTA 677
Mus             CTCTGATGGCATGATGGATTACAGCGGCCCCCCAAGCGGCCCCCGGCGGCAGAATGGCTA 737
                 ** ** ********** ** ********** ** **  ******* * ***  ****

Bos             CGACCGCACTTACTACAGCGAGGCGCCCAACGAACCCCGGCCCGGGAAGAGCGCTGCGGT 951
Ovis            CGACCGCGCTTACTACAGCGAGGCGCCCAATGAACCCCGGCCCGGGAAGAGCGCTGCGGT 1079
Homo            CGAAGGCGCCTACTACAACGAGGCGCCCAGCGAACCCAGGCCCGGGAAGAGTGCGGCGGT 957
Sus             CGACGGCACCTATTACAGCGAGGCGCCCAGCGAACCCCGGCCCGGGAAGAATGCTGCGGT 737
Mus             CGACACCGCCTACTACAGTGAGGCGGCGCGCGAGTCCAGGCCAGGGAAGAGTGCGGCTGT 797
                ***   * * ** ****  ****** *    **  ** **** *******  ** ** **

Bos             GTCGAGCCTCGACTGCCTGTCCAGCATCGTGGAGCGCATCTCCACCGAGAGCCCCGCCGC1011
Ovis            GTCGAGCCTCGACTGCCTGTCCAGCATCGTGGAGCGCATCTCCACCGAGAGCCCCGCAGC 1139
Homo            GTCGAGCCTAGACTGCCTGTCCAGCATCGTGGAGCGCATCTCCACCGAGAGCCCTGCGGC 1017
Sus             GTCGAGCCTCGACTGTCTGTCCAGCATCGTGGAGAGCATCTCCACCGAGAGCCCCGCCGC 797
Mus             GTCGAGCCTCGACTGCCTGTCCAGCATAGTGGAGCGCATCTCCACAGACAGCCCCGCTGC 857
                ********* ***** ********** ****** ********** ** ***** ** **

Bos             GCCCGCGCTTCTGCTAGCCGACGCGCCGCCGGAGTCGTCTCCTGGCCCGCAGGAGGC--- 1068
Ovis            GCCCGCGCTTCTACTGGCCGACGCGCCGCCGGAGTCGTCTCCCGGCCCGCAGGAGGCGGC 1199
Homo            GCCCGCCCTCCTGCTGGCGGACGTGCCTTCTGAGTCGCCTCCGCGCAGGCAAGAGGCTGC 1077
Sus             GCCCGCGCTTCTGCTGGCGGACACGCCGCGGGAGTCGTCTCCGGGCCCGCAAGAGGCGGC 857
Mus             GCCTGCGCTGCTTTTGGCAGATGCACCACCAGAGTCGCCTCCGGGTCCGCCAGAGGGGGC 917
                *** ** ** **  * ** **    **    ****** ****  *   **  ****

Bos             CGCCGGGAGCGAGGTGGAGCGC---GGCACCCCCGCTCCTTCCCCCGACACTGCCCCTCA 1125
Ovis            CGCCGGGAGCGAGGTGGAGTGC---GGCACCCCCGCCCCTTCCCCCGACACTGCCCCTCA 1256
Homo            CGCCCCCAGCGAGGGAGAGAGCAGCGGCGACCCCACCCAGTCACCGGACGCCGCCCCGCA 1137
Sus             CGCCGGGAGCGAGGTCGAGCGC---GGCACCCCCACCCCTTCCCCGGACGCCGCCCCGCA 914
Mus             ATCCCTAAGCGACACAGAACAG---GGAACCCAGACCCCGTCTCCCGACGCCGCCCCTCA 974
                  **   *****    **       **   **   * *  ** ** *** * ***** **

Bos             GGGCCTCGCGGGCGCGAACCCCAACCCGATTTACCAGGTGCTCTGAGGGGTTGTGCGGCC 1185
Ovis            GGGCCTCGCGGGCGCGAACCCCAACCCGATTTACCAGGTGCTCTGAGGGGTTGGGCGGCC 1316
Homo            GTGCCCTGCGGGTGCGAACCCCAACCCGATATACCAGGTGCTCTGAGGGGATG------- 1190
Sus             GTGCCCCGCGAGCGCGAACCCCAACCCTATCTACCAGGTGCTCTGA----------- 960
Mus             GTGTCCTGCAGGCTCAAACCCCAATGCGATTTATCAGGTGCTTTGAGAGATCG-ACTGCA 1033
                * * *  **  *  * ********  * ** ** ******** ***

Bos             TCATCGGGGGGCGCCGCTGCCCACGGGC-CCGAGGGATGGCGCCCTCAGGGTCCCTCGCG 1244
Ovis            TCATCGGGGGGCGCCGCTGCCCACAGGC-CCGAGGGATGGTGCCCCTAGGGTCCCTCGCG 1375
Homo            ------------GTGGCCGCCCACCCGC-CCGAGGGATGGTGCCCCTAGGGTCCCTCGCG 1237
Sus             ------------------------------------------------------------
Mus             GCAGCAGAGGGCGCACCACCGTAGGCACTCCTGGGGATGGTGTCCCT--GGTTCTTCACG 1091


Bos             CCCAAA-GATTGCGCTTAAGTGCCAACCACTCTCCTCCCAACAGCGCTTTAAAAGCGACC 1303
Ovis            CCCAAAAGATTGCGCTTAAGTGCCAACCACTCTCCTCCCAACAGCGCTTTAAAAGCGACC 1435
Homo            CCCAAAAGATTGAACTTAAATGCC-------CCCCTCCCAACAGCGCTTTAAAAGCGACC 1290
Sus             ------------------------------------------------------------
Mus             CCCAAAAGATGAAGCTTAAATGACA------CTCTTCCCAACTGTCCTTTCGAAGCCGTT 1145


Bos             -CTCCCGAGGTAGGAGAGGCGAAGGAACTGTTGTGTTTCCGCCTCCCGCACCCCAGGGCA 1362
Ovis            -CTCCCGAGGTAGGAGAGGCGAGGGAACTGTTGTGTTTCCGCCTCCCGCACCCCAGGGCA 1494
Homo            TCTCTTGAGGTAGGAGAGGCGGGAGAACTG--AAGTTTCCGCC-CCCGCCCCACAGGGCA 1347
Sus             ------------------------------------------------------------
Mus             CTTCCAGAGGGAAGGGAAGAGCAGAAGTC----TGTCCTAGAT---CCAGCCCCAAAGAA 1198


Bos             AGGACATGGTCCTTTTTTTCC------------------CGCCTACTCAGC-GCTCTTCCC 1404
Ovis            AGGACATGGTCCTTCCCCCCCACCACCCCCGCCCCCCCTCCCCACTCAGC-GCTCTTCCC 1553
Homo            AGGACACAGCGCGGTTTTTTC----------------------CACGCAGC-ACCCTTCTC 1385
Sus             ------------------------------------------------------------
Mus             AGGACATAGTCCTTTTTGTTGTTGTTGTTGTAGTCCTTCAGTTGTTTGTTTGTTTTTTCA 1258


Bos             TGAGACCCGCTGTGGCGGCCGTTTGA--------TGTTACTCCGTGGGCCAGAGCTGACC 1456
Ovis            TGAGACCCGCTGTGGTGGCCACTTGA--------TGTTACTCCGTGGGCCAGAGCTGACC 1605
```

```
Homo            GGAGACCCATTGCGATGGCCGCTCCG--------TGTTCCTCGGTGGGCCAGAGCTGAAC 1437
Sus             ------------------------------------------------------------
Mus             TGCGGCTCACAGCGAAGGCCACTTGCACTCTGGCTGCACCTCACTGGGCCAGAGCTGATC 1318

Bos             CTTGAGGAGCCAGGCCCTTCCTCTTCCTCGCCCCCTCCGCCATGGGGGTGTGTGACCCAC 1516
Ovis            CTTGAAGAGCCAGGCCCTTCCTCTTCCTCGCCCCCTCCCCCATGGCGGGGTGGGGCCCAC 1665
Homo            CTTGAGGGGCTAGG---TTCAGCTTTCTCGCGCCCTCCCCCATG--GGGGTGAGACCCTC 1492
Sus             ------------------------------------------------------------
Mus             CTTGAGTGGCCAGG----CGCTCTTCCTTTCCTCATAGCACA----GGGGTGAG--CCTT 1368

Bos             GCAGGCCTAAGCCCCGCCCCCAAGACCCCCGACCGCTTAAAGGGGCGTGCCCCTCCTTTC 1576
Ovis            GCAGGCCTAAGCCCCGCCCCCAAGACCCCCGACCGCTTACAGGGGCGTGCCCCTCCGTTC 1725
Homo            GCAGACCTAAGCCCTGCCCCGGGATGCACCGGTTATTTGGGGGGGCGTG----------- 1541
Sus             ------------------------------------------------------------
Mus             GCACACCTAAGCCCTGCCCTCCACATCCTT--TTGTTTGTCACTTTCTG----------- 1415

Bos             CGGAGCCCCCTGGGGACTTCAGCTGTTTCCCGCCGCTCCCTTCCCAAGTGT-AACAGGTG 1635
Ovis            CGGAGCCCCCTGGTGACTTCAGCTGTTTCCCGCCGCTCCCTTCCCAAGTGTTAACAGGTG 1785
Homo            --AGACCCAGTG---------------------CACTCCGGTCCCAAATGT-AGCAGGTG 1577
Sus             ------------------------------------------------------------
Mus             -GAGCCCTCCTGG----------------CACCCACTTTTCCCCACAG------------ 1446

Bos             TAATGGTAACCACTCCCACCCCCCAACCCCCACCCCCGGTTCAGGACCACTTTTTGTAAT 1695
Ovis            TAACGGTAACCACTCCCACCCCCCAACCCCCACCCCCGGTTCAGGACCACTTTTTGTAAT 1845
Homo            TAACCGTAA-----CCCACCCCC--AACCCGTTTCCCGGTTCAGGACCACTTTTTGTAAT 1630
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------

Bos             ACTTTTGTAATCTATTCCTGTAAATAAGAGTTGCTTTGCCAGAGCAGGAGCCCCTCGGGC 1755
Ovis            ACTTTTGTAATCTATTCCTGTAAATAAGAGTTGCTTTGCCAGAGCAGGAGCCCCTCGGGC 1905
Homo            ACTTTTGTAATCTATTCCTGTAAATAAGAGTTGCTTTGCCAGAGCAGGAGCCCCTGGGGC 1690
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------

Bos             TGTATTTATCTCTGAGGCATGGTGTGCGGTGCGACAGGGACTTTGTATGTTTATACGGCA 1815
Ovis            TGTATTTATCTCTGAGGCATGGTGTGCGGTGCGACAGGGACTTTGTATGTTTATACGGCA 1965
Homo            TGTATTTATCTCTGAGGCATGGTGTGTGGTGCTACAGGGAATTTGTACGTTTATACCGCA 1750
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------

Bos             GGCAGGCGAGCCGCGGGCGCTCGCTCAGGTGTTCGAAATAAAGACGCTAATTTATACAAA 1875
Ovis            GGCAGGCGAGCCGCGGGCGCTCGCTCAGGTGTTCGAAATAAAGACGCTAATTTATAA--- 2022
Homo            GGCGGGCGAGCCGCGGGCGCTCGCTCAGGTGATCAAAATAAAGGCGCTAATTTATAAAAA 1810
Sus             ------------------------------------------------------------
Mus             ------------------------------------------------------------

Bos             AAAAAAAAAAAAAAAAAAAAA 1897
Ovis            ----------------------
Homo            AAAAAAAAAAAAA--------- 1823
Sus             ----------------------
Mus             ----------------------
```

La zona di migliore sovrapposizione tra le sequenze comincia intorno al nucleotide 230 della sequenza di riferimento di Bos taurus.

Questo dato non sembra casuale visto che intorno a quella zona si colloca il codone ATG di inizio traduzione e quindi di inizio della regione codificante. Per disegnare i primer è stata concentrata l'attenzione sulla zona dal nucleotide 230 a circa 1000 della sequenza di riferimento. Sono stati così disegnati quattro primer, due forward e due reverse; uno dei due forward (**MyoDF1)** è degenerato.

La zona intorno alla sequenza dell'ATG mostra delle differenze tra le diverse sequenze prescelte per cui non è possibile scegliere un primer con sequenza definita. Per questo motivo abbiamo disegnato un cosiddetto primer degenerato. Ovvero viene disegnato il primer in modo che venga prodotta una miscela dei primer che differiscono tra loro solo per la cosiddetta base ambigua, ovvero per la base che risulta differente nelle diverse specie. In questo modo durante la reazione di PCR il primer portatore della base giusta potrà garantire la giusta amplificazione [27].

Esiste un codice per indicare le basi ambigue.

CODICE IUPAC-IUB per le BASI AMBIGUE

R (A or G) – Y (C or T) – M (A or C) – K (G or T) – S (G or C) – W (A or T) – H (A or C or T) – B (G or T or C) – V (G or C or A) D (G or T or A) – N (G or A or T or C)

Mediante i primer che coprono la massima sequenza possibile si ha un amplificato su cDNA di circa 1000 bp che copre l'intera regione codificante.

Nome del primer	sequenza	Posizione
MyoDF1	ATGGAGCTRCTGTCGCCGCC	Nt 215
MyoDF2	GATGACCCGTGTTTCGACTC	Nt 305
MyoDR1	TAGTCGTCTTGCGTTTGCAC	Nt 533
MyoDR1	CCCCTCAGAGCACCTGGTA	Nt 1157

N.B: la posizione dei primer è indicata rispetto alla sequenza di riferimento Bos Taurus BC120454

I tentativi di amplificazione usando tutti le possibili combinazioni dei 4 primer su cDNA non hanno dato risultato positivo. Probabilmente tale dato dipende dalla scarsissima rappresentatività del trascritto di MyoD nel tessuto muscolare adulto. Pertanto si è proceduto nelle operazioni di amplificazioni su DNA. Di seguito si riporta la sequenza ottenuta usando i primer **MyoDF2** e **MyoDR1**

Tale sequenza corrisponde a 228 bp nel putativo esone 1 è codifica per 75 aa della putativa proteina.

MyoDF2

GATGACCCGTGTTTCGACTCCCCGGACCTGCGCTTCTTCGAGGACCTGGATCCGCGCCTCGTG
CACGTGGGCGCGCTCCTGAAGCCCGAGGAACACTCGCACTTCCCTGCAGCCGCGCACCCGGCC
CCGGGCGCGCGCGAGGACGAGCATGTGCGCGCGCCCAGCGGGCACCACCAGGCGGGCCGCTGT
TTACTGTGGGCCTGCAAGGC**GTGCAAACGCAAGACGACT**

MyoDR1

```
posizione nucleotide/posizione aminoacido
1/1                                     31/11
GAT GAC CCG TGT TTC GAC TCC CCG GAC CTG CGC TTC TTC GAG GAC CTG GAT CCG CGC CTC
asp asp pro cys phe asp ser pro asp leu arg phe phe glu asp leu asp pro arg leu
61/21                                   91/31
GTG CAC GTG GGC GCG CTC CTG AAG CCC GAG GAA CAC TCG CAC TTC CCT GCA GCC GCG CAC
val his val gly ala leu leu lys pro glu glu his ser his phe pro ala ala ala his
121/41                                  151/51
CCG GCC CCG GGC GCG CGC GAG GAC GAG CAT GTG CGC GCG CCC AGC GGG CAC CAC CAG GCG
pro ala pro gly ala arg glu asp glu his val arg ala pro ser gly his his gln ala
181/61                                  211/71
GGC CGC TGT TTA CTG TGG GCC TGC AAG GCG TGC AAA CGC AAG ACG AC
gly arg cys leu leu trp ala cys lys ala cys lys arg lys thr
```

Per tutti gli altri geni considerati è stato svolto lo stesso tipo di lavoro e per questioni di brevità riportiamo le informazioni principali.

Miostatina

Il gene della miostatina svolge un ruolo importante nello sviluppo muscolare [4]. Tale gene appartiene alla famiglia dei fattori di crescita e differenziazione TGF-B. La proteina è costituita da 110 aa e concorre a regolare la proliferazione miocitica e di conseguenza contribuisce a regolare le dimensioni del muscolo. Nei bovini una singola sostituzione nucleotidica che porta ad una singola sostituzione amminoacidica porta al fenotipo "double muscle" tipico delle razze Belgian Blue e Piemontese.

Al momento della nostra analisi non c'era in banca dati nessuna sequenza relativa a tale gene per bufalo, ed abbiamo scelto come sequenza di riferimento la sequenza di Bos taurus la AB07643, lunga 1128 bp. Tale sequenza è stata immessa in Blast ed in questo modo abbiamo individuato le omologie con le sequenze di Homo sapiens (AF104922, 2823 bp), Mus musculus (AY204900, 2676bp) Ovis aries (AY918121), Sus scrofa (AY527153), capra (AY827576). Anche in questo caso le sequenze sono state allineate mediante ClustalW e sulle regioni di maggiore conservazione sono stati disegnati i primer.

Nome del primer	sequenza	Posizione
BMyo F	ATGCAAAAACTGCAAATC	Nt 134
BMyo2 F	TCGGACGGACATGCACTAA	Nt 344
BMYOF3	CTAACATCAGCAAAGATG	Nt 283
BMYOF4	CATGATCTTGCTTGTAACC	Nt 834
BMYOFn	ATGATGCAAAAACTGCAA	Nt 125
BMyo R	ACCCCACAGCGATCTACTACC	Nt 1254
BMyo2 R	GTCTACTACCATGGCTGGAAT	Nt 1215
BMYOR3	GTTAGGAGCTGTTTCCAG	Nt 263
BMYOR4	CACTGTCTTCACATCAAT	Nt 750
BMYORn	TCATGAGCACCCACA	Nt 1244

N.B: la posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus AB076403.

Di seguito si riporta la sequenza, ottenuta mediante amplificazione di cDNA derivato da tessuto muscolare, usando i primer **BMyo F** e **BMyo R**.

Tale sequenza è 1120 bp è corrisponde ai primi 373 aa della putativa proteina.

BMyo F

ATGCAAAAACTGCAAATCTCTGTTTATATTTACCTATTTATGCTGATTGTTGCTGGCCCAGTG
GATCTGAATGAGAACAGCGAGCAGAAGGAAAATGTGGAAAAAGAGGGGCTGTGTAATGCATGT
TTGTGGAGGGAAAACACTACATCCTCAAGACTAGAAGCCATAAAAATCCAAATCCTCAGTAAA
CTTCGCCTGGAAACAGCTCCTAACATCAGCAAAGATGCTATCAGACAACTTTTGCCCAAGGCT
CCTCCACTCCTGGAACTGATTGATCAGTTCGATGTCCAGAGAGATGCCGGCAGTGACGGCTCC
TTGGAAGACGATGACTACCACGCCAGGACGGACGCGGTCATTACCATGCCCACGGAGTCTGAT
CTTCTAACGCAAGTGGAAGGAAAACCCAAATGTTGCTTCTTTCAATTTAGCTCTAAGATACAA
TACAATAAACTAGTAAAGGCCCAACTGTGGATATATCTGAGACCTGTCAAGACTCCTGCGACA
GTGTTTGTGCAAATCCTGAGACTCATCAAACCCATGAAAGACGGTACAAGGTATACTGGAATC
CGATCTCTGAAACTTGACATGAACCCAGGCACTGGTATTTGGCAGAGCATTGATGTGAAGACA
GTGTTGCAAAACTGGCTCAAACAACCTGAATCCAACTTAGGCATTGAAATCAAAGCTTTAGAT
GAGAATGGTCATGATCTTGCTGTAACCTTCCCAGAACCAGGAGAAGATGGACTGACTCCTTTT
TTAGAAGTCAAGGTAACAGACACACCAAAAAGATCTAGGAGAGATTTTGGGCTTGATTGTGAT
GAGCGCTCCACAGAATCTCGATGCTGTCGTTACCCTCTAACTGTGGATTTTGAAGCTTTTGGA
TGGGATTGGATTATTGCACCTAAAAGATATAAGGCCAATTACTGCTCTGGAGAATGTGAATTT
GTATTTTTGCAAAAGTATCCTCATACCCATCTTGTGCACCAAGCAAACCCCAGAGGTTCAGCC
GGCCCCTGCTGCACTCCTACAAAGATGTCTCCAATTAATATGCTATATTTTAATGGCGAAGGA
CAAATAATATATGGGAAGATTCCAGCCAT**GGTAGTAGATCGCTGTGGGT**

BMyo R

posizione nucleotide/posizione aminoacido

```
1/1                                     31/11
ATG CAA AAA CTG CAA ATC TCT GTT TAT ATT TAC CTA TTT ATG CTG ATT GTT GCT GGC CCA
Met gln lys leu gln ile ser val tyr ile tyr leu phe met leu ile val ala gly pro
61/21                                   91/31
GTG GAT CTG AAT GAG AAC AGC GAG CAG AAG GAA AAT GTG GAA AAA GAG GGG CTG TGT AAT
val asp leu asn glu asn ser glu gln lys glu asn val glu lys glu gly leu cys asn
121/41                                  151/51
GCA TGT TTG TGG AGG GAA AAC ACT ACA TCC TCA AGA CTA GAA GCC ATA AAA ATC CAA ATC
ala cys leu trp arg glu asn thr thr ser ser arg leu glu ala ile lys ile gln ile
181/61                                  211/71
CTC AGT AAA CTT CGC CTG GAA ACA GCT CCT AAC ATC AGC AAA GAT GCT ATC AGA CAA CTT
leu ser lys leu arg leu glu thr ala pro asn ile ser lys asp ala ile arg gln leu
241/81                                  271/91
TTG CCC AAG GCT CCT CCA CTC CTG GAA CTG ATT GAT CAG TTC GAT GTC CAG AGA GAT GCC
leu pro lys ala pro pro leu leu glu leu ile asp gln phe asp val gln arg asp ala
301/101                                 331/111
GGC AGT GAC GGC TCC TTG GAA GAC GAT GAC TAC CAC GCC AGG ACG GAC GCG GTC ATT ACC
gly ser asp gly ser leu glu asp asp asp tyr his ala arg thr asp ala val ile thr
361/121                                 391/131
ATG CCC ACG GAG TCT GAT CTT CTA ACG CAA GTG GAA GGA AAA CCC AAA TGT TGC TTC TTT
met pro thr glu ser asp leu leu thr gln val glu gly lys pro lys cys cys phe phe
421/141                                 451/151
CAA TTT AGC TCT AAG ATA CAA TAC AAT AAA CTA GTA AAG GCC CAA CTG TGG ATA TAT CTG
gln phe ser ser lys ile gln tyr asn lys leu val lys ala gln leu trp ile tyr leu
481/161                                 511/171
AGA CCT GTC AAG ACT CCT GCG ACA GTG TTT GTG CAA ATC CTG AGA CTC ATC AAA CCC ATG
arg pro val lys thr pro ala thr val phe val gln ile leu arg leu ile lys pro met
541/181                                 571/191
AAA GAC GGT ACA AGG TAT ACT GGA ATC CGA TCT CTG AAA CTT GAC ATG AAC CCA GGC ACT
lys asp gly thr arg tyr thr gly ile arg ser leu lys leu asp met asn pro gly thr
601/201                                 631/211
GGT ATT TGG CAG AGC ATT GAT GTG AAG ACA GTG TTG CAA AAC TGG CTC AAA CAA CCT GAA
gly ile trp gln ser ile asp val lys thr val leu gln asn trp leu lys gln pro glu
661/221                                 691/231
TCC AAC TTA GGC ATT GAA ATC AAA GCT TTA GAT GAG AAT GGT CAT GAT CTT GCT GTA ACC
ser asn leu gly ile glu ile lys ala leu asp glu asn gly his asp leu ala val thr
721/241                                 751/251
TTC CCA GAA CCA GGA GAA GAT GGA CTG ACT CCT TTT TTA GAA GTC AAG GTA ACA GAC ACA
phe pro glu pro gly glu asp gly leu thr pro phe leu glu val lys val thr asp thr
781/261                                 811/271
CCA AAA AGA TCT AGG AGA GAT TTT GGG CTT GAT TGT GAT GAG CGC TCC ACA GAA TCT CGA
pro lys arg ser arg arg asp phe gly leu asp cys asp glu arg ser thr glu ser arg
841/281                                 871/291
TGC TGT CGT TAC CCT CTA ACT GTG GAT TTT GAA GCT TTT GGA TGG GAT TGG ATT ATT GCA
cys cys arg tyr pro leu thr val asp phe glu ala phe gly trp asp trp ile ile ala
901/301                                 931/311
CCT AAA AGA TAT AAG GCC AAT TAC TGC TCT GGA GAA TGT GAA TTT GTA TTT TTG CAA AAG
pro lys arg tyr lys ala asn tyr cys ser gly glu cys glu phe val phe leu gln lys
961/321                                 991/331
TAT CCT CAT ACC CAT CTT GTG CAC CAA GCA AAC CCC AGA GGT TCA GCC GGC CCC TGC TGC
tyr pro his thr his leu val his gln ala asn pro arg gly ser ala gly pro cys cys
1021/341                                1051/351
ACT CCT ACA AAG ATG TCT CCA ATT AAT ATG CTA TAT TTT AAT GGC GAA GGA CAA ATA ATA
thr pro thr lys met ser pro ile asn met leu tyr phe asn gly glu gly gln ile ile
1081/361                                1111/371
TAT GGG AAG ATT CCA GCC ATG GTA GTA GAT CGC TGT GGG
tyr gly lys ile pro ala met val val asp arg cys gly
```

Miogenina

Il gene della miogenina codifica per un regolatore trascrizionale che in concerto con MyoD1 da cui è regolata, determina la cascata di differenziamento muscolare [28]. Anche la proteina miogenina (224 aa) appartiene alla famiglia delle bHLH. Come MyoD1 viene espresso principalmente durante lo sviluppo embrionale, mantenimento e riparo del muscolo scheletrico.

Al momento della nostra analisi non c'era in banca dati nessuna sequenza relativa a tale gene per bufalo, abbiamo scelto come sequenza di riferimento la sequenza di Bos taurus la AB110600 (735 bp mRNA). Mediante Blast abbiamo identificata le sequenze con cui confrontarla: BC053899 (homo sapiens, 1573 bp), AF433651 (O.aries, 528bp), BC048683 (m. musculus, 1533 bp).

Nome del primer	sequenza	Posizione
Mge1F	TGGGCGTGTAAGGTGTGTAA	Nt 204
Mge2F	TGTCCACCTCCAGGGCTT	Nt 74
Mge5'F	CCATGGAGCTGTATGAGACC	Nt 17
Mge1R	CACTTACCGCCG**B**GTCCC	Nt 415
Mge2R	TTGTGGGCGTCTGTAGGG	Nt 592

N.B: la posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus AB110600.1

Di seguito si riporta la sequenza, ottenuta mediante amplificazione, usando i primer **Mge2F** e **Mge1R**.

Tale sequenza è 415 bp è corrisponde alla sequenza di parte del putativo primo esone e a 36 aa della putativa proteina.

Mge2F

TGTCCACCTCCAGGGCTTCGAGCCGCCAGGCTATGAGCGGGCTGAGCTCAGCCTGAGCCCTGA
GGCTCGCGTGCCCCTTGAAGACAAGGGGCTGGGGCCCGCGGAGCACTGCCCGGGCCAGTGCCT
GCCGTGGGCGTGTAAGGTGTGTAAGAGGAAGTCGGTGTCTGTGGACCGGCGGCGCGCCGCCAC
GCTGAGAGAGAAGCGCAGACTCAAGAAGGTGAATGAAGCCTTCGAGGCTCTCAAGAGGAGCAC
CCTGCTCAACCCCAACCAGCGGCTGCCCAAAGTGGAGATCCTGCGCAGCGCCATCCAGTACAT
AGAGCGCTTGCAGGCCCTGCTCAGCTCCCTCAACCAGGAGGAGCGCGACCTGCGCTACCGAGG
CGGGGGCGGAC**CCCAGCCGGCG**GTAAGTG

Mge1R

posizione nucleotide/posizione aminoacido

```
1/1                                     31/11
TGT CCA CCT CCA GGG CTT CGA GCC GCC AGG CTA TGA GCG GGC TGA GCT CAG CCT GAG CCC
val his leu gln gly phe glu pro pro gly tyr glu arg ala glu leu ser leu ser pro
61/21                                   91/31
TGA GGC TCG CGT GCC CCT TGA AGA CAA GGG GCT GGG GCC CGC GGA GCA CTG CCC GGG CCA
glu ala arg val pro leu glu asp lys gly leu gly pro ala glu his cys pro gly gln
121/41                                  151/51
GTG CCT GCC GTG GGC GTG TAA GGT GTG TAA GAG GAA GTC GGT GTC TGT GGA CCG GCG GCG
cys leu pro trp ala cys lys val cys lys arg lys ser val ser val asp arg arg arg
181/61                                  211/71
CGC CGC CAC GCT GAG AGA GAA GCG CAG ACT CAA GAA GGT GAA TGA AGC CTT CGA GGC TCT
ala ala thr leu arg glu lys arg arg leu lys lys val asn glu ala phe glu ala leu
241/81                                  271/91
CAA GAG GAG CAC CCT GCT CAA CCC CAA CCA GCG GCT GCC CAA AGT GGA GAT CCT GCG CAG
lys arg ser thr leu leu asn pro asn gln arg leu pro lys val glu ile leu arg ser
301/101                                 331/111
CGC CAT CCA GTA CAT AGA GCG CTT GCA GGC CCT GCT CAG CTC CCT CAA CCA GGA GGA GCG
ala ile gln tyr ile glu arg leu gln ala leu leu ser ser leu asn gln glu glu arg
361/121                                 391/131
CGA CCT GCG CTA CCG AGG CGG GGG CGG ACC CCA GCC GGC GGT AAG
asp leu arg tyr arg gly gly gly gly pro gln pro ala val
```

Myf5

Il gene Myf5 codifica per un altro componente della famiglia bHLH [2]. Anche questa proteina (255 aa) svolge un ruolo regolativo centrale nelle prime fasi del differenziamento muscolare.

Al momento della nostra analisi non c'era in banca dati nessuna sequenza relativa a tale gene per bufalo, abbiamo scelto come sequenza di riferimento la sequenza di Bos taurus (M95684 5219bp) .

Tale sequenza è stata confrontata con H.sapiens (BC069373 860 bp) e con M.musculus (AF336978 478 bp).

Nome primer	Sequenza	Posizione
Myf1F	CCTGAAGAAGGTCAACCAGG	Nt 480
Myf2F	CCAACCCTAACCAGAGGCTG	Nt 530
Myf1R	CATGCCATCAGAGCAACTTG	Nt 695
Myf2R	AGCCAACTATCCACCAGTAA	Nt 814

N.B: La posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus la M95684.

Myf1F

CCTGAAGAAGGTCAACCAGGCTTTCGACACGCTCAAGCGATGCACCACGACCAACCCTAACCAGAGGCTG
CCCAAGGTGGAGATCCTCAGGAATGCCATTCGCTACATTGAGAGTCTGCAGGAGCTGCTAAGGGAACAGG
TGGAAAACTACTATAGCCTGCCGGGGCAGAGCTGCTCTGAGCCCACCAGCCCCACCTCAAGTTGCTCTGA
TGGCATGGTAAGAGATGGCTCTGTACCTGCTAGGACCTTCCCAACTTTTATAAAAATCCTTACATCTCAT
TTAGACCAGGTGTAGCAACCAGATATTCAGTGG**TTACTGGTGGATAGTTGGCT**

Myf2R

```
posizione nucleotide/posizione aminoacido

1/1                                     31/11
CCT GAA GAA GGT CAA CCA GGC TTT CGA CAC GCT CAA GCG ATG CAC CAC GAC CAA CCC TAA
pro glu glu gly gln pro gly phe arg his ala gln ala met his his asp gln pro OCH
61/21                                   91/31
CCA GAG GCT GCC CAA GGT GGA GAT CCT CAG GAA TGC CAT TCG CTA CAT TGA GAG TCT GCA
pro glu ala ala gln gly gly asp pro gln glu cys his ser leu his OPA glu ser ala
121/41                                  151/51
GGA GCT GCT AAG GGA ACA GGT GGA AAA CTA CTA TAG CCT GCC GGG GCA GAG CTG CTC TGA
gly ala ala lys gly thr gly gly lys leu leu AMB pro ala gly ala glu leu leu OPA
181/61                                  211/71
GCC CAC CAG CCC CAC CTC AAG TTG CTC TGA TGG CAT GGT AAG AGA TGG CTC TGT ACC TGC
ala his gln pro his leu lys leu leu OPA trp his gly lys arg trp leu cys thr cys
241/81                                  271/91
TAG GAC CTT CCC AAC TTT TAT AAA AAT CCT TAC ATC TCA TTT AGA CCA GGT GTA GCA ACC
AMB asp leu pro asn phe tyr lys asn pro tyr ile ser phe arg pro gly val ala thr
301/101                                 331/111
AGA TAT TCA GTG GTT ACT GGT GGA TAG TTG GCT
arg tyr ser val val thr gly gly AMB leu ala
```

Di seguito si riporta la sequenza, ottenuta mediante amplificazione, usando i primer **Myf1F** e **Myf2R**

Tale sequenza è 333 bp è corrisponde a 111 aa della putativa proteina.

Tropomiosina

Il nome tropomiosina individua un gruppo di proteine che insieme con la troponina, contribuisce a regolare l'interazione dell'actina e della miosina nel citoscheletro [29].

Esistono numerose isoforme dovute all'esistenza di etero ed omodimeri. Nei vertebrati le catene sono codificate da 4 geni diversi e ciascun locus esistono numerosissimi splicing alternativi. Si riconoscono due gruppi ad alto e basso peso molecolare ed al primo gruppo appartengono le tropomiosine coinvolte nella contrazione muscolare. In considerazione di tale complessità abbiamo scelto di interessarci inizialmente ad un solo locus il TPM3, che codifica per una catena di 284 aa. La sequenza di Bos Taurus di riferimento è AB198072 .

Tale sequenza è stata confrontata con la sequenza di Homo sapiens BC008407 di 1155 bp.

Nome primer	Sequenza	Posizione
TpmF1	AAGCTGAGCAGAAGCAGGCAG	Nt 130
TpmR1	CTCAGCGTCAGCATCACC	Nt 305

N.B: La posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus la AB198072.

Troponina

Tre proteine, Troponina I, Troponina T e Troponina C, insieme alla tropomiosina, regolano l'interazione tra actina e miosina durante la contrazione muscolare [30]. Anche in questo caso la situazione genomica è alquanto complessa ed esistono numerose varianti. Come punto di partenza

per poter poi iniziare una più precisa caratterizzazione siamo partiti da un singolo trascritto, ovvero quello relativo alla Troponina T di Bos Taurus AF175558 (993bp) che codifica per una proteina di 284 aa. Dopo l'analisi mediante Blast, tale sequenza è stata confrontata con la sequenza di M.musculus BC003747 (1101 bp).

La discontinuità della conservazione dei nucleotidi ha consentito la produzione di primer degenerati. A partire dal frammento così ottenuto si potrà, sempre utilizzando primer degenerati si potrà allargare le dimensioni del frammento di PCR.

Nome primer	Sequenza	Posizione
TrpF1	AAGGACCTCAWCGAGCTSCA	Nt 330
TrpR1	CTTGGCCTTWTCCCTCAGCT	Nt 741

N.B: La posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus la AF175558.

Miosina

Il termine miosina individua nei mammiferi una ampia superfamiglia di proteine con funzioni anche molto diversificate. Una funzione meglio caratterizzata è quella di componente portante del sarcomero muscolare **[31]**.

Anche rispetto a questa funzione la situazione genomica è alquanto complessa in quanto esistono diverse varianti.

Abbiamo preso in considerazione la catena pesante espressa preferenzialmente nel muscolo scheletrico adulto MYH1 e la catena leggera

MYL. Per entrambi i dati genomici sono stati abbastanza limitati.i fino al 2006 e solo per la pesante nel 2007 è stata immessa in banca dati una sequenza di Bos taurus. Pertanto per la catena pesante siamo partiti dalle sequenze di M. musculus BC108329 (6058 bp) ed Homo sapiens BC114545 (5870 bp) e di recente l'abbiamo confrontata con quella di B.taurus NM_174117 (5987 bp).

Catena pesante

Nome primer	Sequenza	Posizione
MYHF1	TTGCATCCCTAAAGGCAGGC	Nt 30
MYHF2	ACCTCCGGAAGTCTGAAAAG	Nt 109
MYHF3	CCAGTGTATAACGCAGAGGT	Nt 479
MYHR1	ACCTCTGCGTTATACACACTGG	Nt 459
MYHR2	CGCTGGATGGCATCCGTCTC	Nt 4197
MYHR3	CCAAGCTGCTGCTGCTTCAGAG	Nt 2757

N.B: La posizione dei primer è indicata rispetto alla sequenza di riferimento Bos taurus la NM174117

Catena leggera

In questo caso è stata usata come sequenza di riferimento quella di O. aries (DQ152977 362 bp). Tale sequenza è stata confrontata con quella predetta di AACAGGGTCACCAACCAGCC

Nome primer	Sequenza	Posizione
MYLF1	ATGACGCCAAGGATTTTATC	Nt 4
MYLF2	CAAGAACATGGAGGCCAAGA	Nt 104
MYLR1	GGGTTTTACATGGGGCTTCT	Nt 362
MYLR2	GGCTGGTTGGTGACCCTGTT	Nt 273

N.B: La posizione è indicata rispetto alla sequenza di riferimento O. aries DQ152977.

MLCK

La Myosin Light Chain Kinase (MLCK) è una proteina che fosforila la catena leggera della miosina durante la contrazione cellulare muscolare e non muscolare, durante la citochinesi ed in generale nella organizzazione del citoscheletro [1]. Le sequenze disponibili in rete sono solo quelle relative ad uomo e ad altri primati per cui dobbiamo sospendere l'analisi.

Mioglobina

La mioglobina è una proteina globulare che lega l'ossigeno in maniera reversibile [32]. E' un trasportatore intracellulare dell'ossigeno in cellule specializzate, quail le cellule muscolari. Tali cellule necessitano infatti di quantità costanti di ossigeno per poter espletare la loro funzione contrattile in condizioni aerobiche. La componente proteica è un polipeptide di 154 aa.

Homo NM_005368 1078 bp e Sus scrofa NM_214236.

Nome primer	Sequenza	Posizione
MgbF1	GCCATGGGGCTCAGCGACGG	Nt 78
MgbF2	TCTGGGGGAAGGTGGAGGCTGA	Nt 100
MgbR1	GAGATGAACTCCAGGTACTT	Nt 399
MgbR2	CCTGGAAGCCCAGCTCCTTGTA	Nt 520

N.B: La posizione è indicata rispetto alla sequenza di riferimento H.sapiens NM_005368 .

3.2 Progettazione basi di dati

Questa parte dei risultati, nella fase preliminare della progettazione, era indirizzata all'elaborazione di una base di dati relazionale dedicata alla raccolta, catalogazione e collegamento di informazioni relative al bufalo. Infatti, il nostro particolare interesse era di ottenere un sistema in grado di relazionare i dati genomici, ottenuti sperimentalmente in laboratorio, con aspetti fisologici e manageriali.

L'analisi delle specifiche esigenze e la dissezione dei requisiti hanno portato alla formulazione di un progetto di base di dati che può in qualche modo risultare più versatile rispetto a quelli esistenti da noi conosciuti. Infatti, è stato organizzato in modo da poter collegare dati relativi a specifici polimorfismi genomici con particolari caratteristiche biologiche, e questo non solo relativamente alla specie bufalina ma anche per altre specie di interesse zootecnico.

3.2.1 Analisi dei requisiti.

Requisiti espressi in linguaggio naturale

a. Si vuole progettare una basi di dati, allo scopo di automatizzare la gestione di tutto il materiale, o riferimenti a tale materiale, utile al lavoro sperimentale e potenzialmente anche ad altre attività correlate a qualsiasi specie animale;

b. In prima istanza, bisognerà catalogare i dati molecolari e funzionali che via via si produrranno; (sequenze genomiche), e dati derivanti dal confronto con sequenze di altre specie, presenti nelle banche dati biologiche (percentuali di conservazione - geni noti - lunghezza sequenza simile, numero di esoni e introni, lunghezza della proteina).

Delle sequenze ottenute a partire da un tessuto scelto, si dovrà effettuare uno studio comparativo con altre sequenze presenti nelle banche dati biologiche. Dal raffronto risulterà una duplice possibilità, ovvero che le sequenze prodotte possono essere note (sequenze geniche) oppure sconosciute.

Se siamo in presenza di **sequenze geniche** si dovrà registrare:

- La lunghezza;
- L'organizzazione strutturale (numero esoni e introni);
- Il nome del gene;
- Descrizione del gene
- Percentuale di conservazione – nome della specie – lunghezza sequenza simile;
- Lunghezza della proteina;

- Eventuali polimorfisimi;

mentre per le **sequenze sconosciute** si dovrà registrare:

- Lunghezza;
- Lunghezza dell'ORF;
- Descrizione della sequenza

c. Indipendentemente dai dati prodotti in laboratorio, scelta la specie e a seconda del tipo di ricerca che si vuole effettuare su essa, bisognerà:

c1. catalogare i dati relativi ai **prodotti di origine animale**, associato ad un particolare tessuto, ovvero:

Nome del prodotto;

- Descrizione del prodotto;
- Composizione Biochimica (% di proteine - % di Acidi grassi insaturi -%di Acidi grassi saturi - % di Colesterolo);
- Geni associati la prodotto

c2. catalogare i dati relativi al comparto **allevamento**, ovvero:

- Condizioni ambientali (descrizione della nutrizione – descrizione dello spazio dell'allevamento);

c3. catalogare i dati relativi alle **funzioni biologiche** coinvolte in un certo tessuto, ovvero:

- Nome della funzione biologica;
- I geni coinvolti;
- Gli ormoni coinvolti.

d. Successivamente, bisognerà tenere traccia di dati relativi alle funzioni biologiche con i dati prodotti in laboratorio, ovvero, la base di dati dovrà specificare se i geni coinvolti in una particolare funzione biologica sono stati caratterizzati in laboratorio.

e. Bisognerà inoltre tenere traccia di specifici aspetti biologici dei prodotti di origine animale (geni associati), con i dati relativi alle funzioni biologiche. Ovvero se i geni coivolti in un particolare prodotto di origine animale sono presenti in una particolare funzione biologica.

f. Bisognerà traccia di specifici aspetti biologici dei prodotti di origine animale (geni associati), con i dati prodotti in laboratorio. Ovvero se i geni coivolti in un particolare prodotto di origine animale sono stati caratterizzati in laboratorio.

g. Bisognerà inoltre tenere traccia degli aspetti relativi all'animale – o più in generale della specie – con il comparto allevamento. Ovvero a

partire da un animale, risalire ai dati relativi all'allevamento (condizioni ambientali) e all'azienda gestore dell'allevamento.

3.2.2 Analisi delle specifiche

Analizzando la raccolta dei requisiti, allo scopo di definire dettagliatamente l'organizzazione della base di dati, abbiamo che:

Per ogni sequenza prodotta in laboratorio (ottenuta da un tessuto) a seconda se essa è nota o sconosciuta sono state individuate rispettivamente due tipi di entità: **Sequenza genica** con i seguenti attributi specifici:

- gli attributi atomici [ID_gene (PK)], [nome_gene], [descr_gene], [sequenza], [ID_seq_lab](attributo booleano) e [lunghezza];
- l'attributo composto [organizzazione strutturale] con i relativi attributi semplici [Numero_esoni], [Numero_introni];
- l'attributo composto [Percentuale di conservazione] con i relativi attributi semplici [Nome_specie_simile], [% similarità];
- gli attributi atomici [lunghezza sequenza simile], [Lunghezza_proteina], [descrizioni varie], [Polimorfismi].

Mentre per il tipo di entità **Sequenza sconosciuta** sono stati individuati i seguenti attributi specifici:

- gli attributi atomici: [ID_sequenza_sconosciuta], [Lunghezza] [Lunghezza ORF] e [descr_seq_sconosciuta];

Per tipo di entità **Tessuto** sono stati individuati i seguenti attributi atomici:

- [ID_tessuto (PK)];
- [nome_tessuto];
- [Descrizioni_varie].

Per il tipo di entità **specie**, la base dati tiene informazioni sul [ID_specie], [nome_specie].

Per il tipo di entità **Animale**, la base dati tiene informazioni sul [ID_animale (PK)], [nome_animale], [razza], [sesso], [età], [descr_varie].

Per tipo di entità **Allevamento** la base dati tiene informazioni sui seguenti attributi:

- L'attributo atomico [ID_allevamento];
- l'attributo composto [condizioni ambientali] con i relativi attributi semplici [descr_nutizione], [descr_spazio_allevamento].

Per il tipo di entità **Azienda**, la base dati tiene informazioni sul [ID_azienda (PK)], [nome_azienda], [descr_azienda].

Per tipo di entità **Prodotti di origine animale** la base dati tiene informazioni sui seguenti attributi:

- gli attributi atomici [ID_Prodotto (PK)], [Nome_Prodotto], [Descr_Prodotto] ;
- l'attributo composto [Composizione_Biochimica] con i relativi attributi semplici [% colesterolo], [% Proteine], [% Acidi grassi saturi], [% Acidi grassi insaturi];

- l'attributo composto-multivalore [Nomi_geni_prod] con i relativi attributi semplici [ID_gene_prodotto], [Descr_gene_prod].

Per tipo di entità **Funzione biologica** la base dati tiene informazioni sui seguenti attributi:

- gli attributi atomici [ID_funzione], [nome_funzione];
- l'attributo composto-multivalore [ormoni] con i relativi attributi semplici [ID_ormoni], [Descr_ormoni];
- l'attributo composto-multivalore [Nomi_geni_funz] con i relativi attributi semplici [ID_gene_funz], [Descr_gene_funz].

All'interno della base di dati sono state individuate le seguenti dodici associazioni:

- **[ottenutaSGT]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione totale*, che lega il tipo di entità **sequenza genica** al tipo di entità **tessuto**;
- **[ottenutaSST]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione totale*, che lega il tipo di entità **sequenza sconosciuta** al tipo di entità **tessuto**;
- **[coinvolta]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **tessuto** a **[coinvolta]**) , che lega il tipo di entità **Funzione biologica** al tipo di entità **tessuto**;
- **[associato]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **tessuto** a **[associato]**) , che lega

il tipo di entità **Prodotti di origine animale** al tipo di entità **tessuto**;

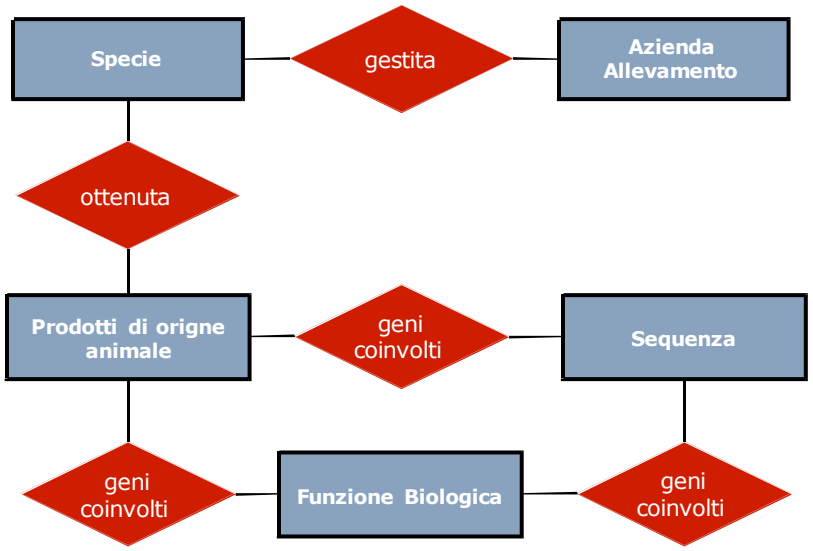
- **[coinvolti P-FB]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale*, che lega tra loro i tre tipi di entità: **Prodotti di origine animale**, **Funzioni biologiche** e **sequenze geniche.**
- **[coinvolti P-SG]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **Sequenze geniche** a **[coinvolti P-SG]**) , che lega il tipo di entità **Prodotti di origine animale** al tipo di entità **Sequenze geniche**;
- **[coinvolti FB-SG]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **Funzioni biologiche** a **[coinvolti FB-SG]**) , che lega il tipo di entità **Funzioni biologiche** al tipo di entità **Sequenze geniche**;
- **[ottenuti]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **Specie** a **[ottenuti]**), che lega il tipo di entità **Specie** al tipo di entità **Prodotti di origine animale**;
- **[di]** con *rapporto di cardinalità* del tipo (m:n) e *vincolo di partecipazione parziale* (*totale* di **Funzioni biologiche** a **[di]**), che lega il tipo di entità **Specie** al tipo di entità **Funzioni biologiche**;
- **[appartieneSA]** con *rapporto di cardinalità* del tipo (1:n) e *vincolo di partecipazione totale*, che lega il tipo di entità **Specie** al tipo di entità **Animale**;

- **[appartieneAA]** con *rapporto di cardinalità* del tipo (n:1) e *vincolo di partecipazione parziale* (*totale* di **Allevamento** a **[appartieneAA]**) , che lega il tipo di entità **Animale** al tipo di entità **Allevamento**;
- **[gestito]** con *rapporto di cardinalità* del tipo (n:1) e *vincolo di partecipazione totale*, che lega il tipo di entità **Allevamento** al tipo di entità **Azienda**;

3.2.3 Programmazione Concettuale

Per la progettazione del modello *EER*, si è deciso di utilizzare la strategia *inside-out* in cui l'attenzione è focalizzata su un insieme centrale di concetti di maggiore rilevanza, costruendo uno schema scheletro.

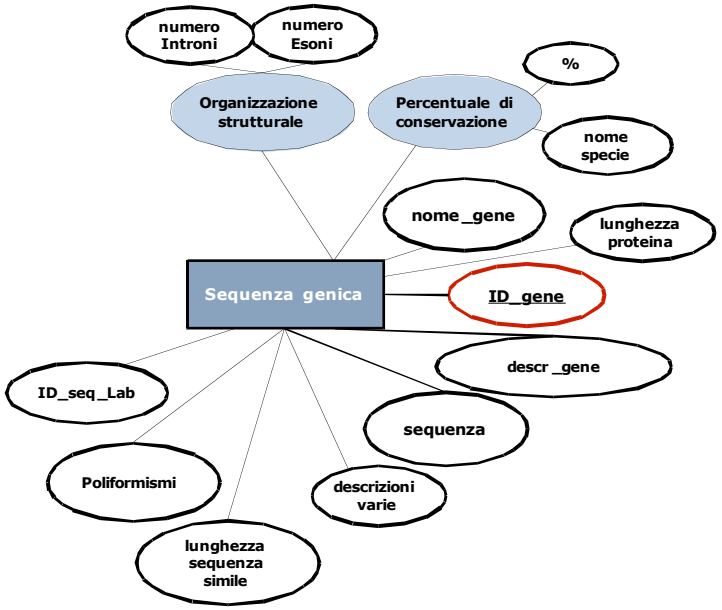
Da una prima analisi fatta sulle specifiche viste precedentemente, si è deciso di fissare il seguente schema:



In seguito, la modellazione si espande verso l'esterno considerando concetti vicini a quelli esistenti. Potremmo specificare i tipi di entità coinvolti nello schema aggiungendo successivamente le associazioni specifiche per ottenere il diagramma *EER* finale.

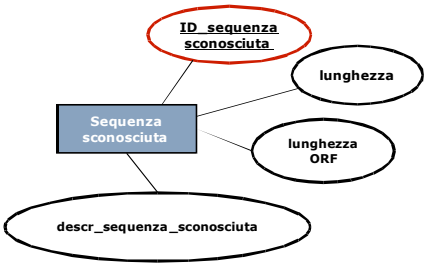
3.2.4 Sviluppo delle Entità presenti nella basi di dati (Inside-out)

Tipo di Entità: Sequenza genica

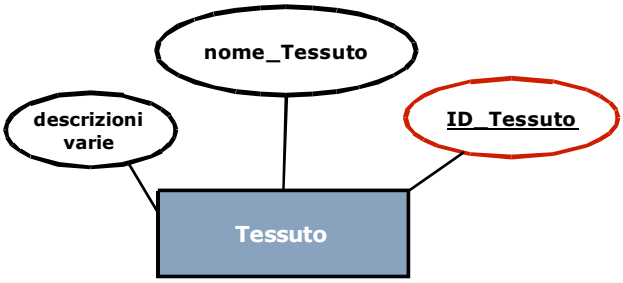


Del tipo di entità **Sequenza genica** si segnala l'attibuto atomico ID_seq_lab, indispensabile alla caratterizzazione dei geni sequenziali in laboratorio composto.

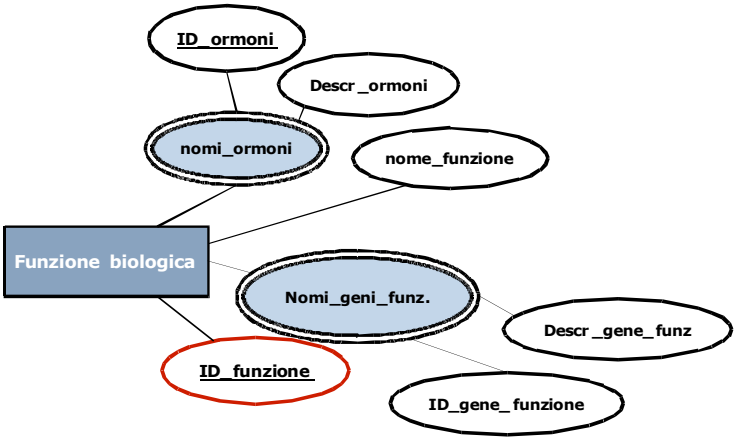
Tipo di Entità: **Sequenza sconosciuta**



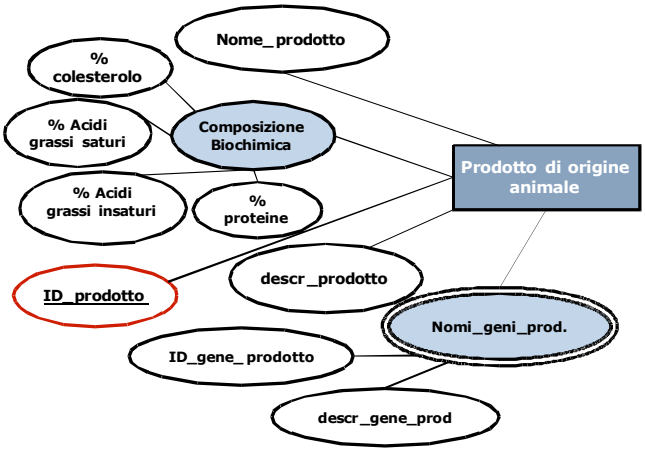
Tipo di Entità: **Tessuto**



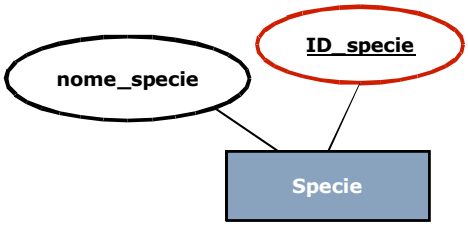
Tipo di Entità: **Funzione Biologica**



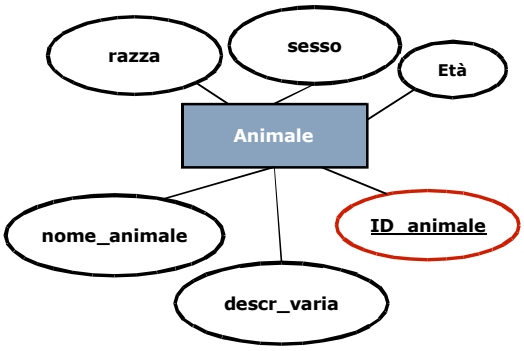
Tipo di Entità: **Prodotti di origine animale**



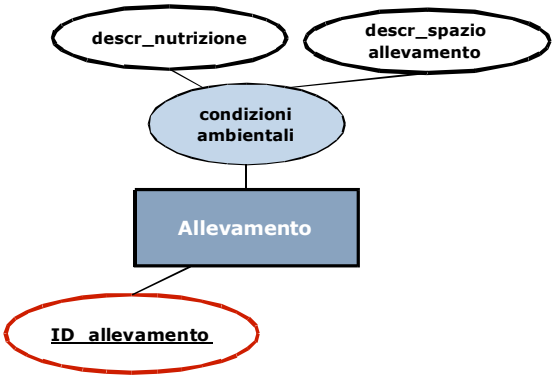
Tipo di Entità: **Specie**



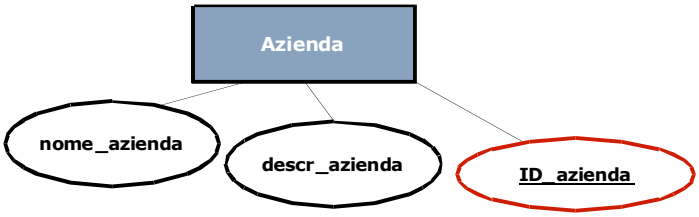
Tipo di Entità: **Animale**



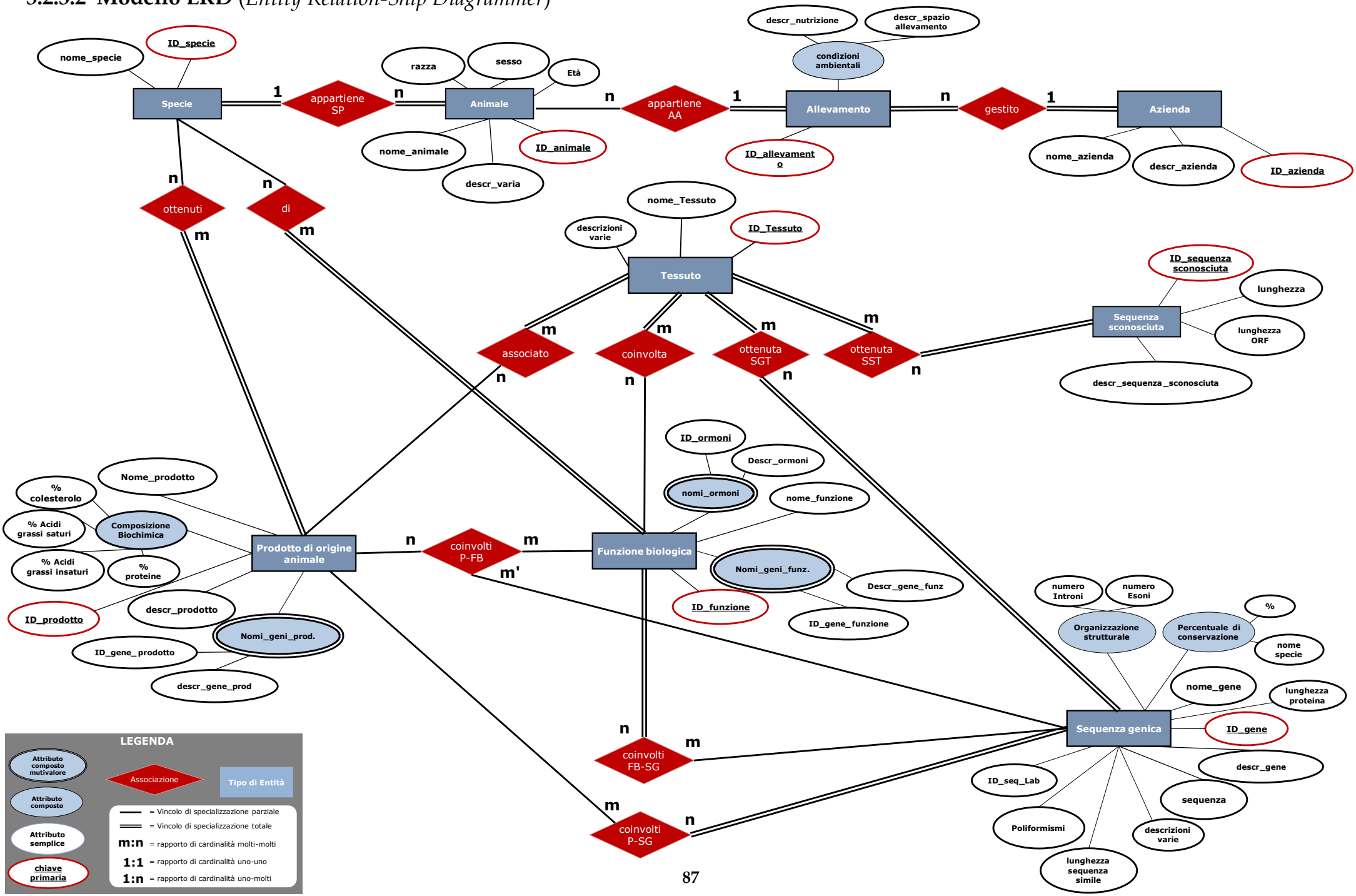
Tipo di Entità: **Allevamento**



Tipo di Entità: **Azienda**



3.2.3.2 **Modello ERD** (*Entity Relation-Ship Diagrammer*)



3.2.4 Dizionario dei dati

Uno schema *EER* va corredato con una documentazione di supporto, per facilitare l'interpretazione dello schema stesso e per descrivere proprietà dei dati che non possono essere espresse direttamente dai costrutti del modello **[9]**.

La documentazione dei vari concetti rappresentati in uno schema, ovvero le regole di tipo descrittivo, può essere prodotta facendo uso di un *dizionario dei dati*. Esso è composto da due tabelle: la prima descrive le Entità dello schema con il nome, una definizione informale in linguaggio naturale, l'elenco di tutti gli attributi e i possibili identificatori, la seconda invece descrive le associazioni dello schema con il nome, una descrizione in linguaggio naturale, le Entità coinvolte (e i relativi vincoli) ed eventuali attributi si associazione.

L'uso del dizionario dei dati è particolarmente importante nei casi in cui lo schema è complesso (molti concetti collegati in maniera articolata) e risulta pesante specificare direttamente sullo schema tutti gli attributi di entià e relazioni.

Di seguito, sono riportate le tabelle relative ai tipi di Entià coinvolte e le Associazioni nella nostra basi di dati.

Entità	*Descrizione*	*Attributi*	*Identificatore*
Sequenza genica	Sequenze nucleonica conosciuta contenete un gene, ottenuta in laboratorio a partire da un particolare tessuto	ID_gene, nome_gene, descr_gene, sequenza, nome specie simile, % di conservazione, lunghezza proteina, lunghezza sequenza simile, polimorfismi, descrizione varie	ID_gene
Sequenza sconosciuta	Sequenze nucleonica sconosciuta	ID_sequenza_sconosciuta, lunghezza, lunghezzaORF	ID_sequenza_sconosciuta
Tessuto	Tessuti oggetti di studio.	ID_tessuto, nome_tessuto, descrizioni varie	ID_tessuto
Funzione biologica	Funzioni del tipo: metabolismo, respirazione, riproduzione, escrezione, digestione	ID_funzione, nome_ funzione, nomio_geni_funzione (ID_gene_funzione, descr_gene_funzione), nomi_ormoni (ID_ormoni-descr_ormoni),	ID_funzione
Prodotto di origine animale	Prodotti derivanti da animali. Ad esempio (latte, Carne)	ID_prodotto, nome_prodotto, composizione biochimica (%colesterolo, %acidi grassi satuti, %acidi grassi insaturi, %proteine) descr_prodotto, nomi geni_prodotto(ID_gene_prod,descr_gene_prod)	ID_prodotto
Specie	Specie animale oggetta dello studio	ID_specie, nome_specie	ID_specie
Animale	Animale oggetto dello studio	ID_animale, nome_animale, sesso, età, razza, descr_varie	ID_animale
Allevamento	Allevamento dove vengono allevati gli animali (o specie) oggetto dello studio	ID_allevamento ,Condizioni ambientali (descr_nutrizione, descr_spazio allevamento)	ID_allevamento
Azienda	Azienda che gestisce l'allevamento	ID_azienda, nome_azienda, descr_azienda	ID_azienda

Associazione	*Descrizione*	*Entità coinvolte*	*Attributi*
OttenutaSST	Associa ad ogni sequenza sconosciuta il tessuto da cui è stata ottenuta.	Tessuto-Sequenza sconosciuta (m:n)	-
OttenutaSGT	Associa ad ogni sequenza genica - sequenziata in laboratorio o proveniente da altra fonte - il tessuto da cui è stata ottenuta.	Tessuto-Sequenza genica (m:n)	
Coinvolta	Associa ad ogni funzione biologica il tessuto coinvolto.	Tessuto-Funzioni biologiche (m:n)	-
Associato	Lega ad ogni prodotto di origine animale il tessuto a cui è associato.	Tessuto-Prodotto di origine animale (m:n)	-
Coinvolte P-FB	Seleziona i geni coinvolti nello specifico prodotto di origine animale, e i geni coinvolti in funzioni biologiche	Prodotto di origine animale – Funzione biologica – sequenze genica (m:n)	-
Coinvolte P-G	Seleziona i geni coinvolti nello specifico prodotto di origine animale, e i geni sequenziali in laboratorio	Prodotto di origine animale – geni (m:n)	-
Coinvolte FB-G	Seleziona i geni coinvolti nelle funzioni biologoche, e i geni sequenziali in laboratorio	Funzione biologica – geni (m:n)	-
Ottenuti	Associa i prodotti di origine animale alla specie da cui provengono	Prodotto di origine animale-Specie (m:n)	-
Di	Associa una funzione biologica di una particolare specie	Specie- Funzione biologica (m:n)	-
Appartiene SA	Associa il singolo capo (animale) a quale specie appartiene	Specie-Animale (1:n)	-
Appartiene AA	Associa il singolo capo (animale) a quel allevamento appartiene	Animale -Allevamento (n:1)	-
Gestito	Associa l'allevamento a quale azienda appartiene	Allevamento-Azienda (n:1)	-

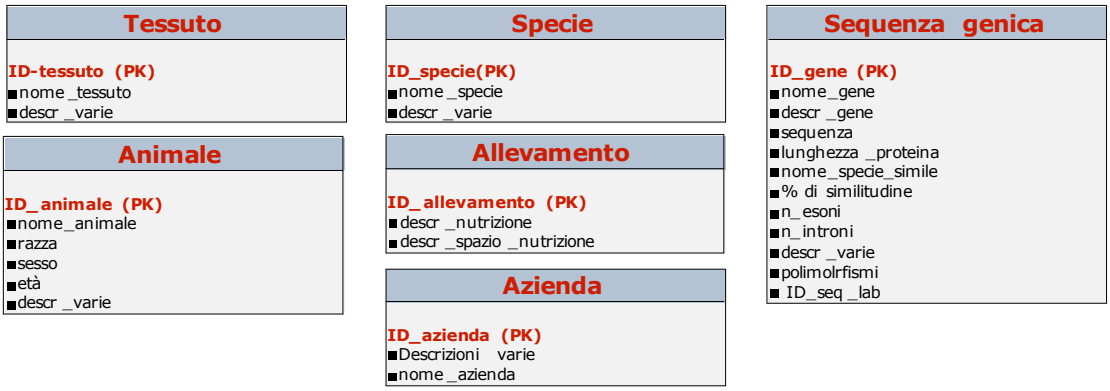
3.2.6 Mapping del modello dei dati - progettazione logica della base di dati -

A partire dallo schema concettuale progettato, si progetterà uno schema di basi di dati relazionale. Analizzeremo in particolare i passi di un algoritmo per la traduzione da uno schema *EER* a uno schema relazionale.

Passo 1: Traduzione di tipi di entità.

Per ogni tipo di entità forte *E*, visto nello schema *EER*, si deve costruire una relazione (tabella) che contenga tutti gli attributi semplici *E*. Di un attributo composto si inseriscano solo gli attributi componenti semplici. Di un attributo multivalore, si costruisce una relazione a parte. Dopo un analisi dell'entità *E*, si scelga come chiave primaria uno degli attributi chiave di *E*.

Nel nostro caso avremo le seguenti relazioni:



Passo 3: Traduzione associazione con rapporto di cardinalità del tipo 1:n

Per ogni tipo di associazione binaria 1:n, nello schema *EER*, si individua le relazione *S* corrispondenti al *lato-n* dell'associazione *R*.

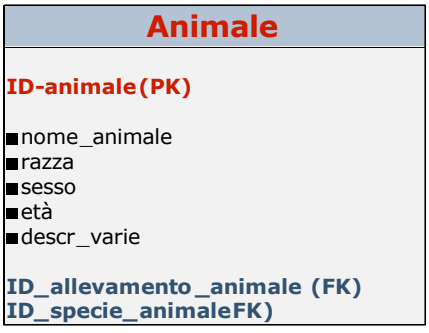
S'inserisce come chiave esterna la chiave primaria della relazione *T* che rappresenta l'altro tipo di entità parteciapante all'associazione *R*; ciò perché ogni istanza di entità *lato-n* è collegata al più un istanza al *lato-1* del tipo di associazione *R*.

S'inseriscano fra gli attributi di *S* tutti gli attributi semplici o componenti di attributi composti (qualora ce ne fossero) del tipo di associazione.

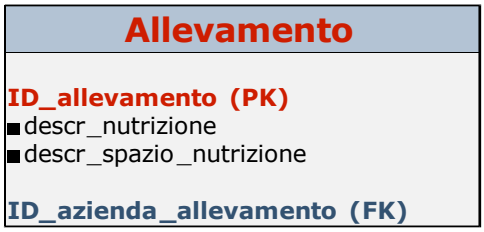
Nel nostro caso abbiamo tre associazione con queste caratteristiche **[appartieneSA], [appartieneAA], [gestito]**.

La prima associazione **[appartieneSA]** è tra **Specie** e **Animale**. L'entità che partecipa al *lato-n* è **Animale**, per cui s'inserisce come chiave esterna la chiave primaria di **Specie** rinominata in [ID_specie_animale].

La seconda associazione **[appartieneAA]** è tra **Animale** e **Allevamento**. L'entità che partecipa al *lato-n* è **Animale**, per cui s'inserisce come chiave esterna la chiave primaria di **Allevamento** rinominata in [ID_allevamento_animale].



La terza associazione **[gestito]** è tra **Allevamento** e **Azienda**. L'entità che partecipa al *lato-n* è **Allevamento**, per cui s'inserisce come chiave esterna la chiave primaria di **Azienda** rinominata in [ID_azienda_allevamento].



Passo 4: Traduzione associazione con rapporto di cardinalità del tipo m:n

Per ogni tipo di associazione binaria m:n, nello schema *EER*, si costruisce una nuova relazione *S* che rappresenti il tipo di associazione.

S'inserisce come attributi di chiave esterna *S* le chiavi primarie delle relazioni che rappresentano i tipi di entità partecipanti. Come chiavi primarie si utilizzeranno una combinazione delle due.

S'inseriscano fra gli attributi di *S* tutti gli attributi semplici o componenti di attributi composti (qualora ce ne fossero) del tipo di associazione m:n.

Questa nuova relazione creata prende il nome di ***relazione associazione***.

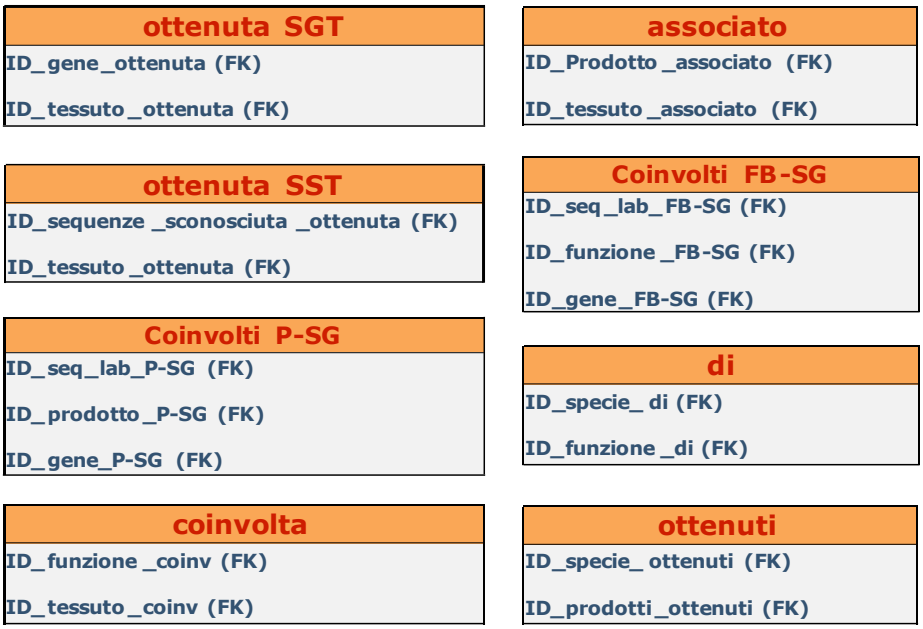
Nel nostro caso, per i tipi di associazione **[coinvolta]** , **[associato]** , **[ottenutaSGT]** , **[ottenutaSST]**, **[di]** , **[ottenuti]** ridenominiamo le chiavi primarie di tutti i tipi di entità che fanno parte delle associazioni, rispettivamente in:

- [ID_funzione_coiniv] [ID_tessuto_coinv];

- [ID_prodotto_associato] [ID_tessuto_associato];
- [ID_gene_ottenuta] [ID_tessuto_ottenuta];
- [ID_sequenze_sconosciuta_ottenuta] [ID_tessuto_ottenuta];
- [ID_specie_di][ID_funzione_di];
- [ID_specie_ottenuti][ID_prodotti_ottenuti];

Invece per i tipi di associazione **[coinvolti FB-SG]** , **[coinvolti P-SG]** bisogna fare una precisione. Poiché nel tipo di entità sequenza genica sono memorizzati tutti i geni coinvolti nella base di dati, bisogna trovare un modo per tenere traccia dei geni che sono sequenziali in laboratorio. All'occorrenza è stato creato l'attributo specifico booleano [ID_seq_lab] dove assumerà il valore VERO per tutti i geni sequenziati in laboratorio FALSO per tutti gli altri casi. Nello specifico, quando bisogna selezionare i geni che sono coinvolti rispettivamente nella funzione biologica e sequenza genica prodotta in laboratorio (**[coinvolti FB-SG]**) e prodotto di origne animale e sequenza genica prodotta in laboratorio (**[coinvolti P-SG]**), bisognerà tenere conto dell' attributo aggiuntivo ci sarà [ID_seq_lab] ridenominiamo in [ID_seq_lab_FB-SG] per la relazione **[coinvolti FB-SG]** mentre [ID_seq_lab_P-SG] per la relazione **[coinvolti P-SG]**. Concludendo, per le relazioni **[coinvolti FB-SG]** e **[coinvolti P-SG]** avremo i rispettivi attributi ridenominati

- [ID_seq_lab_FB-FG] [ID_funzione_FB-FG] [ID_gene_FB-FG];
- [ID_seq_lab_P-SG] [ID_prodotto_P-SG] [ID_gene_P-SG];



Passo 5: Traduzione di tipi di associazione N-arie

Per ogni tipo di associazione n-aria *R*, dove , si costruisce una nuova relazione *S* per rappresentare *R*. Si inserisce come chiave esterna, le chiavi primarie delle relazioni che rappresentano i tipi di entità partecipanti. S'inseriscano fra gli attributi di *S* tutti gli attributi semplici o componenti di attributi composti (qualora ce ne fossero) del tipo di associazione n-aria *R*. La chiave primaria di *S* è di solito una combinazione di tutte le chiavi esterne che riferiscono le relazioni rappresentanti i tipi di entità partecipanti.

Nel nostro caso, siamo in presenza di una relazione 3-aria **[coinvolti P-FB]** che coinvolge i tipi di entità **Prodotto di orgine animale – Funzione Biologica-Sequenza genica**. Infatti l'associazione **[coinvolti P-FB]** seleziona i geni coinvolti in un prodotto di orgine animale e una funzione biologica. Tenendo conto

della regola di formazione della relazione di una associazione N-aria, ridenominiamo le rispettivi chiavi primarie in:

- [ID_prodotto_P-FG] [ID_funzione_P-FG] [ID_gene_P-FG]

Coinvolti P-FB
ID_prodotto_P-FB (FK)
ID_funzione_P-FB (FK)
ID_gene_P-FP (FK)

Passo 6: Traduzione Relazione-attributi multivalore (composto)

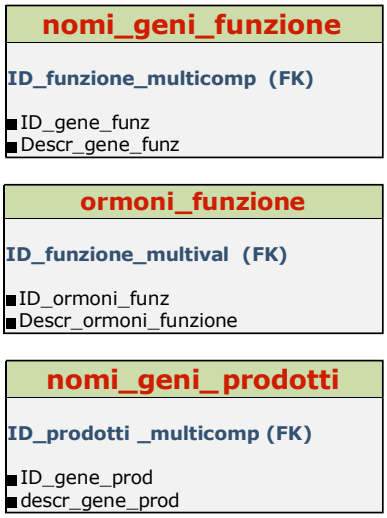
Per ogni attributo multivalore *A* si deve costruire una nuova relazione *R*. Questa nuova relazione comprenderà un attributo corrispondente ad *A*, più l'attributo di chiave primaria *K* (come chiave esterna della nuova relazione *R*) della relazione che rappresenta il tipo di entità che ha *A* come attributo.

Se l'attributo multivalo è composto, si considerano le sue componenti semplici.

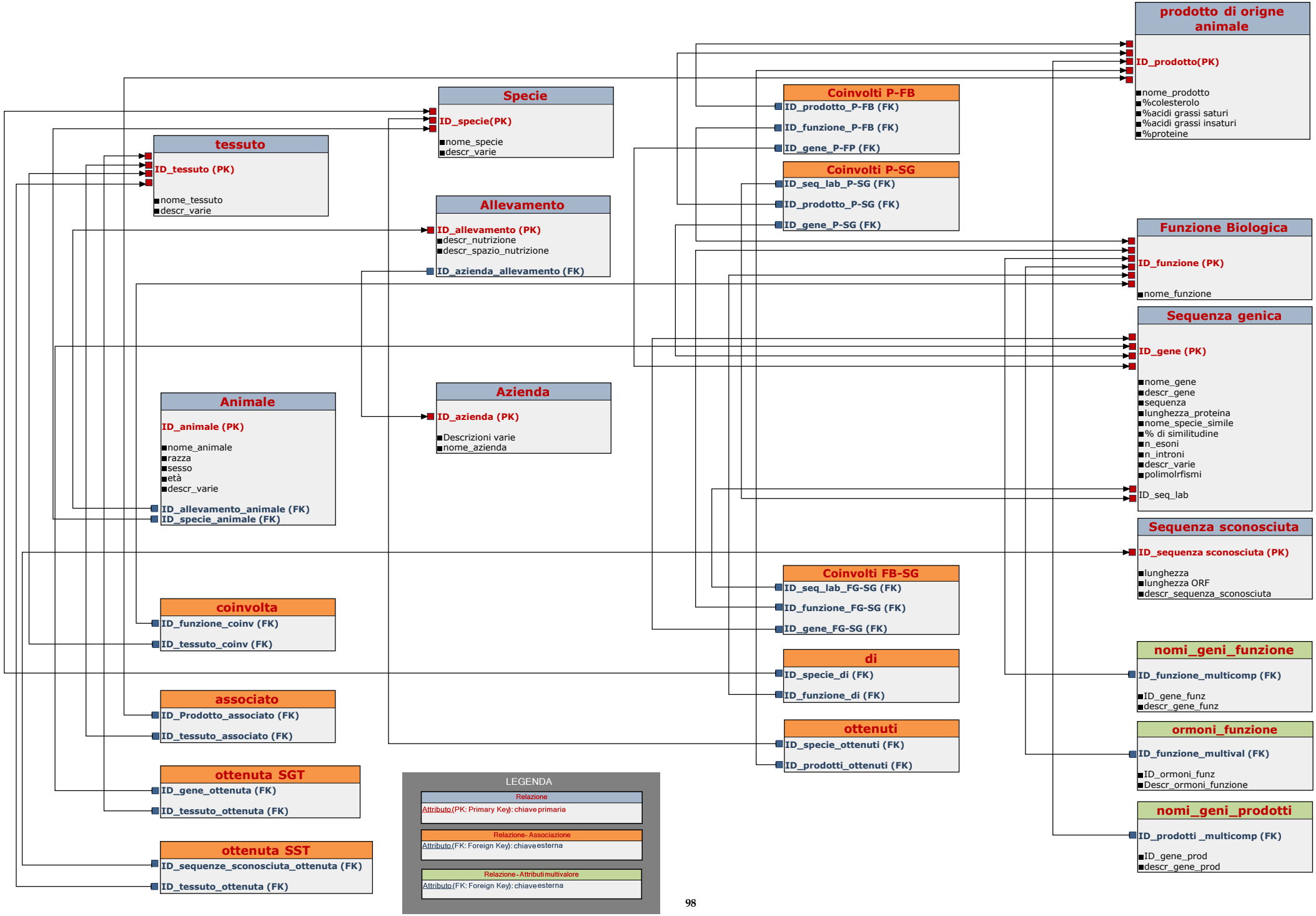
Nel nostro caso l'unico attributo multivalore è [ormoni] per il tipo di Entità **Funzione biologiche**, mentre abbiamo due attributi multivalore-composto, uno per il tipo di Entità **Prodotto di origine animale** [nomi_geni_prod] con le sue componenti semplici [ID_gene_prod] [Descr_gene_prod] e uno per il tipo di Entità **Funzione biologica** [nomi_geni_funz] con le sue componenti semplici [ID_gene_funz] [Descr_gene_funz].

Ridenominiamo le chiavi primarie rispettivamente in:

- [ID_funzione_multicomp];
- [ID_funzione_multival];
- [ID_prodotti_multicomp].



Mapping del modello dei dati (*traduzione dal modello concettuale al modello relazionale)*



3.2.6 Codifica SQL- definizione delle tabelle

```
create table TESSUTO (
     ID_tessuto char(20) not null,
     nome_tessuto char(20) not null,
     descr_varie char (200)not null,
primary key (ID_tessuto),
)
create table ANIMALE(
     ID_animale char(20) not null,
     nome_animale char(20) not null,
     razza char(20) not null,
     sesso char (1) not null,
     età numeric(4)not null,
     descr_varie char(200)not null,
     ID_allevamento_animale char(20) not null,
     ID_specie_animale char(20) not null,
primary key (ID_animale),
foreign key (ID_allevamento_animale) references ALLEVAMENTO(ID_allevamento),
foreign key (ID_specie_animale) references SEQUENZA(ID_gene),
)
create table COINVOLTA(
     ID_funzione_coinv char(20) not null,
     ID_tessuto_coinv char(20) not null,
primary key (ID_funzione_coinv),
primary key (ID_tessuto_coinv),
foreign key (ID_funzione_coinv) references FUNZIONE BIOLOGICA(ID_funzione),
foreign key (ID_tessuto_coinv) references TESSUTO(ID_tessuto),
)
create table ASSOCIATO(
     ID_prodotto_associato char(20) not null,
     ID_tessuto_associato char(20) not null,
```

```
primary key (ID_prodotto_associato),
primary key (ID_tessuto_associato),
foreign key (ID_prodotto_associato) references PRODOTTO DI ORIGINE
ANIMALE(ID_prodotto),
foreign key (ID_tessuto_associato) references TESSUTO(ID_tessuto),
)

create table OTTENUTA SGT(
     ID_gene_ottenuta char(20) not null,
     ID_tessuto_ottenuta char(20) not null,
primary key (ID_gene_ottenuta),
primary key (ID_tessuto_ottenuta),
foreign key (ID_gene_ottenuta) references SPECIE(ID_specie),
foreign key (ID_tessuto_ottenuta) references TESSUTO(ID_tessuto),
)

create table OTTENUTA SST(
     ID_sequenze_sconosciuta_ottenuta char(20) not null,
     ID_tessuto_ottenuta char(20) not null,
primary key (ID_sequenze_sconosciuta_ottenuta),
primary key (ID_tessuto_ottenuta),
foreign key (ID_sequenze_sconosciuta_ottenuta) references SEQUENZE
SCONOSCIUTE(ID_sequenze_sconosciute),
foreign key (ID_tessuto_ottenuta) references TESSUTO(ID_tessuto),
)

create table SPECIE (
     ID_specie char(20) not null,
     nome_specie char(20) not null,
     descr_varie char (200)not null,
primary key (ID_specie),
)
```

```
create table ALLEVAMENTO (
     ID_allevamento char(20) not null,
     descr_nutrizione char (200)not null,
     descr_spazio_nutrizione char (200)not null,
     ID_azienda_allevamento char(20),
primary key (ID_alleavamento),
foreign key (ID_azienda_allevamento) references AZIENDA(ID_azienda),
)

create table AZIENDA (
     ID_azienda char(20) not null,
     nome_azienda char (20) not null;
     descr_varie char (200)not null,
primary key (ID_azienda),
)

create table COINVOLTI P-FB (
     ID_prodotto_P-FB char(20) not null,
     ID_funzione_P-FB char(20) not null,
     ID_gene_P-FB char(20) not null,
primary key (ID_prodotto_P-FB),
primary key (ID_funzione_P-FB),
primary key (ID_gene_P-FB),
foreign key (ID_prodotto_P-FB) references PRODOTTO DI ORIGINE ANIMALE(ID_prodotto),
foreign key (ID_funzione_P-FB) references FUNZIONE BIOLOGICA(ID_funzione),
foreign key (ID_gene_P-FB) references SEQUENZA GENICA(ID_gene),
)

create table COINVOLTI P-SG (
     ID_seq_lab_P-SG boolean(1) not null,
     ID_prodotto_P-SG char(20) not null,
     ID_gene_P-SG char(20) not null,
```

```
primary key (ID_seq_lab_P-SG),
primary key (ID_prodotto_P-SG),
primary key (ID_gene_P-SG),
foreign key (ID_seq_lab_P-SG) references SEQUENZA GENICA(ID_seq_lab),
foreign key (ID_prodotto_P-SG) references PRODOTTO DI ORIGINE ANIMALE(ID_prodotto),
foreign key (ID_gene_P-SG) references SEQUENZA GENICA(ID_gene),
)

create table COINVOLTI FB-SG (
     ID_seq_lab_FB-SG boolean(1) not null,
     ID_funzione_FB-SG char(20) not null,
     ID_gene_FB-SG char(20) not null,
primary key (ID_seq_lab_FB-SG),
primary key (ID_funzione_FB-SG),
primary key (ID_gene_FB-SG),
foreign key (ID_seq_lab_FB-SG) references SEQUENZA GENICA(ID_seq_lab),
foreign key (ID_funzione_FB-SG) references FUNZIONE BIOLOGICA(ID_funzione),
foreign key (ID_gene_FB-SG) references SEQUENZA GENICA(ID_gene),
)

create table DI (
     ID_specie_di char(20),
     ID_funzione_di char(20),
primary key (ID_specie_di),
primary key (ID_funzione_di),
foreign key (ID_specie_di) references TESSUTO(ID_tessuto),
foreign key (ID_funzione_di) references FUNZIONE BIOLOGICA(ID_funzione),
)

create table PRODOTTO DI ORIGINE ANIMALE (
     ID_prodotto char(20) not null,
     nome_prodotto char (20) not null;
```

```
     %colesterolo decimal(10,3)not null,
     %acidi grassi saturi decimal(10,3)not null,
     %acidi grassi insaturi decimal(10,3)not null,
     %proteine decimal(10,3)not null;
primary key (ID_prodotto),
)

create table FUNZIONE BIOLOGICA (
     ID_funzione char(20) not null,
     nome_funzione char (20) not null;
primary key (ID_funzione),
)

create table SEQUENZA GENICA (
     ID_gene char(20) not null,
     nome_gene char (20) not null,
     descr_gene char(200)not null,
     sequenza char(300)not null,
     lunghezza_proteina int(5)not null,
     nome_specie_simile char (20) not null;
     %similitudine decimal(10,3)not null,
     n_esoni int(5)not null,
     n_introni int(5)not null,
     descr_varia char(200)not null,
     polimorfismi char(200)not null,
     ID_seq_lab boolean (1)not null,
primary key (ID_gene),
primary key (ID_seq_lab),
)

create table SEQUENZA SCONOSCIUTA (
     ID_sequenza_sconosciuta char(20) not null,
```

```
     lunghezza int(5)not null,
     lunghezzaORF int(5)not null,
     descr_ sequenza_sconosciuta char(200)not null,
primary key (sequenza_sconosciuta),
)

create table NOMI_GENI_FUNZIONE (
     ID_funzione_multicomp char(20) not null,
     ID_gene_funz char(20) not null,
     descr_gene_funz char (200) not null,
 primary key (ID_funzione_multicomp),
 foreign key (ID_funzione_multicomp) references FUNZIONE BIOLOGICA(ID_funzione),
)

create table ORMONI (
     ID_funzione_multival char(20) not null,
     ID_ormoni_funz char(20) not null,
     descr_ormoni_funz char (200) not null,
 primary key (ID_ funzione_multival),
 foreign key (ID_ funzione_multival) references FUNZIONE BIOLOGICA(ID_funzione),
)

create table NOMI_GENI_PRODOTTI(
     ID_prodotti_multicomp char(20) not null,
     ID_gene_prod char(20) not null,
     descr_gene_prod char (200) not null,
 primary key (ID_prodotti_multicomp),
 foreign key (ID_prodotti_multicomp) references PRODOTTI DI ORIGINE
ANIMALE(ID_prodotto),
```

4. Discussione

La valorizzazione di prodotti agroalimentari è attualmente oggetto di un notevole interesse proveniente da diversi settori socio-economici e culturali.

E' ormai chiaro infatti che la valorizzazione ed il potenziamento delle produzioni agroalimentari ed in particolare di prodotti tipici locali, può costituire fonte di sviluppo economico, turistico e culturale per la zona di produzione e può portare alla creazione di indotti con ulteriori ricadute. Non è peraltro secondario il ruolo che può avere nella salvaguardia della salute pubblica la produzione di prodotti alimentari che siano rispettosi di importanti parametri nutrizionali e che nel contempo tengano conto della sostenibilità ambientale.

In questo quadro è importante che progetti scientifici volti alla valorizzazione di prodotti agroalimentari, si basino su approcci di ampio respiro e che tengano conto di tutti i possibili strumenti tecnologici, scientifici e culturali attualmente disponibili. Il lavoro esposto in questa tesi è parte di un progetto multisciplinare volto alla valorizzazione della carne del bufalo campano.

Questo lavoro è stato svolto in collaborazione con un gruppo di lavoro che si occupa di genomica funzionale che è interessato fra l'altro alla caratterizzazione molecolare del tessuto muscolare bufalino. Tale caratterizzazione rappresenta un indispensabile strumento per lo studio della biologia di tale tessuto e per lo sviluppo di strumenti tecnologici utili per

progettare e monitorare sperimentazioni in campo. Un esempio di un simile strumento è un microarray di trascritti muscolo specifici di bufalo da cimentare con RNA estratti da muscoli provenienti da animali cresciuti in differenti condizioni di allevamento (diete differenti, spazio disponibile etc) o di diversa età e sesso. Inoltre un simile strumento è utile anche per valutare da un punto di vista molecolare fine le principali differenze con altre specie produttive quali ad esempio altri bovini ed ovini.

In questo contesto, l'utilizzo degli strumenti informatici ha svolto un ruolo specifico nell'ottenimento di informazioni genomiche esistenti per altre specie vicine (bue, pecora, uomo etc) e nel loro trasferimento allo studio della specie bufalina per la quale esistono ancora poche informazioni. Come già visto dal lavoro svolto su altre specie da allevamento, (ref. J marshall Graves), l'uso degli strumenti informatici applicati alla genomica comparata e l'interrogazione mirata, attraverso la rete internet, dei siti di collezione ed organizzazione dei dati molecolari e biologici esistenti, ha consentito di avviare studi di genomica della specie bufalina che possono allargare le possibilità di studio di specie di interesse economico.

Nel corso di questo lavoro è però emersa l'importanza di organizzare i dati, talvolta disparati, prodotti in laboratorio in maniera quanto più efficiente possibile, per poterne così estrarre il massimo valore informativo. Inoltre si è evidenziato che il lavoro fatto in laboratorio sulla specie bufalina può essere notevolmente arricchito se continuamente confrontato con le informazioni

esistenti per altre specie di interesse che possono essere molto informative quali bue, pecora, uomo, topo etc. In particolare è utile poter confrontare le caratteristiche di prodotti bufalini, quali ad esempio la carne o il latte, con gli stessi prodotti in altri bovini o ovini etc .

In una visione più allargata, può essere estremamente informativo confrontare funzioni biologiche quali la riproduzione o il metabolismo o la risposta immunitaria con quanto noto in altre specie di mammifero per le quali magari questi studi sono più avanzati o comunque possono fornire spunti interessanti. Le informazioni ottenute da questo tipo di confronti, se opportunamente organizzate e potenzialmente collegabili fra loro, può produrre ulteriori informazioni o spunti per analisi successive

Una possibile soluzione a questa problematica può derivare dalla messa a punto di un sistema di catalogazione su base informatica. Queste considerazioni hanno fornito lo spunto per il lavoro che costituisce la seconda parte dei risultati esposti in questa tesi. Tali risultati riguardano infatti il lavoro di progettazione di un database che consenta la catalogazione dei dati prodotti in laboratorio e dei dati scientifici d'interesse ricavabili dalla letteratura e/o dal web, consentendone oltre ad una facile consultazione anche il confronto dinamico. Questo progetto è partito prendendo in considerazione principalmente i dati molecolari relativi alla specie bufalina ma nel corso del lavoro, l'analisi delle specifiche esigenze e la dissezione dei requisiti hanno portato alla formulazione di un progetto di database che possa in qualche

modo risultare più versatile. Infatti è stato organizzato in modo da poter contenere dati relativi anche ad altre specie animali e non solo quella bufalina, e con la possibilità di relazionare i dati relativi agli aspetti di interesse molecolare con quelli fisiologici, di allevamento, aziendali etc.

Il prosieguo del lavoro qui esposto dovrà essere rappresentato dallla progettazione fisica della basi di dati, ossia il processo di scelta delle strutture di memorizzazione e di accesso ai file della base di dati al fine di garantire buone prestazioni in relazione alle varie applicazioni. Completata l'attività relativa alla progettazione fisica, si passa all'implementazione della basi di dati; di solito questo compito tocca ai programmatori in collaborazione con i progettisti della base di dati. Le istruzioni del linguaggio nel DDL vengono compilate e usate per creare schemi e query d'interrogazione. Successisamente si passa alla popolazione delle basi di dati o alla migrazione qualora i dati già esistano su un'altra base.

Anche il lavoro bioinformatico potrà avere un ulteriore sviluppo passando ad un approccio su larga scala, ovvero l'analisi dei prodotti di sequenziamento di cDNAteche tessuto-specifiche. In tale tipo di lavoro è infatti necessario attribuire alla sequenza ottenuta in laboratorio una possibile identità e questo lo si può fare interrogando le banche dati genomiche per valutare la "somiglianza" della sequenza ignota con altre già caratterizzate e ottenerne quindi una possibile identificazione. Laddove la sequenza non trovi riscontro con altre note è comunque possibile usare dei tools bioinformatici che

diano almeno qualche informazione parziale (potenziale codifica di una proteina, esistenza di particolari strutture secondarie, esistenza di domini strutturali e funzionali etc.).

Riferimenti Bibliografici

[1] Pinney DF, Pearson-White SH, Konieczny SF, Latham KE, Emerson CP Jr. *Myogenic lineage determination and differentiation: evidence for a regulatory gene pathway.* Cell. 1988 Jun 3;53(5):781-93.

[2] Tapscott SJ.*The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription*. Development. 2005 Jun;132(12):2685-95. Review.

[3] Buckingham M, Bajard L, Chang T, Daubas P, Hadchouel J, Meilhac S, Montarras D, Rocancourt D, Relaix F. *The formation of skeletal muscle: from somite to limb*. J Anat. 2003 Jan;202(1):59-68. Review.

[4] Bellinge RH, Liberles DA, Iaschi SP, O'brien PA, Tay GK. *Myostatin and its implications on animal breeding*: a review. Anim Genet. 2005 Feb;36(1):1-6

[5] MV Olson *The Human Genome Project PNAS*, May 1993; 90: 4338.

[6] J. Craig Venter, et al Science, *The Sequence of the Human Genome* Feb 2001; 291: 1304.

[7] Bornholdt S. *Modeling genetic networks and their evolution: a complex dynamical systems perspective.* Biol Chem. 2001 Sep;382(9):1289-99. Review.

[8] O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA. *The promise of comparative genomics in mammals*.

[9] Kadarmideen HN, von Rohr P, Janss LL. *From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding*. Mamm Genome. 2006 Jun;17(6):548-64. Epub 2006 Jun 12. Review.

[10] Huang H, Hu ZZ, Arighi CN, Wu CH. *Integration of bioinformatics resources for functional analysis of gene expression and proteomic data*. Front Biosci. 2007 Sep 1;12:5071-88. Review.

[11] Comicioli V. *Biomatematica* – Apogeo Editore.

[12] Valle G., Helmer Citterich M., Attimonelli M., Pesole G. *Introduzione alla Bioinformatica* - Zanichelli Editore.

[13] Tramontano A. *Bioinformatica - Zanichelli Editore.*

[14] S B Needleman , C D Wunsch: *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. J Mol Biol. 1970 Mar ;48 (3):443-53 5420325 [Cited: 463]

[15] Smith TF, Waterman MS (1981). *Identification of Common Molecular Subsequences*. Journal of Molecular Biology 147: 195-197

[16] Dayhoff, M. O. 1978. *Survey of new data and computer methods of analysis*. In M. O. Dayhoff, ed., Atlas of Protein Sequence and Structure, vol. 5, supp. 3, pp. 29, National Biomedical Research Foundation, Silver

Springs, Maryland.

[17] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, "*Basic Local Alignment Search Tool*", 1990.

[18] Pearson W.R. and Lipman D.J. (1988): *Improved tools for biological sequence comparison*. Proc. Natl. Acad. Sci. USA 85, 2444-2448.

[19] Vitacolonna N. *Allineamento di coppie di sequenze* – Univesità di Udine – http://www.dimi.uniud.it/~vitacolonna.

[20] Vitacolonna N. *Allineamento multiplo di sequenze* – Univesità di Udine – http://www.dimi.uniud.it/~vitacolonna.

[21] Lesk A.M. *Introduzione alla Bioinformatica* – McGraw-Hill Editore;

[22] D. F. Feng e R. F. Doolittle. *Progressive sequence alignment as a prerequisite to correct phylogenetic trees*. Journal of Molecular Evolution , 25:351–360, 1987.

[23] J. D. Thompson, D. G. Higgins e T. J. Gibson. CLUSTAL W*: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice*. Nucleic Acid Research , 22:4673–4680, 1994.

[24] P.Atzeni, S. Ceri, S. Paraboschi, *Basi di dati. Modelli e linguaggi di interrogazione* - McGraw-Hill

[25] Ramez A. Elmasri , Shamkant B. Navathe *Fundamentals of Databases Systems* - The Benjamin/Cummings Pubblishing Company Inc.

[26] E.F. Codd - *A Relational Model of Data for Large Shared Data Banks*,

Communications of the ACM 13, No. 6, June 1970

[27] Libro per la PCR.

[28] Wright WE, Sassoon DA, Lin VK. *Myogenin, a factor regulating myogenesis, has a domain homologous to MyoD*. Cell. 1989 Feb 24;56(4):607-17

[29] Gunning PW, Schevzov G, Kee AJ, Hardeman EC. *Tropomyosin isoforms: divining rods for actin cytoskeleton function*. Trends Cell Biol. 2005 Jun;15(6):333-41. Review

[30] Squire JM, Morris EP. *A new look at thin filament regulation in vertebrate skeletal muscle*. FASEB J. 1998 Jul;12(10):761-71. Review. Erratum in: FASEB J 1998 Sep;12(12):1252.

[31] Craig R, Woodhead JL. *Structure and function of myosin filaments*. Curr Opin Struct Biol. 2006 Apr;16(2):204-12. Epub 2006 Mar 24. Review.

[32] Ordway GA, Garry DJ. *Myoglobin: an essential hemoprotein in striated muscle*. J Exp Biol. 2004 Sep;207(Pt 20):3441-6. Review.