ISSN online: 2282-5665

Consiglio Nazionale delle Ricerche

CERIS ISTITUTO DI RICERCA SULL'IMPRESA E LO SVILUPPO

Ottobre 2006

Rapporto tecnico N.01

Progettazione di un database per la gestione delle analisidi biologia molecolare: MLO

Giancarlo Birello, Ivano Fucile, Valter Giovanetti

RAPPPORTO TECNICO CERIS-CNR
Anno 1, N° 1 – ottobre 2006

Direttore Responsabile
Secondo Rolfo

Direzione e Redazione
Ceris-Cnr
Istituto di Ricerca sull'Impresa e lo Sviluppo
Via Real Collegio, 30
10024 Moncalieri (Torino), Italy
Tel. +39 011 6824.911
Fax +39 011 6824.966
segreteria@ceris.cnr.it
http://www.ceris.cnr.it

Sede di Roma
Via dei Taurini, 19
00185 Roma, Italy
Tel. 06 49937810
Fax 06 49937884

Sede di Milano
Via Bassini, 15
20121 Milano, Italy
tel. 02 23699501
Fax 02 23699530

Segreteria di redazione
Maria Zittino
m.zittino@ceris.cnr.it

Progettazione di un database per la gestione delle analisi di biologia molecolare: MLO

[Planning of a database for the management of the analyses of molecular Biology: MLO]

Giancarlo Birello, Ivano Fucile, Valter Giovanetti
(Ceris-Cnr-Ufficio IT)

Ceris-Cnr
Ufficio IT
Strada delle Cacce, 79
10100 Torino – Italy
Tel.: 011 3977303/388/512
Autore corrispondente: Giancarlo Birello, G.Birello@ceris.cnr.it

ABSTRACT. The goal of the plan "MLO" is to offer a computer science solution, based on the creation, organisation and management of a database built with data coming from the analyses of molecular Biology. We think of facilitating the management and the use of data. This proposal will focus on the following points: − Client/server infrastructure − Input of the analysis data − Identifying management − data research.
The *IT Office* of the Turin Area of Research (National Council of Research) will carry out this plan, and will deal with the existing data and their migration to the new system and every necessary operation.

I N D I C E

1. OBIETTIVI DEL PROGETTO 5
2. ELENCO DEI REQUISITI 5
2.1 Infrastruttura 5
2.2 Lato server 5
2.3 Lato client 6
3. DESCRIZIONE DELLA SOLUZIONE 6
3.1 Descrizione dell'architettura centrale 6
3.2 Descrizione delle postazioni utente 6
3.3 Modello dati 7
3.4 Moduli gestione dati 12
3.5 Moduli consultazione dati 13
3.6 Backup 13
3.7 Sicurezza 13
4. MESSA IN FUNZIONE E ADDESTRAMENTO 14
5. MANUTENZIONE 14
6. COSTI E TEMPI 14
7. CONCLUSIONI 15
ACRONIMI 15
INDICE DELLE FIGURE 16

OBIETTIVI DEL PROGETTO

L'obiettivo del progetto "MLO" è offrire una soluzione informatica, basata su un database centrale, per la gestione dei dati provenienti dalle analisi di biologia molecolare, facilitandone la gestione ed il successivo utilizzo.

La proposta gestirà i seguenti punti:

- Infrastruttura client/server
- Inserimento dati analisi
- Gestione anagrafiche
- Ricerche dati

Il progetto, qualora approvato, sarà realizzato dall'Ufficio IT dell'Area di Ricerca, che provvederà a fornire al gruppo di ricerca il prodotto finito. Nella fornitura è inclusa la migrazione dei dati esistenti nel nuovo database centrale e la configurazione dei server che ospiteranno gli elementi necessari al funzionamento del MLO.

1. ELENCO DEI REQUISITI

1.1 Infrastruttura

Per il funzionamento di un sistema *client/server* come quello proposto è necessaria un'infrastruttura di rete affidabile e di buone prestazioni.

I laboratori del gruppo di biologia molecolare sono connessi in fibra ottica a 100 Mbit ai server centrali dell'Ufficio IT, infrastruttura più che sufficiente al buon funzionamento del prodotto.

1.2 Lato server

I server centrali dovranno ospitare il database e le pagine web dinamiche di accesso ai dati. Presso l'Ufficio IT sono già disponibili un paio di server adatti allo scopo.

Un primo server ha installato il SW MS-SQL2000, necessario per ospitare il database MLO. Presenta tutte le caratteristiche richieste dal progetto quali affidabilità, sicurezza e buone prestazioni in rapporto alla quantità di dati che saranno trattati.

Un secondo server fornisce servizi web ed è visibile anche dall'esterno della LAN. È in grado di accedere ed interrogare il database per offrire pagine web dinamiche con i dati di ricerche all'interno del MLO.

Entrambe le macchine sono dotate dell'HW opportuno e dimensionato per le prestazioni richieste.

1.3 Lato client

Sui PC degli utenti, in particolare degli addetti alla gestione del database, dovrà essere installata la versione 2003 del prodotto MS-Access, incluso nella suite MS-Office. Tale versione è richiesta per compatibilità con il SW lato server e con le maschere per l'inserimento dei dati che dovranno essere in grado di interagire direttamente col server.

2. DESCRIZIONE DELLA SOLUZIONE

Questo capitolo contiene la descrizione della soluzione prescelta che dovrà essere realizzata per soddisfare tutti gli obiettivi del progetto.

2.1 Descrizione dell'architettura centrale

SERVER DATABASE

Il server che ospiterà il database è posizionato all'interno della LAN ed è inserito nel dominio AD TOCNR, lo stesso in cui sono registrati tutti gli account utente. In questo modo e grazie al prodotto MS-SQL2000 è possibile controllare in modo capillare l'accesso ai dati, attribuendo gli opportuni permessi di modifica/inserimento solo agli addetti ai lavori.

SERVER WEB

Il SERVER WEB è inserito invece nella DMZ, zona controllata dal *firewall* ed accessibile pubblicamente dall'esterno. È anch'esso inserito nel dominio TOCNR e quindi in grado di interrogare il server del database. In questo modo le pagine web per la ricerca di dati all'interno del MLO potranno essere ulteriormente personalizzate, fornendo scelte e risultati differenti secondo l'utente che li richiede.

2.2 Descrizione delle postazioni utente

Il personale addetto all'inserimento dei dati potrà compiere tutte le operazioni di gestione del database da una qualsiasi postazione all'interno e solo all'interno della LAN, purché il PC soddisfi le seguenti caratteristiche:

- connessione di rete e logon al dominio
- MS-Access 2003

Chiunque potrà invece accedere alle pagine web di ricerca purché dotato di un PC connesso ad internet e di un browser web.

2.3 Modello dati

Il modello dati è stato derivato dal database ora in uso presso il gruppo di ricerca, apportando gli aggiornamenti che gli addetti hanno suggerito sulla base della loro esperienza d'uso.

TABELLE

Il modello dati si compone di 23 tabelle relazionate tra loro:

- Campioni
- DB
- Enzimi
- Famiglie Piante
- File PCR
- Generi Piante
- Gruppi Risultati
- Metodi DB
- Metodi Estr
- Operatori
- PCR
- PCR Risultati
- Piante
- Primers
- Provenienze
- Regioni
- RFLP
- Risultati DB
- Sintomi
- Soggetti
- Sonde
- Valutazioni
- Valutazioni finali

Le figure seguenti mostrano i vari campi contenuti nei record di ognuna delle tabelle. Sono anche evidenziate le chiavi primarie utilizzate nelle relazioni.

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
	[ID Campione]	int	4	
	[Anno Campione]	smallint	2	
	[ID Provenienza]	int	4	✓
	[Codifica Mittente]	nvarchar	50	✓
	[ID Pianta]	int	4	✓
	[Data di Arrivo]	datetime	8	✓
	[ID Sintomo]	int	4	✓
	[ID Valutazione]	smallint	2	✓
	[ID Valutazione Finale]	smallint	2	✓
	[ID Soggetto]	int	4	✓
	Note	ntext	16	✓
	[ID Metodo Estr]	int	4	✓
	[ID Operatore]	int	4	✓
	[Esito comunicato]	bit	1	✓
	[Data comunicazione]	datetime	8	✓
🔑	IDC	int	4	

Fig. 3.1 – Tabella "Campioni"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
	[Cod DB]	nvarchar	6	✓
	[ID Sonda]	int	4	
	[ID Metodo DB]	int	4	
🔑	[ID Cod DB]	int	4	

Fig. 3.2 – Tabella "DB"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Enzima]	int	4	
	[Descr Enzima]	nvarchar	50	✓

Fig. 3.3 – Tabella "Enzimi"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Famiglia Pianta]	int	4	
	[Descr Famiglia Pianta]	nvarchar	40	✓

Fig. 3.4 – Tabella "Famiglie Piante"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID File PCR]	int	4	
	[Descr File PCR]	nvarchar	50	✓
	TA	smallint	2	✓

Fig. 3.5 – Tabella "File PCR"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Genere Pianta]	int	4	
	[Descr Genere Pianta]	nvarchar	50	✓

Fig. 3.6 – Tabella "Generi Piante"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Gruppo Risultati]	int	4	
	[Descr Gruppo Risultati]	nvarchar	50	✓

Fig. 3.7 – Tabella "Gruppi Risultati"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Metodo DB]	int	4	
	[Descr Metodo DB]	nvarchar	50	✓

Fig. 3.8 – Tabella "Metodi DB"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Metodo Estr]	int	4	
	[Descr Metodo Estr]	nvarchar	50	✓
	[Volume finale]	smallint	2	✓

Fig. 3.9 – Tabella "Metodi Estr"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Operatore]	int	4	
	[Descr Operatore]	nvarchar	50	✓

Fig. 3.10 – Tabella "Operatori"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
	[Cod PCR]	nvarchar	7	✓
	[ID Primers]	int	4	
	[ID File PCR]	int	4	
	[Cod PCR Nested]	nvarchar	7	✓
🔑	[ID Cod PCR]	int	4	
	[ID Cod PCR nested]	int	4	✓

Fig. 3.11 – Tabella "PCR"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Pianta]	int	4	
	[ID Famiglia Pianta]	int	4	✓
	[ID Genere Pianta]	int	4	✓
	[Specie Pianta]	nvarchar	20	✓
	[Nome Comune]	nvarchar	20	✓
	[Sigla Pianta]	nvarchar	4	✓
	Serra	bit	1	✓

Fig. 3.12 – Tabella "Piante"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Primers]	int	4	
	[Descr Primers]	nvarchar	50	✓
	[Quantità Mix]	smallint	2	✓

Fig. 3.13 – Tabella "Primers"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Provenienza]	int	4	
	[Descr Provenienza]	nvarchar	50	✓
	[Descr Riferimento]	nvarchar	50	✓
	[ID Regione]	int	4	✓
	Laboratorio	bit	1	✓

Fig. 3.14 – Tabella "Provenienze"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Regione]	int	4	
	[Descr Regione]	nvarchar	50	✓
	[Nazione italia]	bit	1	✓

Fig. 3.15 – Tabella "Regioni"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID RFLP]	int	4	
	[ID Risultato PCR]	int	4	
	[ID Enzima]	int	4	✓
	[ID Gruppo Risultati]	int	4	✓

Fig. 3.16 – Tabella "RFLP"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Risultato DB]	int	4	
	[ID Campione]	int	4	✓
	[Anno Campione]	smallint	2	✓
	[Risultato DB]	bit	1	✓
	[Cod DB]	nvarchar	6	✓
	[Risultato DB - Simbolico]	nvarchar	5	✓
	[Sicurezza Dati]	bit	1	✓
	[ID Cod DB]	int	4	✓
	IDC	int	4	✓

Fig. 3.17 – Tabella "Risultati DB"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Risultato PCR]	int	4	
	IDC	int	4	
	[ID Cod PCR]	int	4	
	[numero tubo]	smallint	2	✓
	[lettera tubo]	char	1	✓
	Risultato	bit	1	✓
	Sicuro	bit	1	✓
	[Qta Campione]	char	8	✓
	Inibisce	bit	1	✓

Fig. 3.18 – Tabella "PCR Risultati"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Sintomo]	int	4	
	[Descr Sintomo]	nvarchar	50	✓

Fig. 3.19 – Tabella "Sintomi"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Soggetto]	int	4	
	[Descr Soggetto]	nvarchar	50	✓

Fig. 3.20 – Tabella "Soggetti"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Sonda]	int	4	
	[Descr Sonda]	nvarchar	50	✓

Fig. 3.21 – Tabella "Sonde"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Valutazione]	smallint	2	
	[Descr Valutazione]	nvarchar	50	✓

Fig. 3.22 – Tabella "Valutazioni"

	Nome colonna	Tipo di dati	.unghezza	Ammetti Null
🔑	[ID Valutazione Finale]	smallint	2	
	[Descr Valutazione Finale]	nvarchar	50	✓

Fig. 3.23 – Tabella "Valutazioni finali"

RELAZIONI

Per la natura di questo tipo di database (DB relazionale), le relazioni sono fondamentali per il corretto funzionamento delle operazioni d'inserimento ed estrazione dei dati.

Nelle figure seguenti sono illustrate le relazioni che intercorrono tra le varie tabelle dei dati.

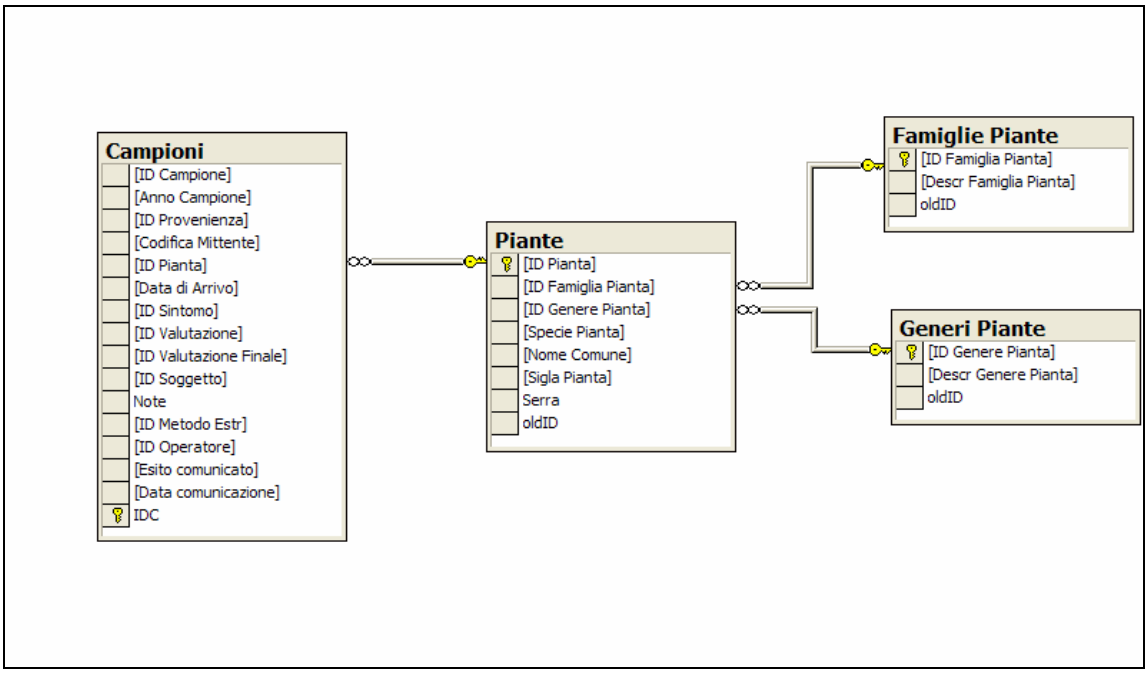

Fig. 3.24 – Diagramma "Cam/Pian/Fam/Gen"

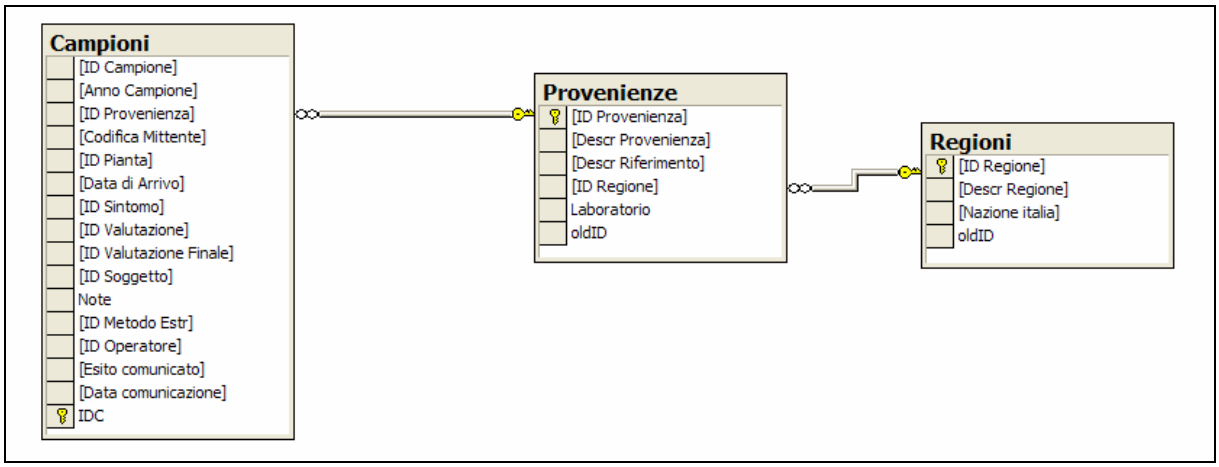

Fig. 3.25 – Diagramma "Camp/Prov/Reg"

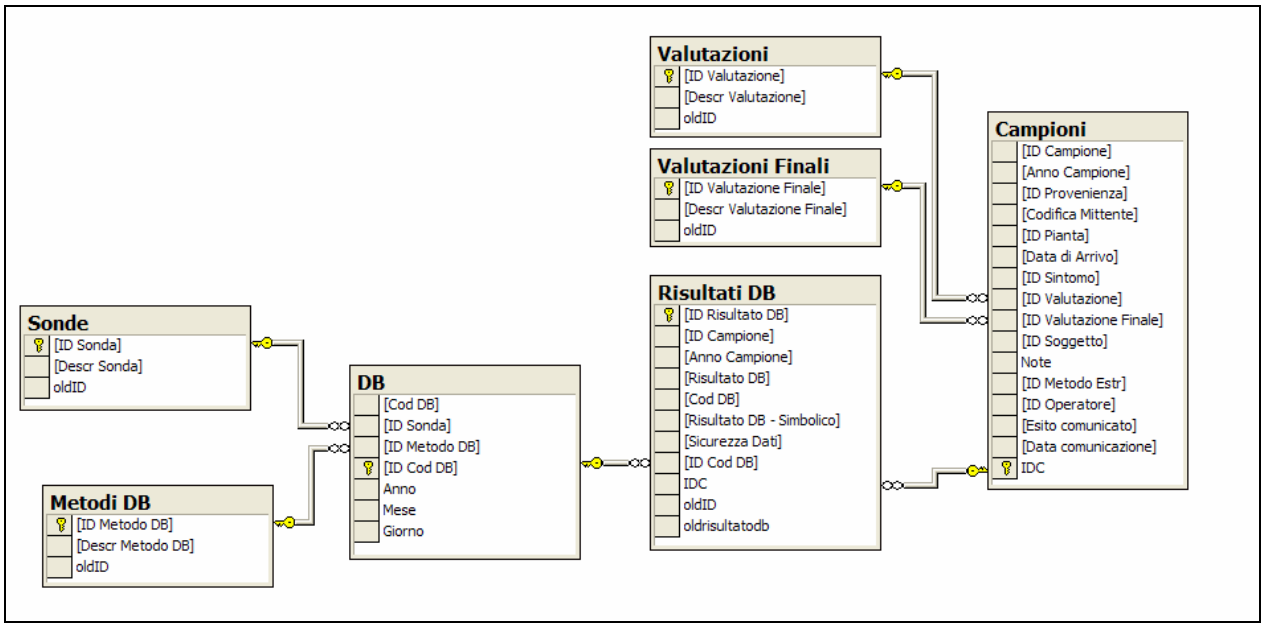

Fig. 3.26 – Diagramma "Camp/Ris/DB"

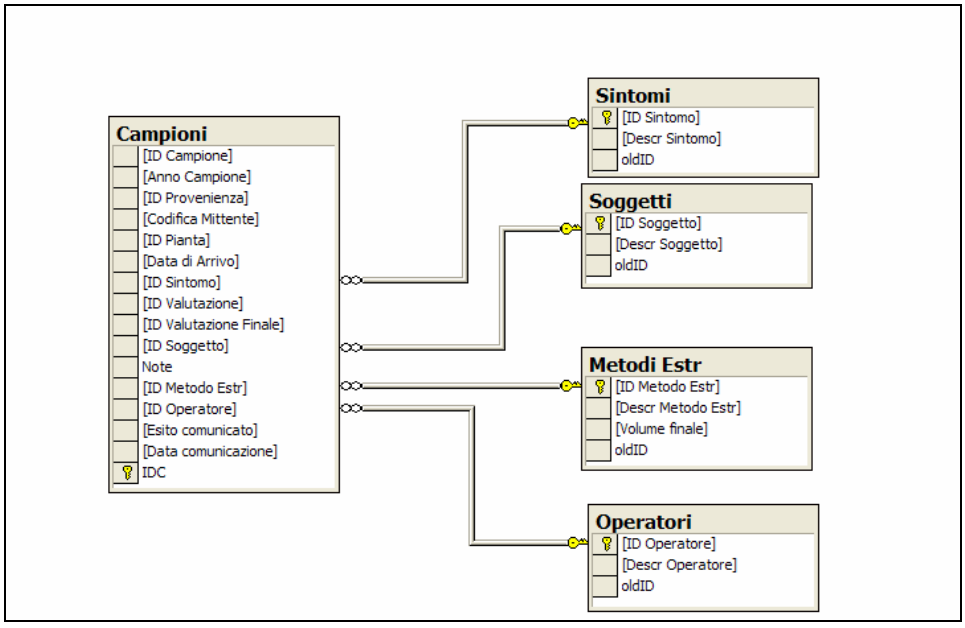

Fig. 3.27 – Diagramma "Camp/Sin/Sog/met/Ope"

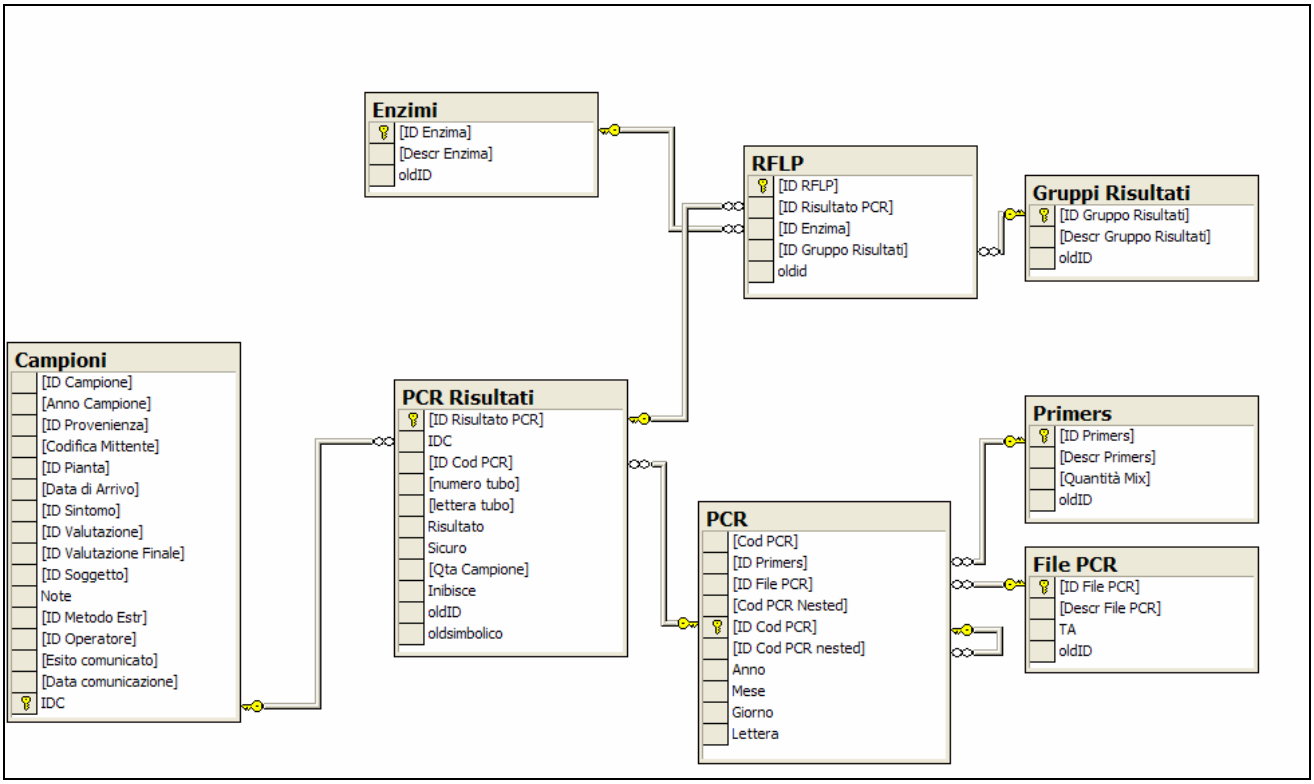

Fig. 3.28 – Diagramma "RFLP/PCR"

2.4 Moduli gestione dati

L'inserimento e la modifica dei dati sono riservati al personale del gruppo responsabile del database. Gli utenti abilitati avranno il controllo completo sui dati, ma non potranno modificare la struttura del database.

L'accesso alle maschere di gestione dei dati sarà possibile da un qualsiasi PC che soddisfi i requisiti indicati al punto 2.3.

Dalla maschera iniziale sarà possibile accedere alle seguenti procedure:

- gestione anagrafiche e tabelle di riferimento
- inserimento/modifica campioni
- gestione DB
- gestione PCR
- ricerche e *report* specifici riservati agli operatori

2.5 Moduli consultazione dati

Nella versione attuale del database, che risiede in un file locale e non su un server, non è possibile accedere ai dati se non direttamente dal PC dell'operatore. Grazie alla struttura client/server qui proposta, sarà invece possibile permettere l'accesso ai dati a chiunque connesso ad internet tramite semplici pagine web.

Sul SERVER WEB già descritto risiederanno le pagine web di ricerca dati all'interno del database MLO. Si potranno strutturare più livelli di ricerca, personalizzarli in base all'utente e controllare quali dati presentare, in funzione delle regole di privacy che il gruppo riterrà più opportune.

Le pagine web di ricerca dovranno essere inserite in un sottoweb indipendente dal resto del sito dell'Istituto ma direttamente raggiungibili da quest'ultimo.

2.6 Backup

Il DATABASE MLO, poiché risiede su uno dei server dell'Ufficio IT, è soggetto alle stesse procedure di *backup* e *disaster recovery* applicate agli altri server centrali. Lo stesso dicasi delle PAGINE WEB di ricerca poiché risiedono anch'esse su uno dei server dell'Ufficio IT.

In particolare la procedura prevede due *backup* indipendenti settimanali dei database e dei server. La settimana successiva avviene la sovrascrittura dei due *backup* precedenti.

I backup settimanali vengono effettuati su unità NAS per velocizzare le operazioni. Trimestralmente i dati sono memorizzati su nastri poi depositati in cassaforte in luogo distinto dalla sala macchine dell'Ufficio IT.

2.7 Sicurezza

La presenza di un *firewall* a livello d'infrastruttura di rete e la posizione dei server in DMZ e nella LAN garantiscono un elevato livello di sicurezza dei dati e riducono la possibilità di intrusioni che potrebbero rendere inattivi i server stessi.

L'accesso al database è permesso solo agli utenti autenticati nel dominio AD ed è possibile controllare in modo capillare quali utenti possono modificare i dati.

Questa soluzione permette individuare ruoli differenti tra gli stessi operatori, riservando ad alcune operazioni non permesse ad altri. Si potranno inoltre attribuire responsabilità differenti su singole tabelle del database MLO.

3. MESSA IN FUNZIONE E ADDESTRAMENTO

Il progetto MLO è l'evoluzione di un prodotto già esistente ed utilizzato attualmente. Nei limiti del possibile si cercherà di mantenere immutate le procedure di gestione dei dati per rendere minimi i tempi necessari all'entrata in funzione del nuovo prodotto.

Sarà necessario un certo tempo di inattività per permettere la migrazione dei dati nel nuovo database. Indicativamente l'operazione richiederà 2/3 giorni lavorativi.

Per le pagine web si prevede la loro messa in funzione in un tempo successivo all'entrata in funzione del database. Richiederanno un periodo di messa a punto maggiore poiché nuove rispetto al prodotto esistente.

Sui PC degli utenti non è necessaria alcuna installazione quindi sarà subito utilizzabile il nuovo prodotto appena trasferiti i dati nel database sul server.

4. MANUTENZIONE

La manutenzione ordinaria dei server rientra tra i compiti dell'Ufficio IT che ne garantisce il funzionamento e la stabilità. Rientrano quindi tra queste operazioni la manutenzione ordinaria del database e delle pagine web.

Sempre a carico dell'Ufficio IT sono anche gli interventi straordinari quali modifiche strutturali del database e modifica o aggiunte di nuove pagine web.

Si conferma la massima disponibilità dell'Ufficio IT nel considerare tutte le richieste di modifica/aggiornamento che gli saranno sottoposte dagli operatori durante l'utilizzo del prodotto.

5. COSTI E TEMPI

I costi lato server sono nulli in termini di HW e SW in quanto le apparecchiature, il SW di gestione del database e relative licenze, il server web e relative licenze sono già in possesso dell'Ufficio IT e come tali a disposizione degli Istituti afferenti all'Area di Ricerca.

I PC degli operatori sono già stati aggiornati alla versione MS-Office 2003 quindi non sono necessari ulteriori investimenti.

I mesi/uomo teorici per la realizzazione del progetto non sono molti, indicativamente 2 mesi/uomo potrebbero essere sufficienti.

L'Ufficio IT offre principalmente un servizio di help desk a tutta l'utenza piemontese

quindi l'attività è molto diversificata ed in generale non molto prevedibile. Non è possibile quindi dedicarsi a tempo pieno alla realizzazione del progetto, ciò significa un'estensione dei tempi di realizzazione a circa 6 mesi/uomo.

La realizzazione di questo progetto rientra tra le attività istituzionali che l'Ufficio IT offre agli utenti degli istituti afferenti. Non saranno richiesti contributi economici per i mesi/uomo che saranno impiegati nella realizzazione del progetto.

6. CONCLUSIONI

Grazie alla struttura attuale dei dati, il progetto non si presenta molto complesso. La fattibilità è favorita inoltre dalla presenza e dalla disponibilità di risorse centrali già operative.

Il progetto significa per il gruppo di biologia molecolare un'interessante evoluzione del database attuale, soprattutto per la potenziale visibilità esterna dei dati.

Per l'Ufficio IT rappresenta un'applicazione pratica delle risorse e conoscenze sviluppate in questi anni e un opportuno sfruttamento delle strutture esistenti.

Infine l'eventuale realizzazione del progetto sarebbe un importante momento di collaborazione tra settori molto differenti dello stesso ente.

ACRONIMI

Acronimo	*Definizione*
AD	Active Directory
DB	Dot Blot
DMZ	De-Militarized Zone
HW	Hardware
IT	Information Technology
LAN	Local Area Network
MLO	Micoplasm Like Organism
MS	Microsoft
NAS	Network Attached Storage
PC	Personal Computer
PCR	Polymerase Chain Reaction
SW	Software

INDICE DELLE FIGURE

Fig. 3.1 – Tabella "Campioni" 8
Fig. 3.2 – Tabella "DB" 8
Fig. 3.3 – Tabella "Enzimi" 8
Fig. 3.4 – Tabella "Famiglie Piante" 8
Fig. 3.5 – Tabella "File PCR" 8
Fig. 3.6 – Tabella "Generi Piante" 8
Fig. 3.7 – Tabella "Gruppi Risultati" 8
Fig. 3.8 – Tabella "Metodi DB" 8
Fig. 3.9 – Tabella "Metodi Estr" 8
Fig. 3.10 – Tabella "Operatori" 8
Fig. 3.11 – Tabella "PCR" 8
Fig. 3.12 – Tabella "Piante" 9
Fig. 3.13 – Tabella "Primers" 9
Fig. 3.14 – Tabella "Provenienze" 9
Fig. 3.15 – Tabella "Regioni" 9
Fig. 3.16 – Tabella "RFLP" 9
Fig. 3.17 – Tabella "Risultati DB" 9
Fig. 3.18 – Tabella "PCR Risultati" 9
Fig. 3.19 – Tabella "Sintomi" 9
Fig. 3.20 – Tabella "Soggetti" 9
Fig. 3.21 – Tabella "Sonde" 9
Fig. 3.22 – Tabella "Valutazioni" 9
Fig. 3.23 – Tabella "Valutazioni finali" 9
Fig. 3.24 – Diagramma "Cam/Pian/Fam/Gen" 10
Fig. 3.25 – Diagramma "Camp/Prov/Reg" 10
Fig. 3.26 – Diagramma "Camp/Ris/DB" 11
Fig. 3.27 – Diagramma "Camp/Sin/Sog/met/Ope" 11
Fig. 3.28 – Diagramma "RFLP/PCR" 12
}}