5. Ricerca di domini presenti in una sequenza

In questa attività utilizzeremo alcuni programmi basati su HMM che consentono di caratterizzare funzionalmente e strutturalmente una sequenza proteica attraverso l'identificazione di possibili domini e motivi presenti. Si noti come approcci dive rsi possano dare informazioni diverse e complementari. Come caso di studio prenderemo in esame la stessa sequenza utilizzata nell'esercitazione di previsione della struttura secondaria: **P98153**.

- Collegatevi al sito della banca dati **Pfam** (http://pfam.sanger.ac.uk). Ricordiamo che questa banca dati contiene allineamenti multipli di domini proteici caratterizzati funzionalmente e codificati sotto forma di HMM. Inserite nel campo *sequence search* la sequenza in formato FASTA e lanciate la ricerca (il risultato dovrebbe essere visibile in pochi minuti). La **Figura 1** riporta la schermata che restituisce Pfam. In posizione 29-66 della sequenza è stato riconosciuto il dominio del recettore delle lipoproteine a bassa densità di classe A.

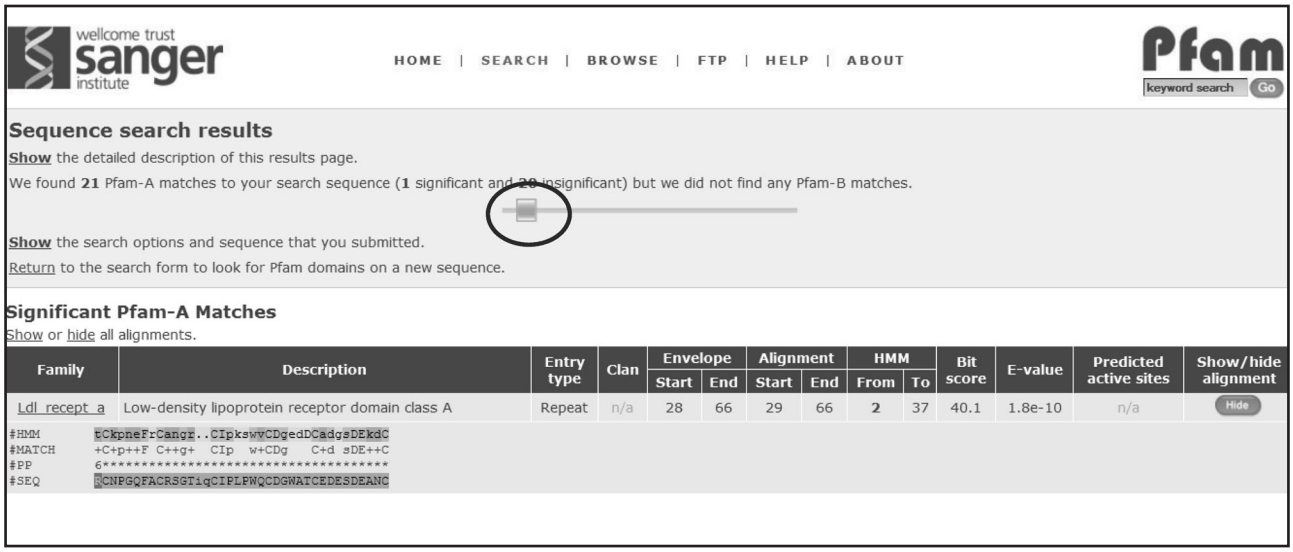
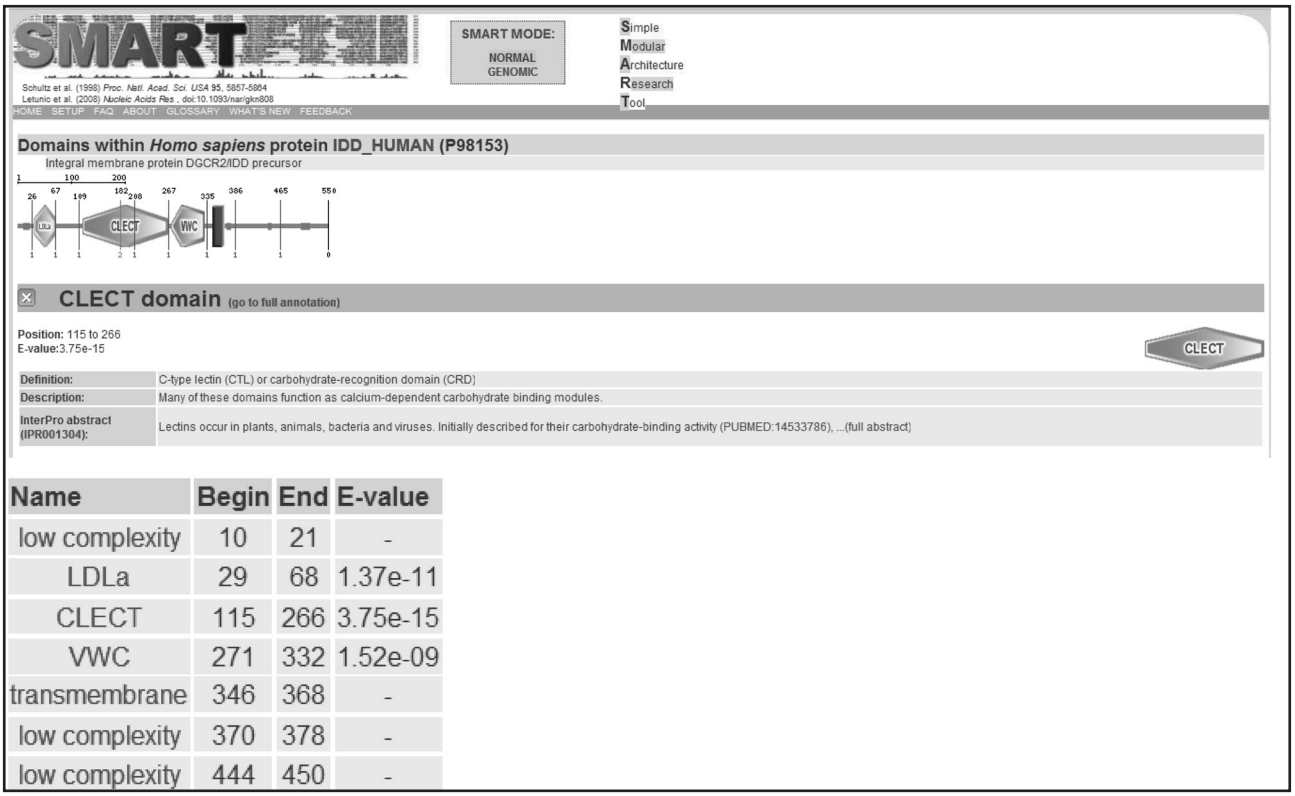


Figura 1 Risultato della ricerca in Pfam con la sequenza sonda. La presenza di un dominio identificato dal programma è evidenziata da un'icona rettangolare cliccabile posta all'interno dello schema della sequenza sonda (cerchio nella figura). In basso sono riportati i dettagli dei risultati della ricerca: in particolare l'E-value e l'allineamento tra la porzione della sequenza sonda e l'HMM del dominio.

- Un'ulteriore analisi può essere effettuata attraverso il sito **SMART** (http://smart.embl-heidelberg.de). Questa banca contiene una collezione di domini funzionalmente caratterizzati curata manualmente. Inserite la sequenza nel campo di immissione e lanciare la ricerca (il risultato dovrebbe essere visibile in pochi minuti). Nella schermata dei risultati viene riportata in forma schematica la sequenza con l'assegnazione dei domini riconosciuti (**Figura 2**).



Name	Begin	End	E-value
low complexity	10	21	-
LDLa	29	68	1.37e-11
CLECT	115	266	3.75e-15
VWC	271	332	1.52e-09
transmembrane	346	368	-
low complexity	370	378	-
low complexity	444	450	-

Figura 2 In alto: schermata con i risultati trovati da SMART. Sullo schema della sequenza sonda sono disegnate le icone che rappresentano i domini identificati; ciascuna icona è attivabile attraverso puntatore del mouse per visualizzare informazioni sul dominio relativo. In basso: elenco dei domini trovati lungo la sequenza con l'indicazione del relativo E-value e della posizione all'interno della sonda.

- Attraverso il sito **SUPERFAMILY** (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY) si possono effettuare ricerche di sequenza in una banca dati di HMM. Dalla pagina iniziale, attivate il collegamento a Sequence search e inserite nel campo apposito la sequenza. Si può scegliere se avere i risultati in tempo reale oppure attraverso notifica per posta elettronica (selezionare la scelta nella casella notification: browser o e-mail). In quest'ultimo caso è necessario inserire il proprio indirizzo di posta elettronica. Nel messaggio elettronico è contenuto un indirizzo attraverso il quale si può accedere direttamente ai risultati che vengono conservati sul sito per 15 giorni. La **Figura 3** riporta una porzione della schermata dei risultati. Si può verificare come siano stati identificati gli stessi domini che erano stati identificati da SMART.

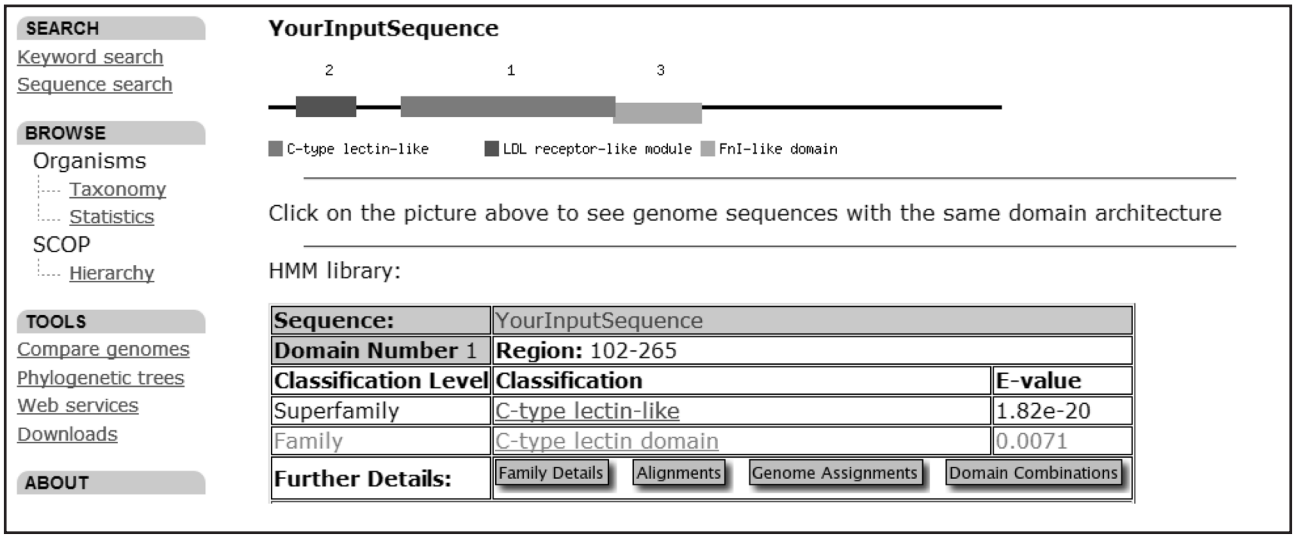


Figura 3 Parte della schermata dei risultati ottenuti con il server SUPERFAMILY. Nella parte superiore è mostrata l'assegnazione dei domini sulla sequenza immessa. Di seguito è riportato l'elenco dei domini con il relativo E-value. I collegamenti attivi consentono di recuperare informazioni strutturali e funzionali sui domini assegnati.

- Un sito che utilizza ricerche con HMM è **HHpred** (http://toolkit.tuebingen.mpg.de/hhpred). Questo sito offre la possibilità di cercare omologie remote tra una sequenza sonda e una libreria di HMM di proteine a struttura nota. L'algoritmo su cui è basato il metodo del sito dovrebbe essere considerato più propriamente un sistema di riconoscimento di fold. Nei risultati è riportata la lista dei domini mappati sulla sequenza (**Figura 4**). Scorrendo lo schermo verso il basso, si arriva alla sezione contenente la lista dettagliata dei domini (**Figura 5**) con le relative informazioni statistiche (per esempio E-value). Ancora più in basso sono riportati gli allineamenti di sequenza con la struttura secondaria prevista e osservata.

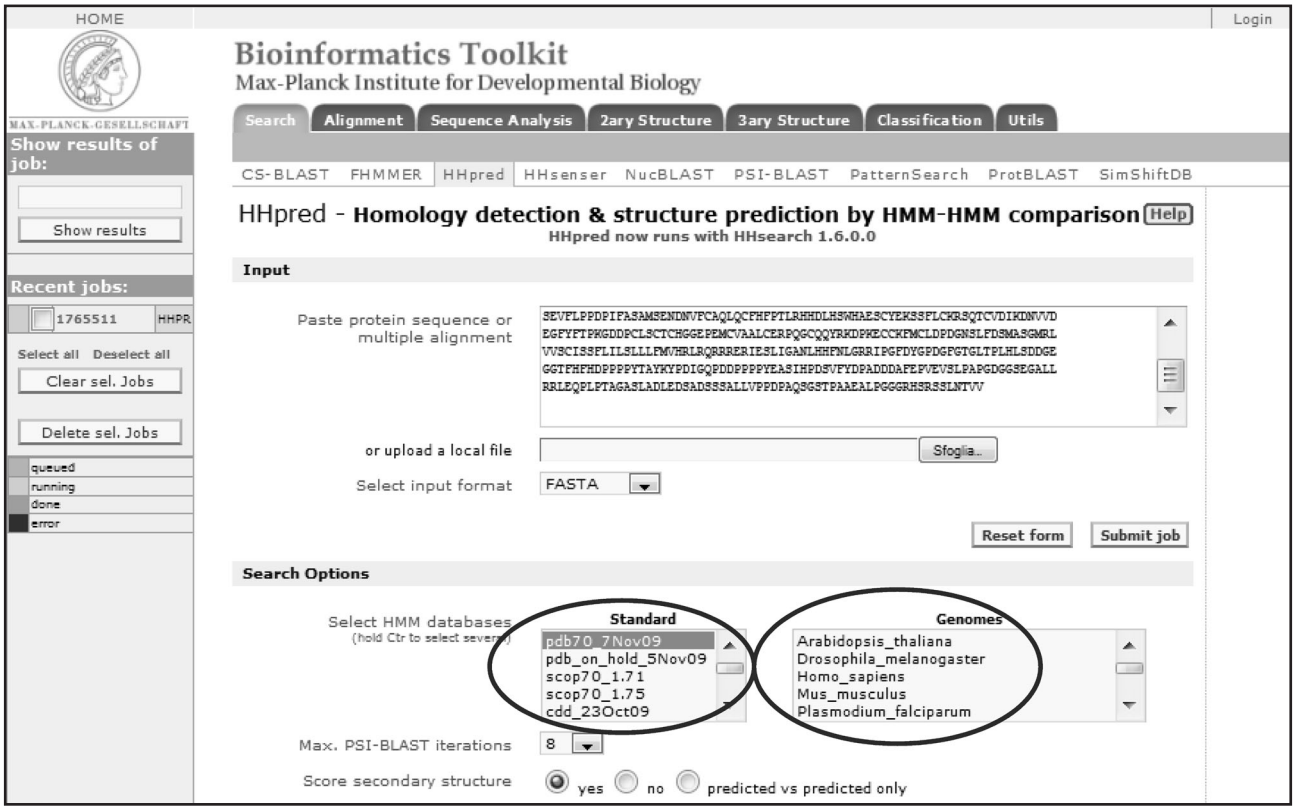


Figura 4 Finestra di ingresso al server HHpred. Nella parte inferiore dello schermo si può selezionare una o più collezioni di HMM derivate da banche dati strutturali (ovale a sinistra) o limitare la ricerca a proteine appartenenti a un particolare genoma (ovale a destra).

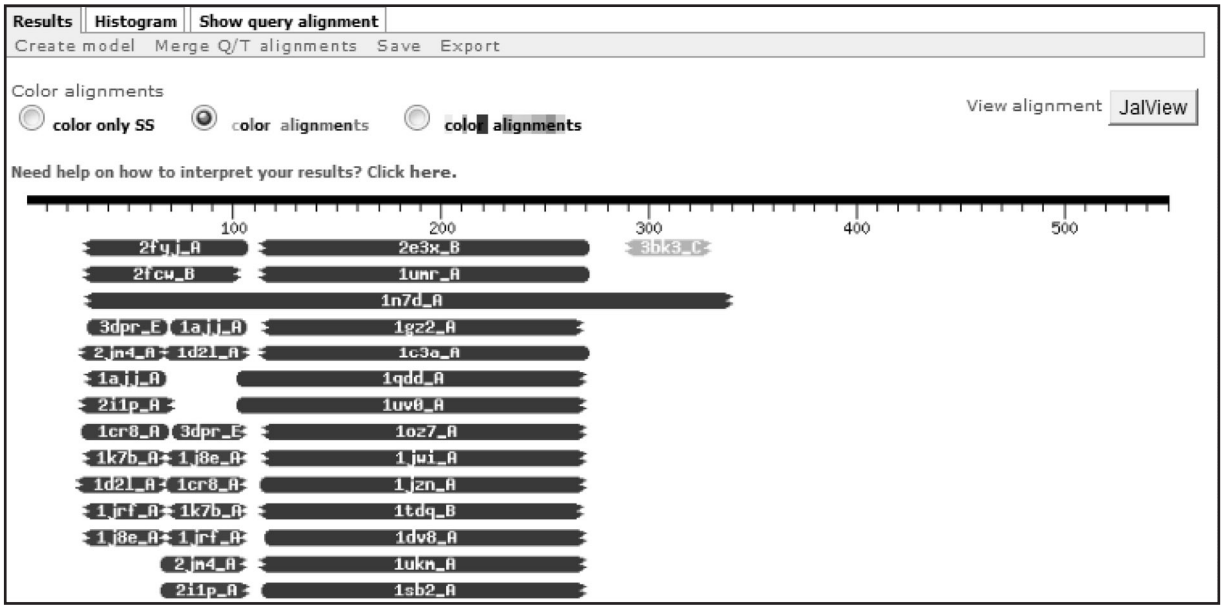


Figura 5 Prima parte della schermata che mostra i risultati della ricerca con HHpred. Sulla sequenza sonda vengono mappati i domini che sono risultati significativamente simili.

Al termine di questa serie di ricerche è possibile avere un quadro sufficientemente dettagliato dell'architettura e della probabile funzione della sequenza. In realtà questo tipo di indagine viene fatta preliminarmente quando la sequenza è inclusa nelle banche dati curate (come Uniprot/Swiss-prot). Le annotazioni della sequenza P98153 recuperata da www.uniprot.org, infatti, contengono l'assegnazione dei domini fin qui trovati e molte altre e dettagliate informazioni.