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.

wellcome sain institute	nger	HOME SEAR	СН В	ROWS	6E I	тр	HEL	P [ABOUT					fgm rd search Go
Sequence se	arch results													
<u>Show</u> the detailed	description of this results page.													
We found 21 Pfam	Ve found 21 Pfam-A matches to your search sequence (1 significant and 29 insignificant) but we did not find any Pfam-B matches.													
<u>Show</u> the search of	ptions and sequence that you subm	nitted.												
<u>Return</u> to the searc	Return to the search form to look for Pfam domains on a new sequence.													
Significant Pf	am-A Matches													
Show or hide all alig	nments.													
Family	Description		Entry type	Clan	Envel Start	lope End	Alignr Start	nent End	HMN From	4 То	Bit score	E-value	Predicted active sites	Show/hide alignment
Ldl recept a Lo	w-density lipoprotein receptor doma	in class A	Repeat	n/a	28	66	29	66	2	37	40.1	1.8e-10	n/a	Hide
#HMM tCkpne #MATCH +C+p++ #PP 6*****	FrCangrCIpkswvCDgedDCadgsDEkdC F C++g+ CIp w+CDg C+d sDE++C													

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**).

Schultz et al. (1998) Proc. Netl. Acc Letunic et al. (2008) Nucleic Acids HOME SETUP FAC ABOUT	d. Sci. L Res. doi: GLOSS	AND 1000	Lew FEEDB	ACK	SMART MODE: NORMAL GENOMIC	Simple Modular Architecture Research Tool
Domains within H Integral membrane pro	267	Sapiens GCR2/IDD pred	465 55	IDD_HUMAN (P98153)	
Position: 115 to 266 E-value:3.75e-15	oma	ain (go to ful	II annotatio	n)		CLECT
Definition:	C-type I	lectin (CTL) or o	arbohydrat	e-recognition domain ((CRD)	
Description: InterPro abstract (IPR001304):	scription: Many of these domains function as calcium-dependent carbohydrate binding modules. erPro abstract R001304): Lectins occur in plants, animals, bacteria and viruses. Initially described for their carbohydrate-binding activity (PUBMED:14533786),(full abstract)					
Name		Begin	End	E-value		
low complexi	tv	10	21	_		
LDLa	ĺ	29	68	1.37e-11		
CLECT		115	266	3.75e-15		
VWC		271	332	1.52e-09		
transmembrane		346	368	-		
low complexi	ty	370	378	-		
low complexi	ity	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.

SEARCH	YourInputSequence	e							
<u>Keyword search</u> <u>Sequence search</u>	2	1 3							
BROWSE Organisms	C-type lectin-like	LDL receptor-like module 🗌 FnI-like domain							
<u>Statistics</u> SCOP	<u>Taxonomy</u> <u>Statistics</u> Click on the picture above to see genome sequences with the same domain architectures of the second sec								
Hibrary:									
TOOLS	Sequence:	YourInputSequence							
Compare genomes	Domain Number 1	Region: 102-265							
Phylogenetic trees	Classification Level	Classification	E-value						
Web services	Superfamily	<u>C-type lectin-like</u>	1.82e-20						
<u>Downloads</u>	Family	<u>C-type lectin domain</u>	0.0071						
ABOUT	Further Details:	Family Details Alignments Genome Assignments	Domain Combinations						

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.

HOME		Login
	Bioinformatics Toolkit Max-Planck Institute for Developmental Biology	
MAX-PLANCK-GESELLSCHAFT	Search Alignment Sequence Analysis 2ary Structure 3ary Structure Classification Utils	
Show results of		
Job.	CS-BLAST FHMMER HHpred HHsenser NucBLAST PSI-BLAST PatternSearch ProtBLAST SimShiftDB	
Show results	HHpred - Homology detection & structure prediction by HMM-HMM comparison Help HHpred now runs with HHsearch 1.6.0.0	
Desert is her	Input	
Recent Jobs: 1765511 HHPR Select all Deselect all Clear sel, Jobs	Paste protein sequence or multiple alignment SEVELPPDPIFASAMSENDNVPCAQLQCPHPPTLRHHDLHSWHAESCYEKSSFLCKRSQTCVDIKUNVVD EFFYFTKRDDPCLSCTCRG62PBMCVALALCERQ6QQYRKDDFKCCKFMCLDPDGSSLFDBAASGMAL VVSCISSFLILSLLFW/NRLRQRRRERIESLGANLHFNLGRRIP6PDGSPGFGGETGLIPLHLSDDGE GGTPHFTDGPPDPTLAXKPDIGQEDDPPPPTLASHPDSVTPDADDDAFEVVEVSLEAFGDG6SEGALL RRLEQPLPTAGASLADLEDSADSSSALLVPPDPAQSGSTPAAEALPGGGRHSRSSLNTVV	
Delete sel. Jobs	or upload a local file Sfoglia	
queued running	Select input format FASTA	
done		
	Reset form Submit job	
	Search Options	
	Select HMM databases Standard Genomes (hold Ctr to select several pdb_on_hold_5Nov09 scop70_1.71 Dr.55 cdd_23Oct09 Plasmodium_falciparum	
	Max. PSI-BLAST iterations 8	
	Score secondary structure $ extbf{0}$ yes $ extbf{0}$ no $ extbf{0}$ predicted vs predicted only	

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).

Results	Histogram	Show query alignment				
Creater	model Men	ge Q/T alignments S	ave Export			
Color ali	gnments r only SS	Color alignments	C color alignments			View alignment JalView
Need help	o on how to in	nterpret your results? Cli	ck here.			
	2fyj_	100 A	200 2e3x_B	300 ≤ 3bk3_C≥	400	500
3	2fcw_l	8 3 4	1unr_A			
:	2	1	ln7d_A			
)	(<u>3dpr_E] 1</u>	ajj_A) E	1gz2_A 3			
4	2_jn4_8 7 1	d21_A	1c3a_A			
	1ajj_A)		Lqdd_A 3			
1	2i1p_A 3		Luv0_A 3			
•	1cr8_A [3	idpr_E	1oz7_A 3			
3	<u>1k7b_A</u> ₹ 1	j8e_A≩ €	1jwi_A 3			
1	1d21_A 1	cr8_A3 (1jzn_A 3			
1	1jrf_A <u>₹</u> 1	.k7b_A3 \$	1tdq_B 3			
1	1.j8e_A 1	jrf_A≩ (1dv8_A 3			
	(2,	jn4_A 🗧	1ukn_A 3			
	(2:	i1p_A 3	1sb2_A 🗧			

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.