Question 10

For a given protein, how can one determine whether it contains any functional domains of interest? What other proteins contain the same functional domains as this protein? How can one determine whether there is a similarity to other proteins, not only at the sequence level, but also at the structural level?

doi:10.1038/ng1198

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics

To demonstrate how to find functional domains within a protein, the human testis-determining factor TDF, also known as the sex-determining protein SRY, will be used as an example.

Although the search could be commenced from the Entrez search box on the NCBI home page, a better way to perform the initial search is from LocusLink[10]. One of the advantages of using LocusLink lies in its standardization of gene and protein names with appropriate cross-referencing, making it more likely that the correct protein will be found on the first attempt. From the NCBI home page at http://www.ncbi.nlm.nih.gov/, choose *LocusLink* from the pull-down menu in the upper left corner, type the gene name, 'TDF', into the query box, and click *Go*. Four loci are returned (Fig. 10.1). The first column gives the Locus ID, which is a stable identifier associated with that gene locus. Clicking on the *LocusID* produces a LocusLink report view; more detailed information on the report view can be found in the LocusLink Help feature and in the literature[15]. The second column, marked *Org*, gives a shorthand version of the organism name. Here, there is one entry from *Drosophila* (*Dm*), one from mouse (*Mm*), one from human (*Hs*) and one from rat (*Rn*). A series of alphabet blocks shown to the right of each entry provide jumping-off points to other database resources. The locus of interest here is the fourth entry in the list, because that is the one for the human form of TDF/SRY. To find additional information on the protein, click on the second *P* (in green) on that line. This takes the user to the protein entries corresponding to that particular LocusLink entry (Fig. 10.2). At this point, the user can click on any of the hyperlinks to look at the raw database information available on any of the proteins listed.

Consider the last entry in the list, an NCBI Reference Protein sequence with accession number NP_003131. To the right of the accession number is a series of hyperlinks. Clicking on the link labeled *BLink* will take the user to the BLink page for the protein of interest (Fig. 10.3). BLink stands for 'BLAST Link' and provides the graphical results of pre-computed BLAST searches that have been performed not just for this protein sequence, but for every protein sequence within the Entrez Proteins data domain. The pre-computed BLAST results for TDF/SRY are shown in the section beginning with the label '204 aa'. Across the top are a number of buttons that allow the user to ask a series of questions regarding their protein of interest. As the object of this question is to find the protein domains present within the TDF/SRY protein, the user can click on *CDD-Search* (Conserved Domain Database Search[18]). Doing this will produce a graphical overview of any domains present within the protein, as well as a sequence alignment of those domains with the query sequence (Fig. 10.4). In this case, one functional domain is found: an HMG box, which is a DNA-binding domain found in many nuclear proteins. The domain was found in all of the databases comprising CDD (Pfam, SMART, and COG), as can be seen by looking at the accession numbers in the hit list.

To determine which other proteins contain this same HMG-box domain, click on the box labeled *Show*, right under the graphical view near the top of the page. This will invoke the domain architecture retrieval tool (DART). DART shows functional domains within a protein and, more importantly, other proteins with a similar domain architecture (Fig. 10.5). The query (the HMG-box) is shown at the top of the page in red. Every other protein in the NCBI's non-redundant sequence database having that same domain is then shown below the query, with the HMG box again colored red. Other domains within the found proteins are also shown, in various colors and shapes, with a key appearing at the bottom of the web page. Clicking on any of the links to the left would provide additional information about these new proteins.

Although a protein domain has now been identified within the query protein, no in-depth information has yet been provided about the function of that domain. Whereas a circuitous path could be followed from the DART page to find this information, an easier method is to use another web-based resource, called InterPro. InterPro is an integrated resource for information about protein families, domains and functional sites, bringing together information from a number of protein domain-based resources, such as PROSITE, PRINTS, Pfam and ProDom[19]. The InterPro Simple Search engine can be accessed from the InterPro home page, at http://www.ebi.ac.uk/interpro. Clicking on *Text Search*, on the left, brings the user to the search page; for this search, type "HMG Box" (with quotes) into the text box and hit *Search*. Two hits are returned (Fig. 10.6). For purposes of this example, follow the link from the first hit, for *high mobility group proteins HMG1 and HMG2* (IPR000135). The resulting InterPro summary page (Fig. 10.7) provides information on the function, intracellular location and, most importantly, metabolic role of this particular protein within the cell, in an executive summary format. References are provided at the bottom of the web page for users who wish for more in-depth information about the domain. Users can also retrieve all of the full-length sequences containing the domain; the reader is referred to the InterPro documentation for more details.

The final part of this question asks whether similarity to the query protein can be found at the structural as well as the sequence level. Answering this question requires a new search against NCBI Structures. From the NCBI home page, change the pull-down menu in the query box at the top of the page to *Structure*, type 'SRY' in the box and hit *Go*. Seven three-dimensional structures are returned, one of which is 1HRY,

At Ensembl, the GeneView links directly to the InterPro domain(s) found in the protein (Fig. 1.10).

the structure of the human SRY–DNA complex solved by NMR. Clicking on the *1HRY* hyperlink takes the user to the Structure Summary page for 1HRY. The summary links to more detailed information about chain A, the protein component of the structure, chain B, the nucleotide component of the structure, and the conserved domain (CD) in the protein, obtained through a CDD search. Click on the chain A graphic to get a list of proteins whose known structures have, using a method called VAST, been deemed similar to that of the original SRY protein; more information on the method and on interpreting the data within the tables can be found elsewhere[15]. Here, the SRY protein is shown to have some structural similarity to a fasciculin 2–mouse acetylcholinesterase complex (1MAH), a protein named V-1 Nef (1AVZ), a heat-shock protein of 70 kD (1QQN), and a myosin motor-domain complex(1BR1) (Fig. 10.8). The VAST program quite often reveals similarities between proteins that are not evident from simple BLAST or FASTA searches, so readers are encouraged to employ this and similar tools when trying to answer questions related to protein families.

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics

Figure 10.1

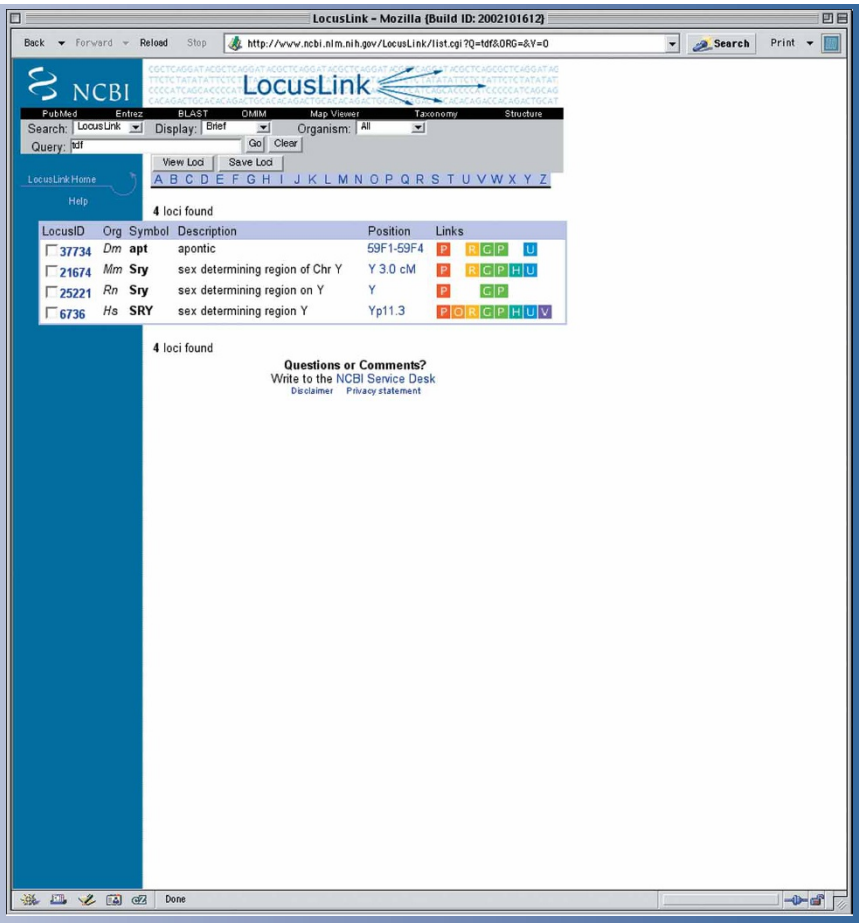


Figure 10.2

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics

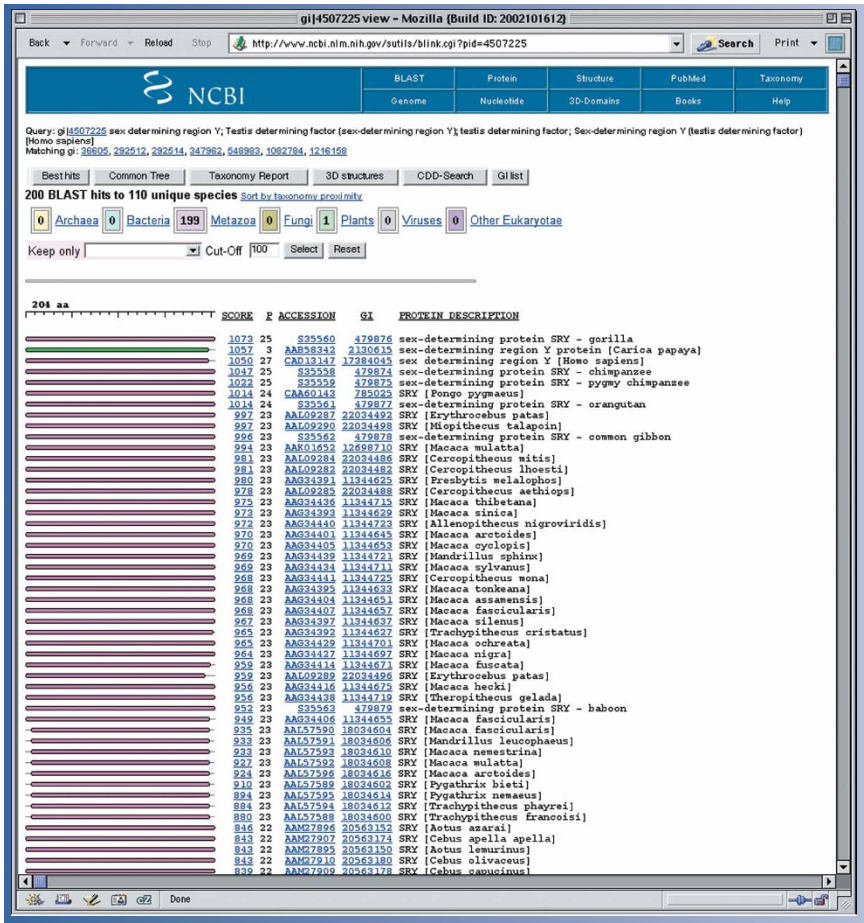


Figure 10.3

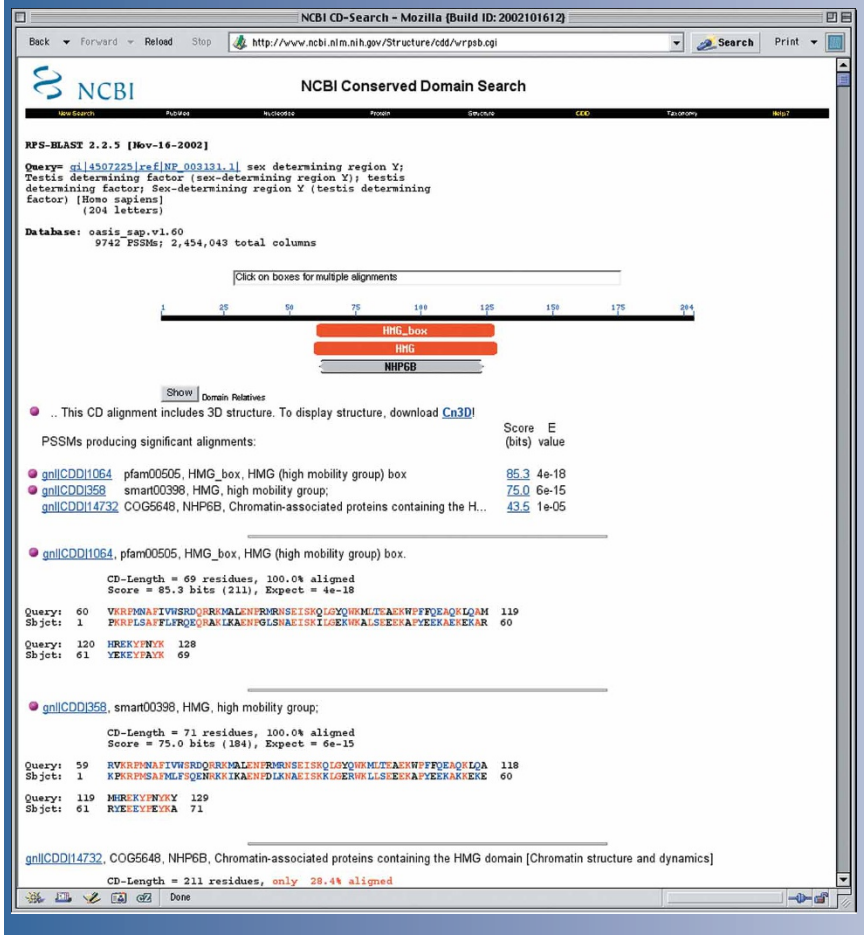


Figure 10.4

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics

Figure 10.5

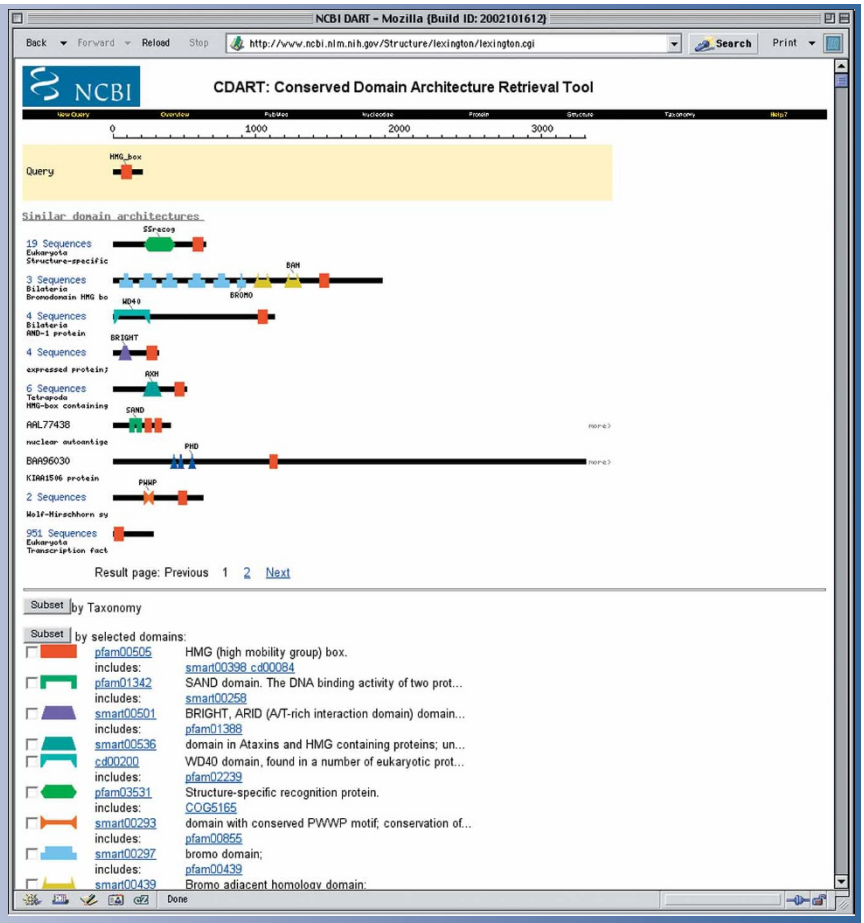
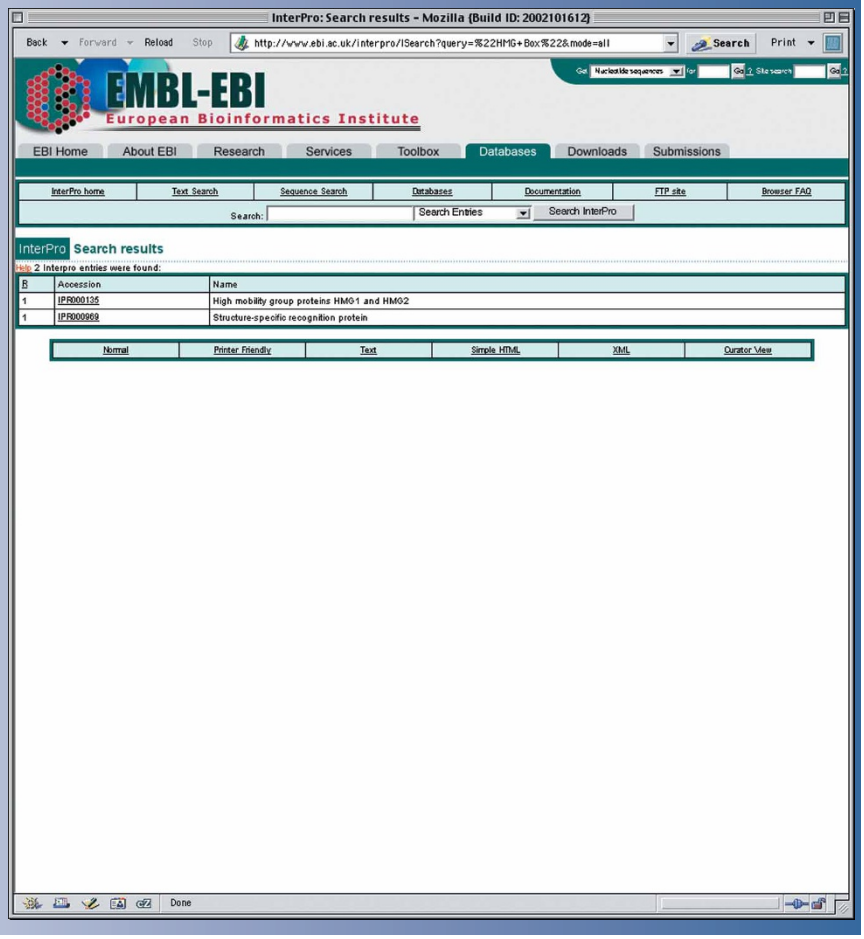


Figure 10.6



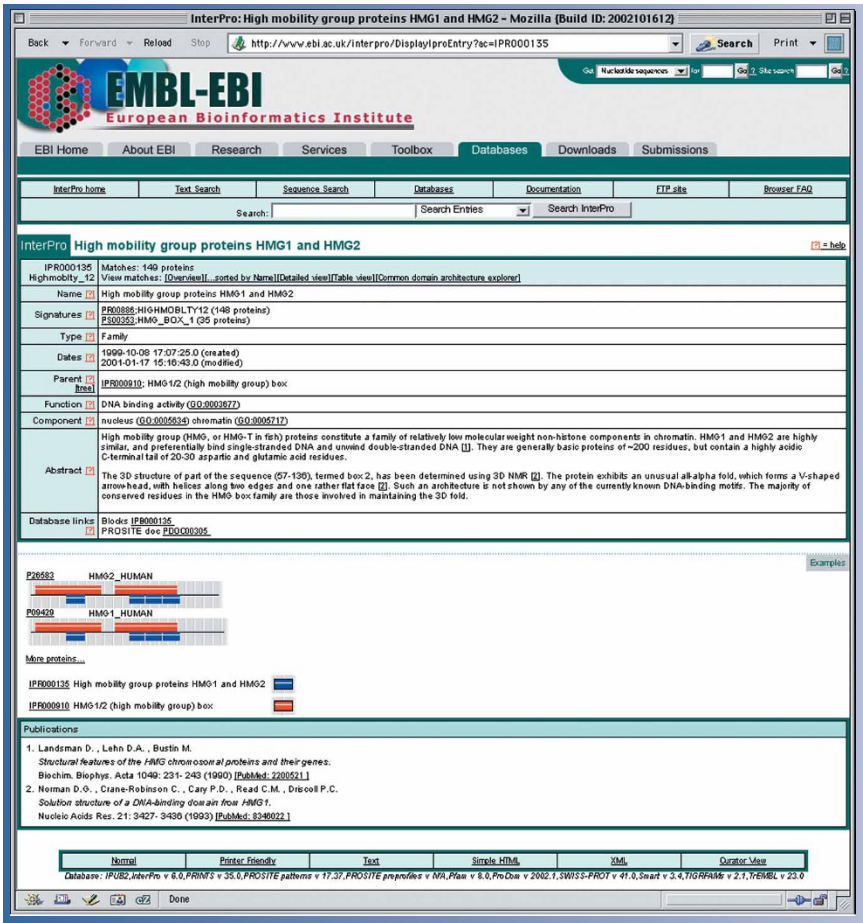


Figure 10.7

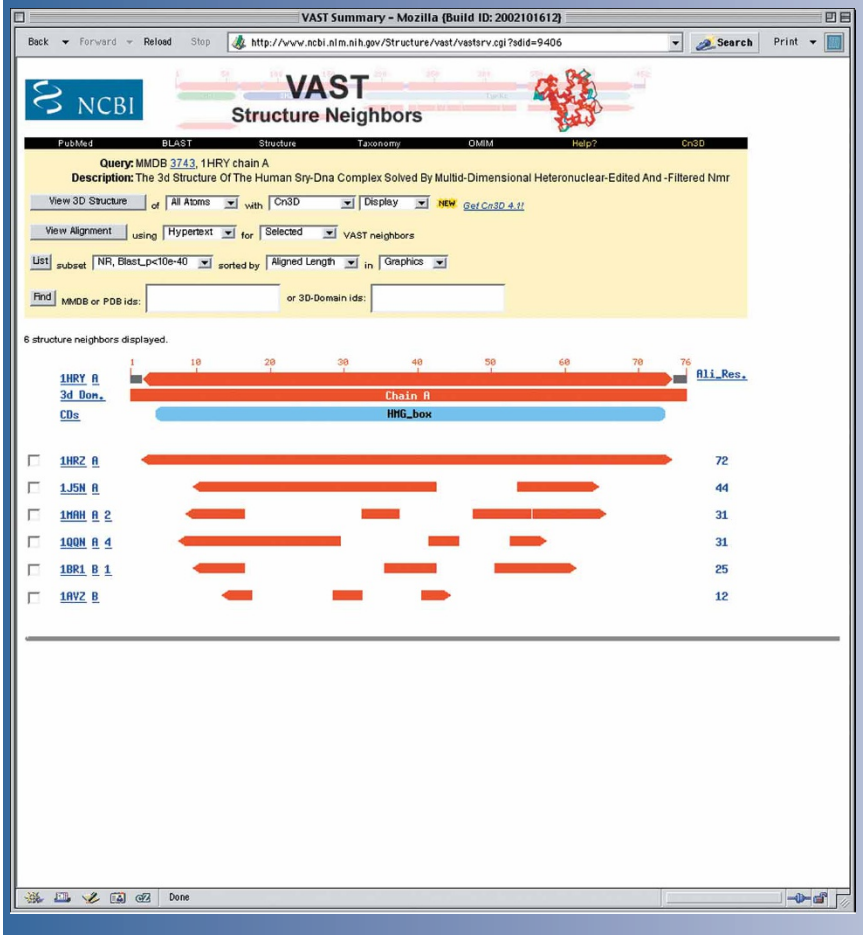


Figure 10.8

© 2003 Nature Publishing Group http://www.nature.com/naturegenetics