The Protein Information Resource

A flood of genomic information is currently pouring from the large sequencing efforts, most prominently from the Human Genome Project. The Protein Information Resource (PIR) is one of the institutions, such as EMBL and others, which channel this flood and thus make it available to researchers. The PIR is built around the Protein Sequence Database (PSD), a rapidly growing stock of currently more than a quarter of a million protein sequences that are freely accessible on the web. These sequences are annotated with additional information of biological or medical relevance, such as short descriptions, references, or hyperlinks to other databases.

The PIR is fostered of by a team of about a dozen specialists at the National Biomedical Research Foundation, which is affiliated to Georgetown University, Washington, DC. This team is to be congratulated because its fosterling is in good shape. A clearly structured and ergonomic entrance web page is the first sign thereof. It gives a compact overview of the most important parts and functions of the PIR and also provides a simple search interface to the PSD. With another mouse click one reaches web interfaces to various databases and to bioinformatics tools. Features are generally well documented and easy to use.

As an example, let us try to find information on the enzyme tyrosine hydroxylase. To this end we type "tyrosine hydroxylase" in the search field at the top left of the entry page and hit the enter key or click the "Go!" button. After a few seconds, we are rewarded with a list of titles of PSD entries that contain our query term. With a further click we can inspect any one of these entries, or modify our query, for example if we think that our query term was too unspecific. For the modification of queries, the PIR provides an elegant function I miss at other database sites (bottom of Figure 1). In our example, we could restrict our search to humans. This is done by choosing the category "Species" beneath the second search field, writing "homo sapiens" into that field, and finally clicking the "Submit" button. Now, our list shrinks from 27 entries to just two, only one of which describes the human tyrosine hydroxylase. Figure 2 shows a section of this entry. Apart from the sequence of amino acids, the entry comprises references to original articles and hyperlinks to abstracts in PubMed, items on genetics and hyperlinks to entries in the OMIM database of hereditary diseases, information on function, modifications and interactions, hyperlinks to three-dimensional structures in the Protein Database (PDB), and more. Furthermore, the sequence of amino acids in the entry can be sent to bioinformatics tools directly, for instance to Psi-Blast or Fasta. Thus, it is easy to perform simple sequence analyses and comparisons, the results of which often give valuable hints to structure and function of a protein. Query results and whole databases can be downloaded in several common formats.

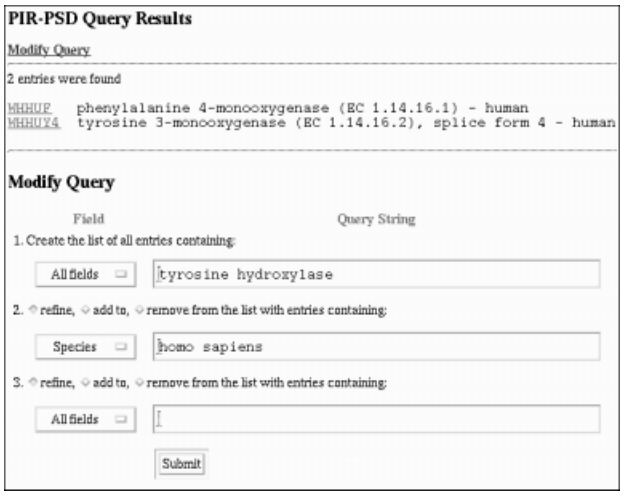

Figure 1. Query result list and refinement form at the Protein Information Resource.

```
FUNCTION
   #description catalyzes the 3'-hydroxylation of tyrosine to 3',
                4'-dihydroxyphenylalanine by tetrahydrobiopterin and
                oxygen
   #pathway     catecholamine biosynthesis
   #note        this is the rate-limiting step in catecholamine
                biosynthesis
CLASSIFICATION  SF000336
   #superfamily phenylalanine 4-monooxygenase
KEYWORDS        alternative splicing; biopterin; catecholamine
                biosynthesis; iron; metalloprotein; monooxygenase;
                oxidoreductase; phosphoprotein

FEATURE
   1-528                #product tyrosine 3-monooxygenase, splice form 4
                        #status predicted #label MAT4\
   1-34,62-528          #product tyrosine 3-monooxygenase, splice form 2
                        #status predicted #label MAT2\
   1-30,35-528          #product tyrosine 3-monooxygenase, splice form 3
                        #status predicted #label MAT3\
   1-30,62-528          #product tyrosine 3-monooxygenase, splice form 1
                        #status predicted #label MAT1\
   8                    #binding_site phosphate (Thr) (covalent) (by
                        unidentified kinase) #status predicted\
```

Figure 2. Tyrosine 3-monooxygenase data.

Suggest a web site or submit a review:
angewandte@wiley-vch.de

The above example demonstrates only a small fraction of the possible usages of the PIR. It offers more databases than the PSD and also a number of bioinformatics tools, even though the range of bioinformatics available at the PIR is somewhat restricted compared to the state of the art, for example for the prediction of structures. Here, an integration (similar to that of Psi-Blast and others) of threading or other similarity-based methods which use known structures for predictions would be helpful.

Daniel Hoffmann
CAESAR, Bonn (Germany)

For further information visit:

http://pir.georgetown.edu/
or contact
pirmail@nbrf.georgetown.edu

  1433-7851/02/4106-1075 $ 17.50+.50/0