

## The Tool *Prosite*



***Prosite*** – a tool for searching motifs of functional or structural importance in a given amino acid sequence, and predicting its activity and function. A motif is a set of amino acids with an essential and conserved sequence or a typical structural element in the protein folding, that is conveyed in a protein sequence and accounts for a biochemical function, structural domain or activity of the protein. In both cases, either when the motif refers to a particular amino acid sequence or a set of contiguous secondary structure elements, it originates in a conserved amino acid sequence. Thus, motifs can be predicted based on the protein sequence only, namely its primary structure. The tool *Prosite* includes a search engine that compares the amino acid sequence of a studied protein (input or query sequence) with sequences in a motif designated database, sequences whose structure or function were formerly experimentally defined and characterized. Based on similarities between regions in the query sequence and motif sequences as deposited in the database, it thereby enables the prediction of motifs that may occur in the studied protein, and defines the protein family it may belong to. This is based on a basic assumption in bioinformatics that similarities between sequences (conserved sequences) usually suggest similarities in structure and function.



***Prosite*** – the tool can be accessed via the URL:

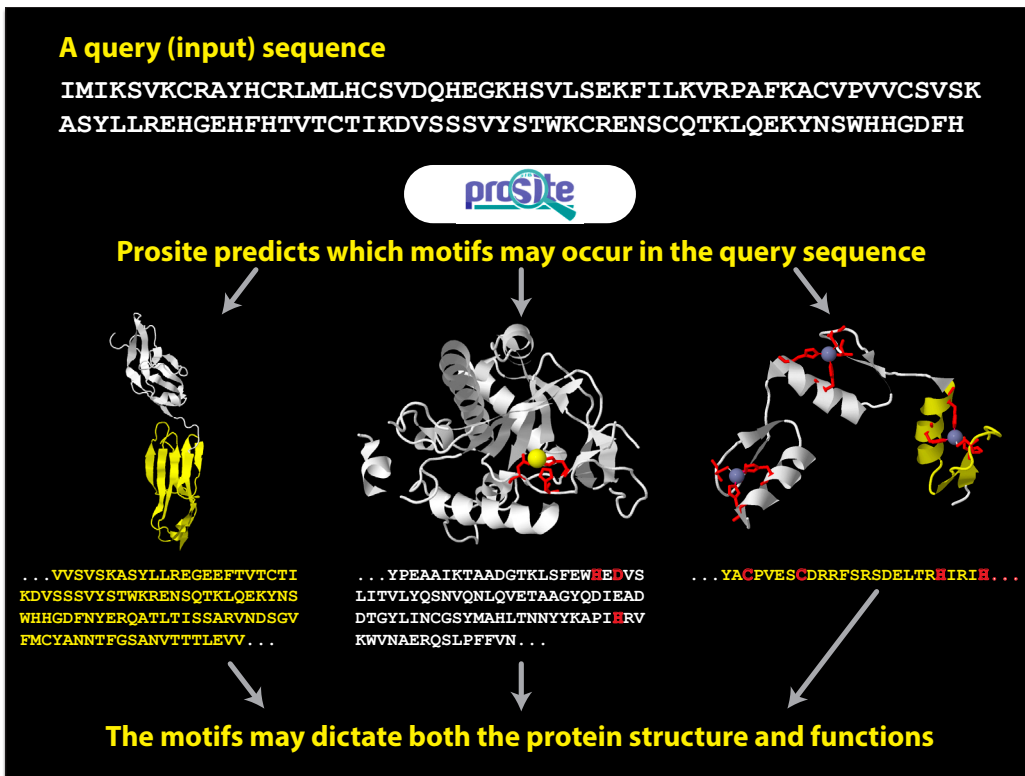
<http://prosite.expasy.org/>



## Welcome to the guided tour of the bioinformatics search tool *Prosite*,

Recently, due to development of novel technologies for sequencing, many sequenced of new proteins are revealed. However, it is not always clear in what processes a new protein is involved in, and which function it possesses. Before investing any resources in exploring such and other issues experimentally in the laboratory, one can learn about its estimated structure, function and activity using the bioinformatics tool *Prosite*.

*Prosite* predicts the motifs that may occur in the studied protein sequence. These motifs may dictate both the protein structure and functions.



A motif is a set of amino acids with an essential and conserved sequence or a typical structural element in the protein folding. In both cases, either when the motif refers to a particular amino acid sequence or a set of contiguous secondary structure elements, motifs can be predicted based on the protein sequence only, without actually determining the protein structure. Determining the three- dimensional structure of a protein is a time consuming, expensive and complicated process.

**A motif is a set of essential and conserved sequence or a typical structural element in the protein folding**

A typical structural element in the protein folding

Essential and conserved amino acids

**The sequence of each motif within the whole protein sequence**

```
MERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFS
RSDHLTTHIRHTTGEKPFACDICGRKFARSDERKRHTKIHLRQKD
```

A motif can consist of a few amino acids that hold a certain role, such as an active site of an enzyme, a metal ion-binding site etc. It can also consist of tens or even hundreds of amino acids folded in a characteristic structure.

**A motif is a set of essential and conserved sequence or a typical structural element in the protein folding**

**A structural motif typical of antibodies**

**A metal ion-binding motif consists of 3 conserved amino acids**

**The motif sequence within the protein sequence**

```
IMIKSVKRAYHRLCLHCSVDQEGKSVLSEKFIKVRPA
FKAVPVVSSKASYLLREGEEFTVCTIKDVSSSVYST
WKRENSQTKLQEKYNSWHHGDFNYERQATLTISSARVN
DSGVFMCYANNTFGANVTTTLEVVDKGFINIFPMINT
TVFVNDGENVDLIVEYEAFPKPEHQQWIYMNRFTDKW
EDYPKSENESNIRYSELHLTRLKGTEGGTY
```

**The motif sequence within the protein sequence**

```
DETKHPGFQDFAEQYYWDVFGLSALLKGYALAL
GKEENFFARHFKPDDTLASVVLIRYPYLDPYPEA
AIKTAADGTKLSFEWEDVSLITVLYQSNVQNLQ
VETAAGYQDIEADDTGYLINCGSYMAHLTNNYYK
APIERVKWVNAERQSLPFFVNLGDYSDVIDPFDPR
EPNGKSDREPLSYGDYLQNGLVSL
```

## The basic principle behind Prosite

*Prosite* is a searching engine that assists in analyzing an amino acid sequence of a protein with an unknown structure or function. The tool compares the query sequence to sequences of motifs which structures and functions were already investigated in previous researches and submitted to an appointed database. If the query sequence harbors regions that are similar to the amino acid sequence of known motifs, we can assume with high certainty that the studied protein conveys these motif. This relies on a basic assumption in bioinformatics that claims that similarities among protein sequences and sequence conservation correlates, in most cases, to functional and structural similarities between the proteins.

## Prosite – the interface

We have an amino acid sequence of a protein which activity and function we wish to study using *Prosite*. This is the interface of the *Prosite* tool. The interface allows the user to perform different searching actions that relate to protein motifs, and we will focus on the part that allows screening for motifs in a query sequence. We will paste in the amino acid sequence, in a FASTA format, into the designated window, and click “Scan”. The tool will now scan the protein sequence, testing, using the motifs database, which motifs are similar to segments in the query protein sequence.

The screenshot displays the Prosite web interface. At the top, there are logos for SIB (Swiss Institute of Bioinformatics) and EXPASY, along with the text 'EXPASY Proteomics Server'. A search bar contains the word 'PROSITE' and has 'Go' and 'Clear' buttons. Below the logos, a breadcrumb trail reads 'You are here: EXPASY CH > Databases > PROSITE'. The main heading is 'prosite Database of protein domains, families and functional sites'. A navigation menu includes 'Home', 'ScanProsite', 'ProRule', 'Documents', 'Downloads', 'Links', and 'Funding'. A paragraph describes Prosite's content: 'PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More details / References / Disclaimer / Commercial users]. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More details].' Below this, it states 'Release 20.68, of 11-Jan-2011 (1599 documentation entries, 1308 patterns, 912 profiles and 898 ProRule)'. The 'PROSITE access' section features a search input field with the example 'e.g. PDOC00022, PS50089, SH3, zinc finger', a 'Search' button, and an 'add wildcard "\*" ' checkbox. To the right, a 'Browse:' section lists options: 'by documentation entry', 'by ProRule description', 'by taxonomic scope', and 'by number of positive hit'. The 'PROSITE tools' section is divided into two columns. The left column is titled 'Scan a sequence against PROSITE patterns and profiles - quick scan' and includes a diagram showing four 'PAN' motifs and one 'TRYPSIN\_DOM' motif. Below the diagram, it says '(Output includes graphical view and feature detection)'. The text 'Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [ help ]:' is followed by a text area containing a protein sequence in FASTA format: '>Protein\nMRLPGAMPALALKGEI...GP\nELVNLVSSSTFVLTCS...ET\nFSSVLTNLTLGLDT...TV\nPDPTVGFPLPNDAEEL...TL\nHEKKGDVALFPVPYDHQRGFSQTFEDRSTYCKRTITQDREVD'. A yellow callout box points to the text area with the text 'Enter the studied amino acid sequence here (FASTA Format)'. Below the text area are 'Scan' and 'Clear' buttons. A red arrow points to the 'Scan' button, and a yellow callout box says 'Click to begin scanning'. The right column of the 'PROSITE tools' section lists: 'ScanProsite - advanced scan', 'PRATT - allows to interactively generate conserved patterns from a series of unaligned proteins.', and 'MyDomains - Image Creator new - allows to generate custom domain figures.' Below this is a diagram showing a flow from 'Custom' to 'Images' to 'of' to 'DOMAINS'.

## Reading the results page

On the first section of the results page, we are presented with general data the tool found for the query sequence: the sequence, its name and length and the motifs that were found in the sequence.

**SIB** Swiss Institute of Bioinformatics **ExPASy** Proteomics Server

Search  for

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Databases > PROSITE

Home [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#)

### ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features ([help](#)).

**Hits for all PROSITE (release 20.68) motifs on sequence Protein :**

found: 6 hits in 1 sequence

PROTEIN (1106 aa)

MRLPGAMPALALKGELLLSLLLLLLEPQISQGLVVTTPGPPELVLVNSSTFVLTCSSGAPVWERM  
QEPPEMAKAQDGTSSVLTNLNLTGLDTGEYFCTHNSRGLTDERKRLYIFVPDPTVGLPND  
EELFIFLITEITEITIPCRVTDPLVVTLEHKKGDVALPVPYDHQRFSGIFEDRSYICKTTIGD  
VSDAYVYRQLQVSSINVSNAVQTVVRQGENITLHCIVIGNEVVNFETYPKESGRLVEPVTD  
LLDMPYHIRSILHIPSAAELEDSTYTCNVTVESVNDHQDEKAINITVVESGYVRLLEGEVGT  
LQFAELHRSRTLQVFEAYPPPTVLFKDNRTLGDSSAGEIALSTRNVSETRYVSELTLVVRV  
KVAEAGHYTMRAFHEDAQVLSFQLQINVPVVRVLELSESHPDSEGTVRCRGRGMPQPNII  
WSACRDLKRCPRELPTLLGNSSEESQLETNVTVYEEEEQEFVUVSTLRLQHVDRPLSVR  
CTLRNAVQDQTQEVIVVPHSLPFKVVVISAILALVVLTIISLIILIMLWQKKPRYEIR  
WKVIESVSSDGHEYIYVDPMLQPYDSTWELPRDQLVLRGTLGSGAFGQVVEATAHGL  
SHSQATMKVAVKMLKSTARSSEKQALMSELKIMSHLGPLHLNVNLLGACTKGGPIYI  
ITEYCRYGDLVDYLHRNKHTFLQHHSKRRRPPSAELYSNALPVGLPLPSHVSLTGES  
DGGYMDMSKDESVDVYVPLDMKGDVKYADIESSNYMAPYDNYVPSAPERTCRATLI  
NESPVLSYMDLVGFSYQVANGMEFLASKNCVHRDLAARNVLICEGKLVKICDFGLARD  
IMRDSNYISRGSTFLPLKWMAPESIFNSLYTTLSDVWSFGILLWEIFTLGGTPYPEL  
PMNEQFYNAIKRGYRMAQPAHASDEIYEIMQKCEWEEKFEIRPPFSQLVLLERLLGEGY  
KKKYQQVDEEFLRSDHPAILRSQARLPGFHGLRSPLDTSSVLYTAVQPNEGDNDYI  
IPLPDPKPEVADEGLEGSPSLASSTLNEVNTSS TISCDSPLEPQDEPEPEPQLE  
LQVEPEPELEQLPDSGCPAPRAEAEDSFL

Number of identified motifs ("hits")

Sequence name and length

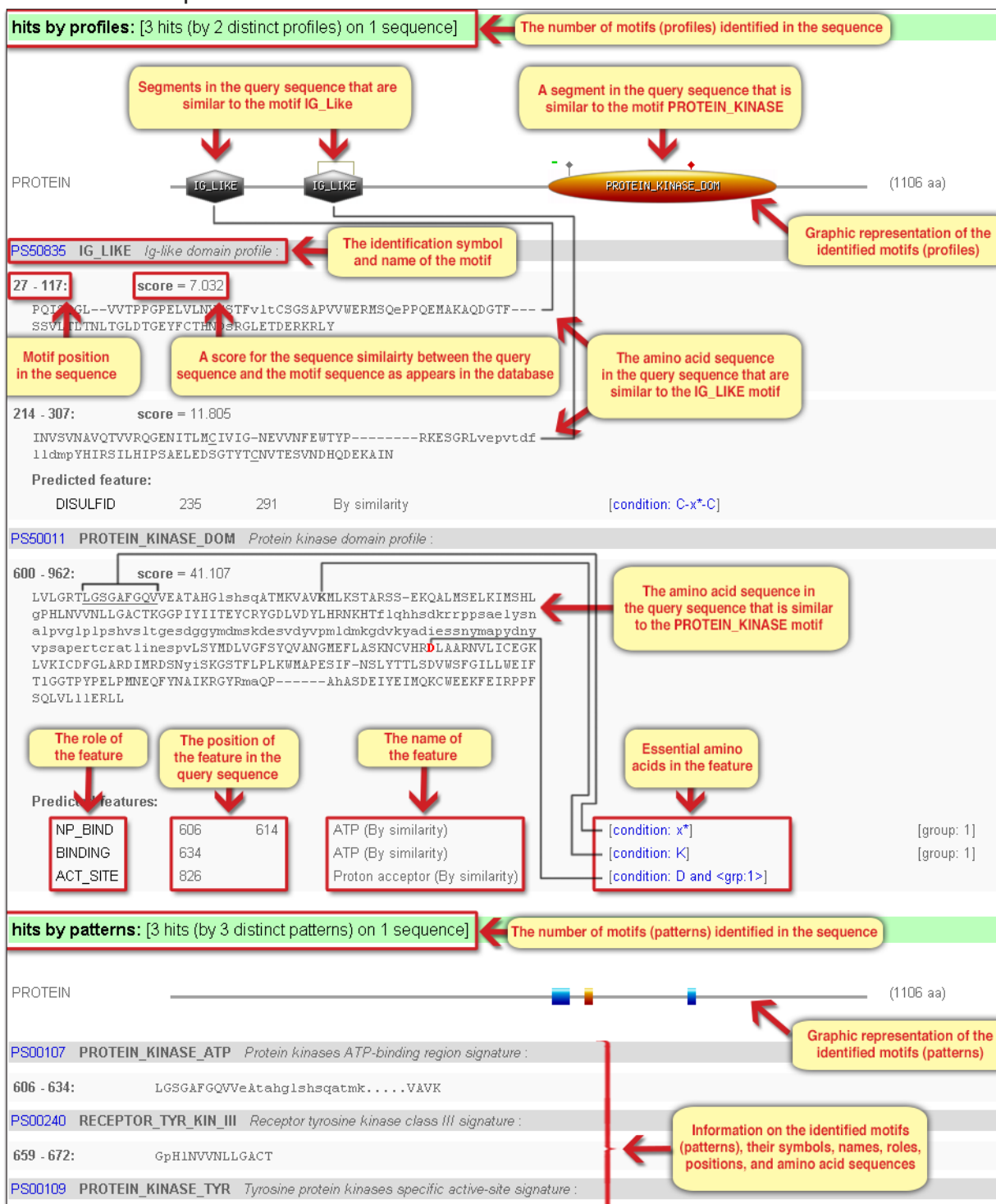
The scanned amino acid sequence

Scale for query sequence length

ruler: 1 100 200 300 400 500 600 700 800 900 1000

Next in the results page, the motifs themselves are listed. The long motifs (also called “profiles”) are presented first, and then the short motifs (also called “patterns”). We will focus only on the long motifs.


In the graphic description that presents the motifs identified within the protein sequence, three segments of the query sequence were identified as similar to motifs of different types. The first two segments are similar to the motif “IG\_LIKE”, and the last one to the motif “PROTEIN\_KINASE”. Next on the results page, the sequences of these segments are described. In addition, for every segment, the tool reports the positions in the sequence where the motif starts and ends. Next to each segment appears a score that indicates the level of similarity between the sequences as appeared in the segment of the query protein and the motif as defined in the database. A high score means that the segment in the protein is more similar to the motif sequence. Sometimes, amino acids within the motif that convey an essential activity are marked in the segment sequence. Their exact location and the function they play role in are reported under “Predicted Feature”, presented immediately after the motif sequence.





What are these motifs and what are they involved in? We can click on one of the symbols of the motif in the graphic description, or on the name of the motif, and a datasheet will open. From the first lines of the datasheet we learn that the motif IG\_LIKE is common in many proteins and many tissues, and that proteins that have this motif (such as antibodies) are involved in binding.

## Information page for the Ig-like motif

 [Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

**PDOC50835**

### Ig-like domain profile

**Description:**


The Ig-like domain is probably the most widespread domain, at least in animals. This domain can be considered as an heterogeneous group built on a common fold. Proteins containing an Ig-like domain differ in their tissue distribution, amino acid composition, and biological role [1,2,3].

All Ig-like domains appear to be involved in binding functions. The ligands range from small molecules (antigens, chromophores), to hormones (growth hormone, interferons, prolactin), up to giant molecules (muscle proteins) [3]. Binding sites are localized either in the loop regions (the most variable parts of the immunoglobulins) or in strands. For instance, distinct areas of the sheets are used to bind the ligands of the MHC, CD8, CD4, and PapD molecules or of the growth hormone receptor (GHR). These binding sites may be formed by a single chain (CD2, CD4), by homodimers (GHR, CD8), or by heterodimers [3].

Classical Ig-like domains are composed of 7 to 10  $\beta$  strands, distributed between two sheets with typical topology and connectivity described as a Greek key  $\beta$ -barrel (see <PDB:3HLA>). The general shape of Ig-like domains is well conserved, but they can differ significantly in their size, owing to high variability of the loops. While a classical domain contains about 100 residues, smaller ones (74-90 residues) have been observed in several Ig-related molecules (CD2, CD4). Large decorations within loops, sometimes including extra domains, are found in hemocyanin (238 amino acids) [4] and transcription factor NFkappaB (201 amino acids) [5]. The schematic representation of the structure of a typical Ig-like domain is shown below.

The other motif that was identified in the query sequence is Protein Kinase. We can thus conclude that the sequence we study is of a protein that belongs to the kinases protein family. The latter is a large and diverse family of proteins that share an enzymatic activity of transferring a phosphate group to other proteins. The kinase proteins carry an important role in activation and inhibition of cellular processes, and we can read more about it in the datasheet.

## Information page for the protein kinases motif

 [Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

**PDOC00100**

### Protein kinases signatures and profile

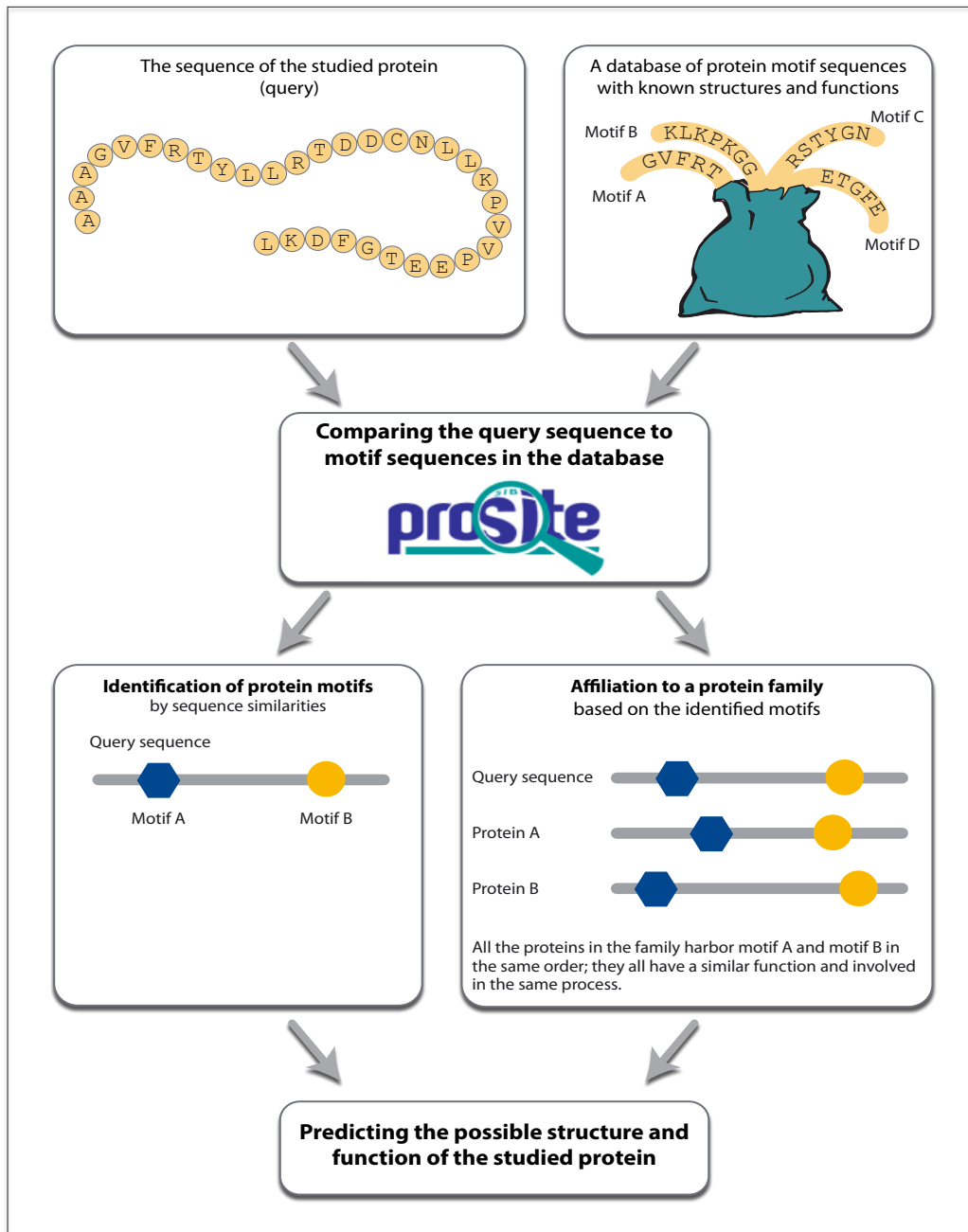
**Description:**

Eukaryotic protein kinases [1,2,3,4,5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. We have selected two of these regions to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme [6]; we have derived two signature patterns for that region: one specific for serine/ threonine kinases and the other for tyrosine kinases. We also developed a profile which is based on the alignment in [1] and covers the entire catalytic domain.

Note:  
If a protein analyzed includes the two protein kinase signatures, the probability of it being a protein kinase is close to 100%.

## Summary

The tool *Prosite* relies on the basic assumption that similarities between protein sequences reflect, in most cases, similarities in protein structure and function. The tool analyzes protein sequences the user submits in order to predict its function. It compares the query amino acid sequence to amino acid sequences of motifs which structures and functions were characterized in previous studies. The tool predicts the motifs the query sequence conveys, and the protein families it might belong to. Based on this information, the structure and functions of the protein can be estimated with high certainty.



We invite you to experiment with the tool and hope you have fun. Good luck!