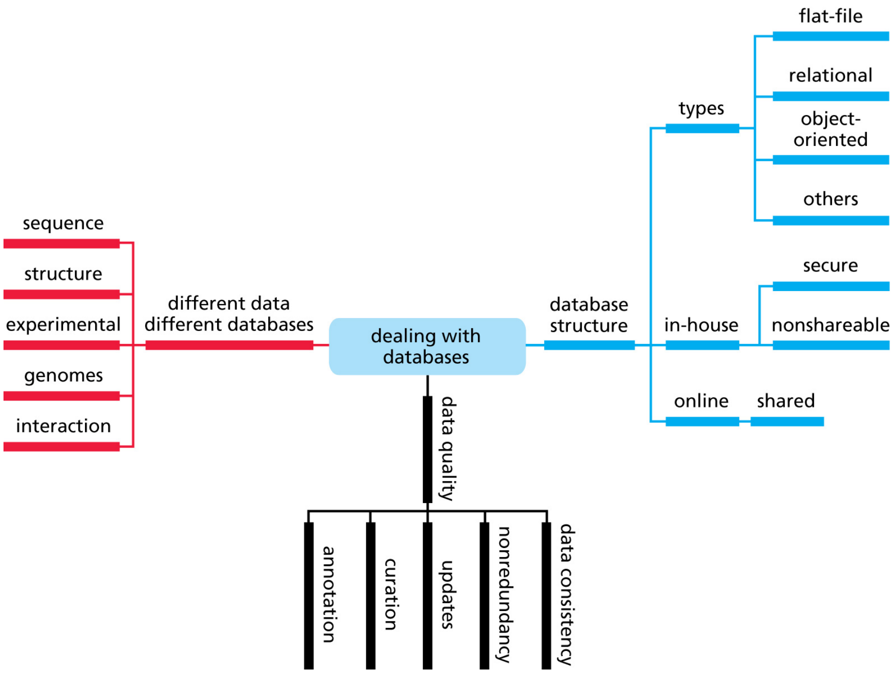
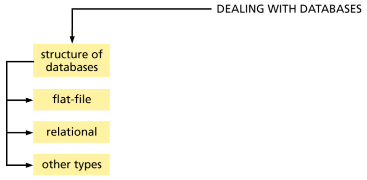
General aspects of databases



Structure of databases



Flat database: it is the simplest form of a database where collections of data (aminoacid sequence) are stored as a large txt file or more than one txt file.

Relational database: it stores the data within a number of tables, each consisting of records and fields. Each table will be linked to at least one other by a shared field called a KEY.

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

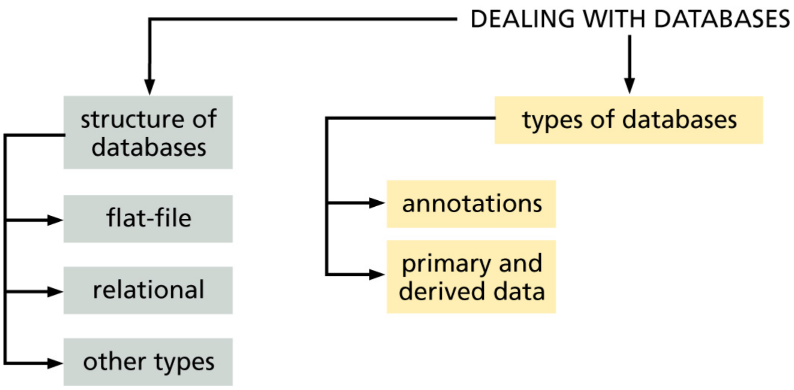
Type of databases

Data: it is the minimal content of a database including data's identity (for example protein name and source) and the author/submitter responsible for the entry.

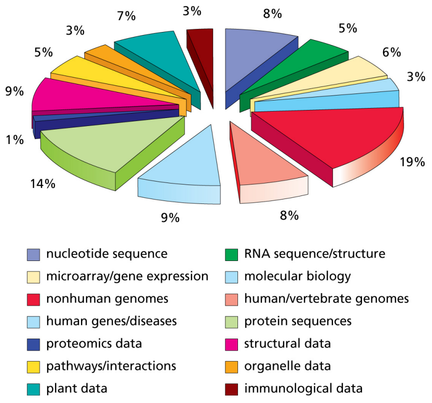
Annotation: provide more information to the data (published papers, lists of entries in other databases, gene structure)

Primary data: they include the raw experimental results.

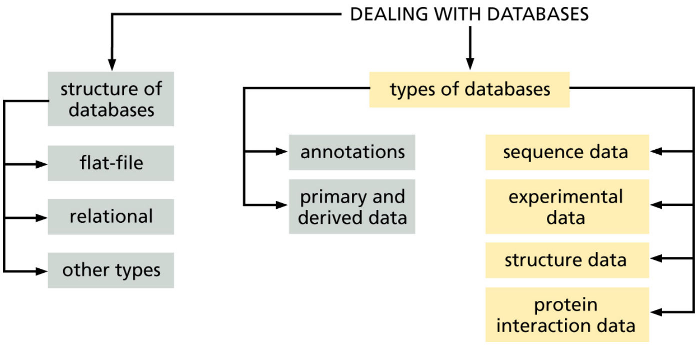
Derived data: based on the data existing at the time (example: conserved protein sequence motifs).



Looking for databases



Distribution of the type of databases as classified at the Nucleic Acid Research (NRA) Molecular Biology Database Collection Web site. In 2006 there were 858 databases listed, classified into 14 main catagories.



Sequence database

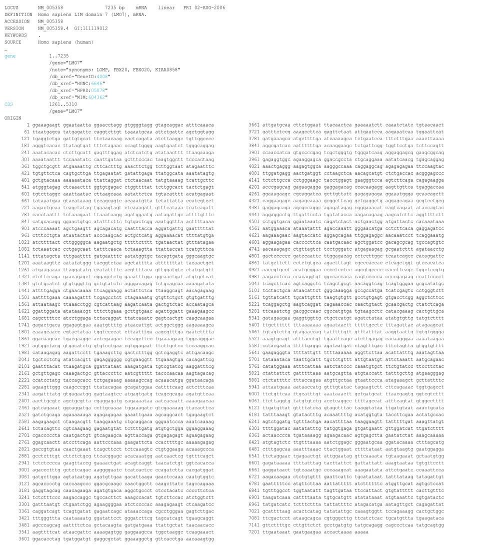
1. DNA sequences:

- Raw genomic sequence (chromosomal DNA)

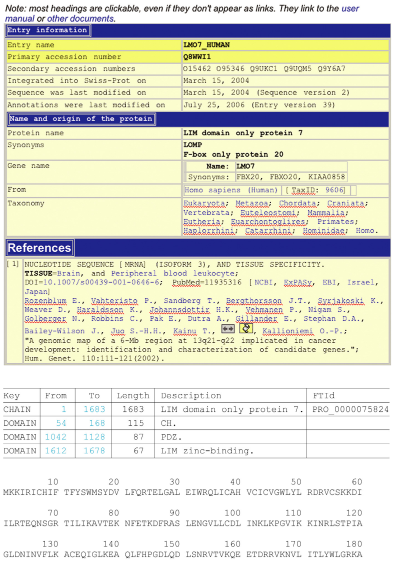
- cDNA (from mRNA)

- Expressed sequence tags (ESTs). Partial cDNA seqeunce.

2. Protein sequences (UniProtKB, Swiss-Prot, NCBI Protein Database



```
LOCUS       NM_005358               7235 bp    mRNA    linear   PRI 02-AUG-2006
DEFINITION  Homo sapiens LIM domain 7 (LMO7), mRNA.
ACCESSION   NM_005358
VERSION     NM_005358.4  GI:111119012
KEYWORDS    .
SOURCE      Homo sapiens (human)
…
gene            1..7235
                /gene="LMO7"
                /note="synonyms: LOMP, FBX20, FBXO20, KIAA0858"
                /db_xref="GeneID:4008"
                /db_xref="HGNC:6646"
                /db_xref="HPRD:05078"
                /db_xref="MIM:604362"
CDS             1261..5310
                /gene="LMO7"
ORIGIN
        1 ggaaagaagt ggaataatta ggaacctagg gtggggtagg gtagcaggac atttcaaaca
       61 ttaatgagca tatgagattc caggtcttgt taaaatgcaa attctgattc agctggtagg
      121 tgaggtctga gattgtgcat ttctaacaag cactcagata atcttaaggc tgttggcccc
      181 agggtcacac ttatagtgat tttctagaac ccagttgggg aagtgaatct tgggcaggag
      241 aaatacacac ctcttgcatt gagtttggag atctcatctg atataacttt ttaagaaaga
      301 aaaataattt tccaaatatc caattgataa gctttcccac taagtggctt tcccactaag
      361 tggctgcgtt atgaaaattg cttcactttg aaacttctgg tcttggtaat atagaatttc
      421 tgtgttctca cagtgcttga ttgagaatat gatattgaga ttatggcata aaatatagtg
      481 gctgtacaaa aaaaaataca ttattaggat ctctaacaat tatgtaaaag tcattgcttc
      541 atgggtagag ctcaaacttt ggtgtgagac ctggttttat tcttggcact tactctgagt
      601 tgtcttaggc aaattaatac cttaagcaaa aatattctca tgtacatttt acatgagaat
      661 tataaatgaa gtacataaag tccagcagtc acaaatgtta tctattatta ccatcgtcct
      721 aagactgcaa tcagctatag tgaaagtagt ctcaaagatt gtttcataaa tcatcagatt
      781 cacctaattt tctaaagaat ttaaataagg agatggaatg aatagattgc attttgtttc
      841 catgcacagg ggaactgtgc atatttcttc tgtgactcgg aaatggttta acttttaaaa
      901 atcccaaaat agctgaagtt agcagacatg caatttacca aggatgattg gaatttttat
      961 ctttcctgta ataatactat acccaagcac actgctcatg aggaaaacat ttttatgtga
     1021 atcttttact cttgggggca aagaatgctg tttttctttt tgataactat gtttatagaa
     1081 tctaaatcac cctgagcaat tatttcaaca tctaaagtta ttattaccat tcatgtttca
     1141 tttatagcta tttgaatttt gatgaatttc aatatggtgc tacagtgata gggcaagtgc
     1201 aaataagttc aatatatggg tacggtctaa agctatttta atttttttat tacaactgct
     1261 atgaagaaaa ttaggatatg ccatattttc acgttttaca gttggatgtc ctatgatgtt
     1321 ctcttccaga gaacagagct cggagctctg gaaatttgga ggcaactgat atgtgctcat
     1381 gtctgcatct gtgtgggttg gctgtatctc agggacagag tctgcagcaa aaaagatata
     1441 attttgagga ctgaacaaaa ttcaggaagg actattctca ttaaggcagt aacagagaag
     1501 aattttgaaa caaaagattt tcgagcctct ctagaaaatg gtgttctgct gtgtgatttg
     1561 attaataagc ttaaacctgg cgtcattaag aagatcaata gactgtctac accaatagca
     1621 ggattggata atataaacgt ttcttgaaa gcttgtgaac agattggatt gaaagaagcc
     1681 cagcttttcc atcctggaga tctacaggat ttatcaaatc gagtcactgt caagcaagaa
     1741 gagactgaca ggagagtgaa aaatgttttg ataacattgt actggctggg aagaaaagca
     1801 caaagcaacc cgtactataa tggtccccat cttaatttga aagcgtttga gaatctttta
     1861 ggacaagcac tgacgaaggc actcgaagac tccagcttcc tgaaaagaag tggcagggac
     1921 agtggctacg gtgacatctg gtgtcctgaa cgtggagaat ttcttgctcc tccaaggcac
     1981 cataagagag aagattcctt tgaaagcttg gactctttgg gctcgaggtc attgacaagc
     2041 tgctcctctg atatcacgtt gagagggggg cgtgaaggtt ttgaaagtga cacagattcg
     2101 gaatttacat ttaagatgca ggattataat aaagatgata tgtcgtatcg aaggatttcg
     2161 gctgttgagc caaagactgc gttacccttc aatcgttttt tacccaacaa aagtagacag
     2221 ccatcctatg taccagcacc tctgagaaag aaaaagccag acaaacatga ggataacaga
     2281 agaagttggg caagcccggt ttatacagaa gcagatggaa cattttcaag actctttcaa
     2341 aagatttatg gtgagaatgg gagtaagtcc atgagtgatg tcagcgcaga agatgttcaa
     2401 aacttgcgtc agctgcgtta cgaggagatg cagaaaataa aatcacaatt aaaagaacaa
     2461 gatcagaaat ggcaggatga ccttgcaaaa tggaaagatc gtcgaaaaag ttacacttca
     2521 gatctgcaga agaaaaaaga agagagagaa gaaattgaaa agcaggcact tgagaagtct
     2581 aagagaagct ctaagacgtt taaggaaatg ctgcaggaca gggaatccca aaatcaaaag
     2641 tctacagttc cgtcaagaag gagaatgtat tcttttgatg atgtgctgga ggaaggaaag
     2701 cgacccccta caatgactgt gtcagaagca agttaccaga gtgagagagt agaagagaag
     2761 ggagcaactt atccttcaga aattcccaaa gaagattcta ccacttttgc aaaaagagag
     2821 gaccgtgtaa caactgaaat tcagcttcct tctcaaagtc ctgtggaaga acaaagccca
     2881 gcctctttgt cttctctgcg ttcacggagc acacaaatgg aatcaactcg tgtttcagct
     2941 tctctcccca gaagttaccg gaaaactgat acagtcaggt taacatctgt ggtcacacca
     3001 agaccctttg gctctcagac aaggggaatc tcatcactcc ccagatctta cacgatggat
     3061 gatgcttgga agtataatgg agatgttgaa gacattaaga gaactccaaa caatgtggtc
     3121 agcacccctg caccaagccc ggacgcaagc caactggctt caagcttatc tagccagaaa
     3181 gaggtagcag caacagaaga agatgtgaca aggctgccct ctcctacatc cccccttctca
     3241 tctctttccc aagaccaggc tgccacttct aaagccacat tgtcttccac atctggtctt
     3301 gatttaatgt ctgaatctgg agaaggggaa atctccccac aaagagaagt ctcaagatcc
     3361 caggatcagt tcagtgatat gagaatcagc ataaaccaga cgcctgggaa gagtcttgac
     3421 tttgggttta caataaaatg ggatattcct gggatcttcg tagcatcagt tgaagcaggt
     3481 agcccagcag aattttctca gctacaagta gatgatgaaa ttattgctat taacaacacc
     3541 aagttttcat ataacgattc aaaagagtgg gaggaagcca tggctaaggc tcaagaaact
     3601 ggacacctag tgatggatgt gaggcgctat ggaaaggctg gttcacctga aacaaagtgg

     3661 attgatgcaa cttctggaat ttacaactca gaaaaatctt caaatctatc tgtaacaact
     3721 gatttctccg aaagccttca gagttctaat attgaatcca aagaaatcaa tggaattcat
     3781 gatgaaagca atgcttttga atcaaaagca tctgaatcca tttctttgaa aaacttaaaa
     3841 aggcgatcac aatttttttga acaaggaagc tctgattcgg tggttcctga tcttccagtt
     3901 ccaaccatca gtgccccgag tcgctgggtg tgggatcaag aggaggagcg gaagcggcag
     3961 gagaggtggc agaaggagca ggaccgccta ctgcaggaaa aatatcaacg tgagcaggag
     4021 aaactgaggg aagagtggca aagggccaaa caggaggcag agagagagaa ttccaagtac
     4081 ttggatgagg aactgatggt cctaagctca aacagcatgt ctctgaccac acgggagccc
     4141 tctcttgcca cctgggaagc tacctggagt gaagggtcca agtcttcaga cagagaagga
     4201 acccgagcag gagaagagga gaggagacag ccacaagagg aagttgttca tgaggaccaa
     4261 ggaaagaagc cgcaggatca gcttgttatt gagagagaga ggaaatggga gcaacagctt
     4321 caggaagagc aagagcaaaa gcggcttcag gctgaggctg aggagcagaa gcgtcctgcg
     4381 gaggagcaga agcgccaggc agagatagag cgggaaacat cagtcagaat ataccagtac
     4441 aggaggcctg ttgattccta tgatatacca aagacagaag aagcatcttc aggttttctt
     4501 cctggtgaca ggaataaatc cagatctact actgaactgg atgattactc cacaaataaa
     4561 aatggaaaca ataaatattt agaccaaatt gggaacatga cctcttcaca gaggagatcc
     4621 aagaaagaac aagtaccatc aggagcagaa ttggagaggc aacaaatcct tcaggaaatg
     4681 aggaagagaa caccccttca caatgacaac agctggatcc gacagcgcag tgccagtgtc
     4741 aacaaagagc ctgttagtct tcctgggatc atgagaagag gcgaatcttt agataacctg
     4801 gactccccc gatccaattc ttggagacag cctccttggc tcaatcagcc cacaggattc
     4861 tatgcttctt cctctgtgca agactttagt cgcccaccac ctcagctggt gtccacatca
     4921 aaccgtgcct acatgcggaa cccctcctcc agcgtgcccc caccttcagc tggctccgtg
     4981 aagacctcca ccacaggtgt ggccaccaca cagtcccca ccccgagaag ccattcccct
     5041 tcagcttcac agtcaggctc tcagctgcgt aacaggtcag tcagtgggaa gcgcatatgc
     5101 tcctactgca ataacattct gggcaaagga gccgccatga tcatcgagtc cctgggtctt
     5161 tgttatcatt tgcattgttt taagtgtgtt gcctgtgagt gtgacctcgg aggctcttcc
     5221 tcaggagctg aagtcaggat cagaaaccac caactgtact gcaacgactg ctatctcaga
     5281 ttcaaatctg gacggccaac cgccatgtga tgtaagcctc catacgaaag cactgttgca
     5341 gatagaagaa gaggtggttg ctgctcatgt agatctataa atatgtgttg tatgtctttt
     5401 ttgctttttt tttaaaaaaa agaataactt tttttgcctc tttagattac atagaagcat
     5461 tgtagtcttg gtagaaccag tatttttgtt gtttatttat aaggtaattg tgtgtgggga
     5521 aaagtgcagt atttacctgt tgaattcagc atcttgagag cacaagggaa aaaataagaa
     5581 cctacgaata tttttgaggc agataatgat ctagtttgac tttctagtta gtggtgtttt
     5641 gaagagggta ttttattgtt ttttaaaaaa aggttcttaa acattatttg aaatagttaa
     5701 tataaataca taattgcatt tgctctgttt attgtaatgt attctaaatt aatgcagaac
     5761 catatggaaa atttcattaa aatctatccc caaatgtgct ttctgtatcc ttccttctac
     5821 ctattattct gatttttaaa aatgcagtta atgtaccatt tatttgcttg atgaagggag
     5881 ctctattttc tttaccagaa atgttgctaa gtaattccca atagaaagct gcttattttc
     5941 attaatgaaa aataaccatg gtttgtatac tagaagtctt cttcagaaac tggtgagcct
     6001 ttctgttcaa ttgcatttgt aaataaactt gctgatgcat ttaacgagtg ggtcgtcttt
     6061 ttcttaggtg tatgtgtctg acctcaggcc ttttagccat atttcagtat gtggcctttt
     6121 ttgatgttat gttttatcca gtagctttac taaggtataa ttgatgtaat aaactgcata
     6181 tatttaaagt gtatactttg acaaattttg acatggtgta taccttcgaa actatgccac
     6241 agtctggatg tgtttactga aacattttaa taaggaagtt tatttttgat aaagttatgt
     6301 ttttggatac aatatatttg tatggtgaga gtgatgaatt gttggatcat ttgatctttt
     6361 actaacccca tgataaaagg agaagacaac agtgagctta gaatatctat aaagcaaaaa
     6421 atgtagtctc ttgtttaaaa aatctggagc gggaatgcaa ggatacaaaa ctttagcatg
     6481 ctttgagcaa aaatttaaac ttactggaat cttttataat aatgtaagtg gaatggagga
     6541 ttctaggaac tgagaactgt attggaatag gttcaaaata tgtaagaaat gctaatgtgg
     6601 gagataaaaa ttttatttag tacttattct gattattatt aaagtaataa tgtgttcctt
     6661 gaggataact tgtcaaatgc cccaaagcat aaagaatata attctgaatc ccaaattcca
     6721 aagacaagaa ctctgtgttt gaattcattc tgcatataat tatttataag tatagattgt
     6781 gaatttttcc atgttcttaa aattattttt atcttttttc atggttgcat agtgctccat
     6841 tgtttggcct tggtaatatt tagttgataa ttccattact gtgtattttt cacttgtttc
     6901 taagatcaaa cattttaata tgtgcatgtt atatataaat atgtaaattc tgtgatactc
     6961 tatgatcatc tctttcttta tattattttc atagacatga aatagttgct cagagattat
     7021 gcattttaag acactcatag tatatattgc caaagtggtt tccagaaagg cactgctggc
     7081 ttcgactcct ataagcagca cgtgggcttg ttcatctcac tgcatgttta tgaagataca
     7141 gttcttttgc cttgttctct gcctgatgtg tatgcagagg cagccctcaa tatgcagtgg
     7201 ttgaataaat gaatgaagaa accactaaaa aaaaa
```



Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**LMO7_HUMAN**
Primary accession number	**Q8WWI1**
Secondary accession numbers	O15462 O95346 Q9UKC1 Q9UQM5 Q9Y6A7
Integrated into Swiss-Prot on	March 15, 2004
Sequence was last modified on	March 15, 2004 (Sequence version 2)
Annotations were last modified on	July 25, 2006 (Entry version 39)
Name and origin of the protein	
Protein name	**LIM domain only protein 7**
Synonyms	**LOMP** **F-box only protein 20**
Gene name	**Name:** **LMO7** Synonyms: FBX20, FBXO20, KIAA0858
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

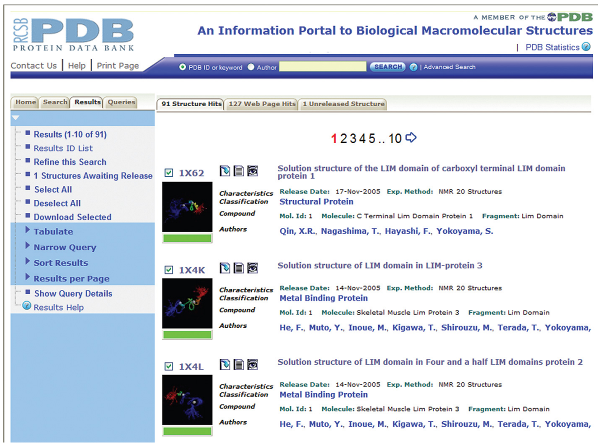
References

[1] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 3), AND TISSUE SPECIFICITY.
TISSUE=Brain, and Peripheral blood leukocyte;
DOI=10.1007/s00439-001-0646-6; PubMed=11935316 [NCBI, ExPASy, EBI, Israel, Japan]
Rozenblum E., Vahteristo P., Sandberg T., Bergthorsson J.T., Syrjakoski K., Weaver D., Haraldsson K., Johannsdottir H.K., Vehmanen P., Nigam S., Golberger N., Robbins C., Pak E., Dutra A., Gillander E., Stephan D.A., Bailey-Wilson J., Juo S.-H.H., Kainu T., ⇔ , Kallioniemi O.-P.;
"A genomic map of a 6-Mb region at 13q21-q22 implicated in cancer development: identification and characterization of candidate genes.";
Hum. Genet. 110:111-121(2002).

Key	From	To	Length	Description	FTId
CHAIN	1	1683	1683	LIM domain only protein 7.	PRO_0000075824
DOMAIN	54	168	115	CH.	
DOMAIN	1042	1128	87	PDZ.	
DOMAIN	1612	1678	67	LIM zinc-binding.	

```
        10         20         30         40         50         60
MKKIRICHIF TFYSWMSYDV LFQRTELGAL EIWRQLICAH VCICVGWLYL RDRVCSKKDI

        70         80         90        100        110        120
ILRTEQNSGR TILIKAVTEK NFETKDFRAS LENGVLLCDL INKLKPGVIK KINRLSTPIA

       130        140        150        160        170        180
GLDNINVFLK ACEQIGLKEA QLFHPGDLQD LSNRVTVKQE ETDRRVKNVL ITLYWLGRKA
```

Structural database



☑ 1X62 — Solution structure of the LIM domain of carboxyl terminal LIM domain protein 1

Characteristics Release Date: 17-Nov-2005 Exp. Method: NMR 20 Structures
Classification Structural Protein
Compound Mol. Id: 1 Molecule: C Terminal Lim Domain Protein 1 Fragment: Lim Domain
Authors Qin, X.R., Nagashima, T., Hayashi, F., Yokoyama, S.

☑ 1X4K — Solution structure of LIM domain in LIM-protein 3

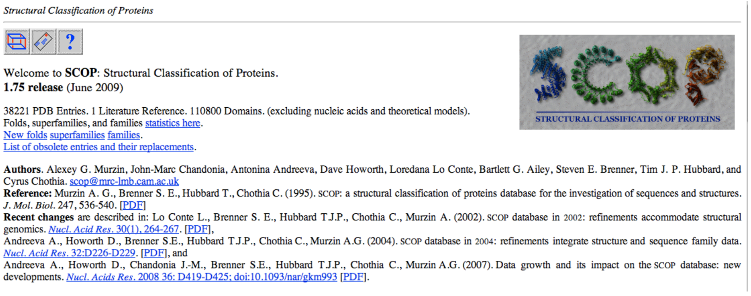
Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures
Classification Metal Binding Protein
Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain
Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

☑ 1X4L — Solution structure of LIM domain in Four and a half LIM domains protein 2

Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures
Classification Metal Binding Protein
Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain
Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

They contain information about the structure of small molecules, proteins, DNA and RNA sequences, carbohydrates.

Protein folds have also been classified according to the conservation of the fold. They include CATH and SCOP.



Structural Classification of Proteins

Welcome to **SCOP**: Structural Classification of Proteins.
1.75 release (June 2009)

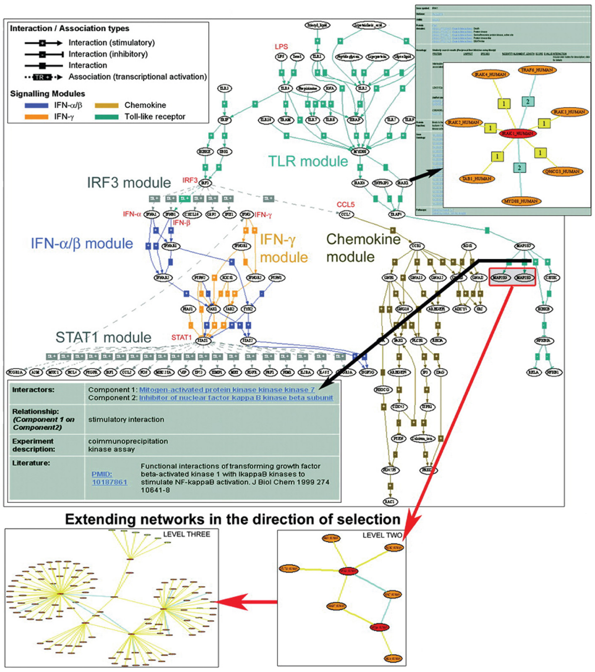
38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF],
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 2008 36: D419-D425; doi:10.1093/nar/gkm993 [PDF].

Protein interaction databases

They provide information about the interactions of proteins with other molecules, including other proteins.

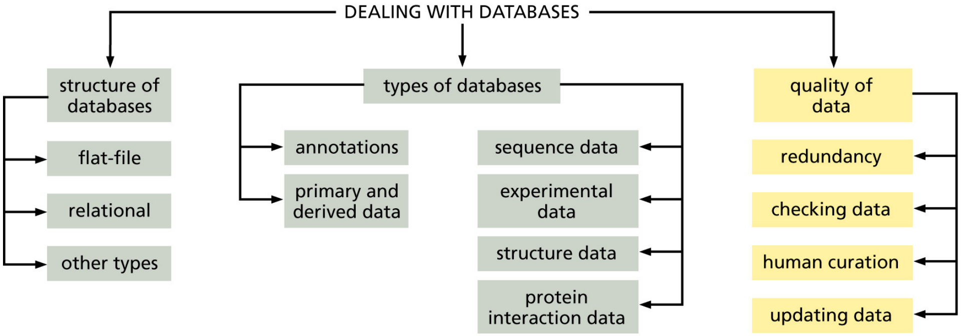
They include: the Database of Interacting Proteins (DIP) and the Molecular INTeraction Database (MINT).



Interactors:	Component 1: Mitogen-activated protein kinase kinase kinase 7 Component 2: Inhibitor of nuclear factor kappa B kinase beta subunit
Relationship: *(Component 1 on Component2)*	stimulatory interaction
Experiment description:	coimmunoprecipitation kinase assay
Literature:	PMID: 10187861 Functional interactions of transforming growth factor beta-activated kinase 1 with IkappaB kinases to stimulate NF-kappaB activation. J Biol Chem 1999 274 10641-8

Extending networks in the direction of selection

Quality of databases



Non reduntant databases: they include all the experimental data (from different labs) in one entry.

Checking data: a DNA seqeunce must contain only A, C, G, T. A protein sequence must correspond to a certain molecular weight according to the amino acids present.

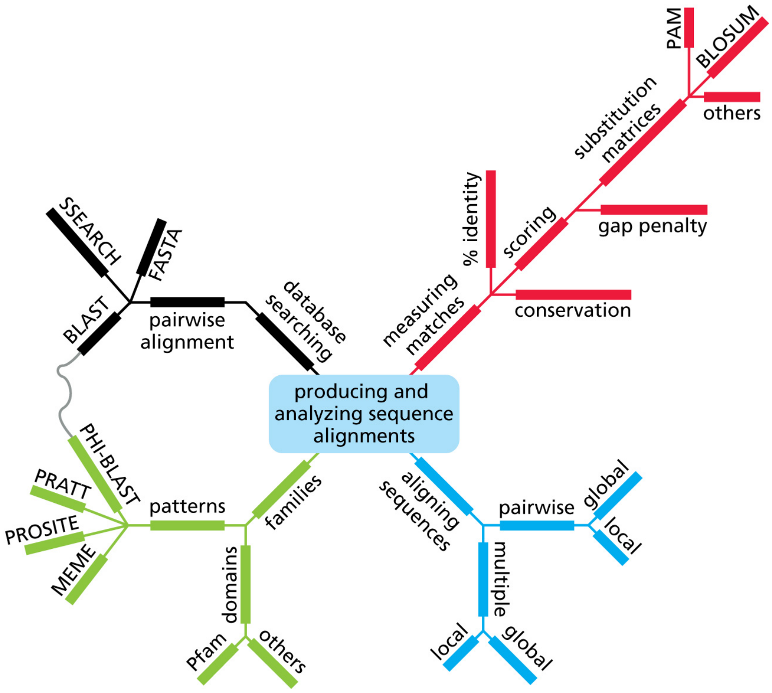
Sequence alignments

Useful for:

-comparing an unknown sequence to all the sequences contained in a database;

- prediction of a protein structure

- construction of phylogenetic trees



Sequence alignments

Alignment is the task of locating equivalent regions of two or more sequences to maximize their similarity.

THIS SEQUENCE

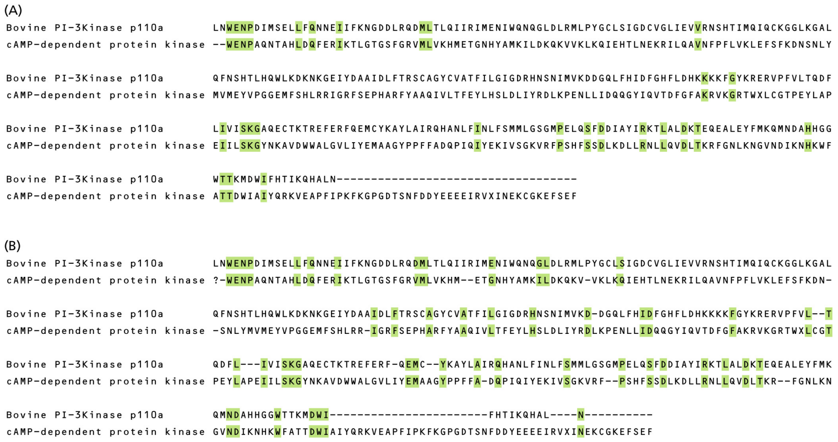
THAT SEQUENCE

The differences in length between two or more sequences can be compensated by the introduction of **GAPS**.

THISISA-SEQUENCE

TH----ATSEQUENCE

Gap penalty: each time a gap is introduced, the penalty is subtracted from the score, decreasing the overall score of the alignment.



```
(A)
Bovine PI-3Kinase p110a        LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase --WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLY

Bovine PI-3Kinase p110a        QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLTQDF
cAMP-dependent protein kinase MVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAP

Bovine PI-3Kinase p110a        LIVISKGAQECTKTREFERFQEMCYKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMKQMNDAHHGG
cAMP-dependent protein kinase EIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWF

Bovine PI-3Kinase p110a        WTTKMDWIFHTIKQHALN-------------------------------------
cAMP-dependent protein kinase ATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF


(B)
Bovine PI-3Kinase p110a        LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase ?-WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHM--ETGNHYAMKILDKQKV-VKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDN-

Bovine PI-3Kinase p110a        QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-DGQLFHIDFGHFLDHKKKKFGYKRERVPFVL--T
cAMP-dependent protein kinase -SNLYMVMEYVPGGEMFSHLRR-IGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGT

Bovine PI-3Kinase p110a        QDFL---IVISKGAQECTKTREFERF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMK
cAMP-dependent protein kinase PEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVRF--PSHFSSDLKDLLRNLLQVDLTKR--FGNLKN

Bovine PI-3Kinase p110a        QMNDAHHGGWTTKMDWI-----------------------FHTIKQHAL----N----------
cAMP-dependent protein kinase GVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF
```

A) An alignment where the gap penalty has been set very high.

B) An alignment with a very long gap penalty. Many more gaps have been introduced.

Sequence alignments

Similarity: the sequences show some degree of match.

Homology: similarity in sequence or structure due to descent from a common ancestor.

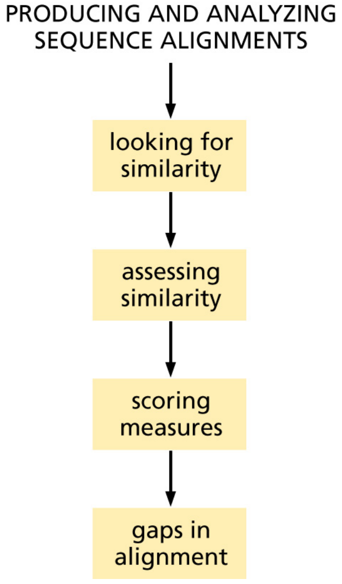
Mutation and selection over millions of years can result in considerable divergence between present-day sequences derived from the same ancestral gene.

Bases at originally same position can change as a result of:

- Mutations
- Insertions
- Deletions
- Gene fusions

Homology ⇒ common ancestor ⇒ common structure or function?

Not always……



PRODUCING AND ANALYZING SEQUENCE ALIGNMENTS

looking for similarity

assessing similarity

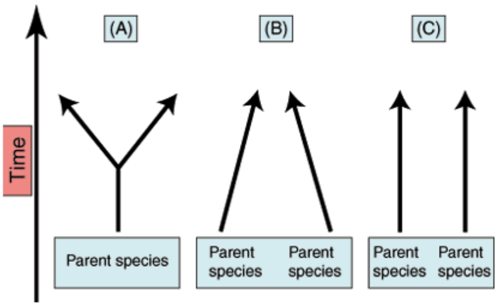
scoring measures

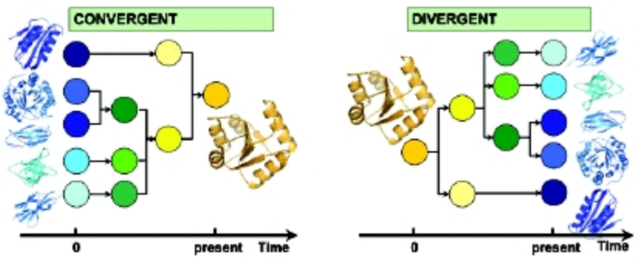
gaps in alignment

Sequence alignments

<u>Divergent evolution</u>: mutation and selection can generate proteins with new functions but relatively little changes in sequence. Therefore, sequence similarity does not always imply a common function.

<u>Convergent evolution</u>: proteins with very little sequence similarity to each other but in which a common protein fold and function are preserved.





It is easier to compare to detect homology when comparing protein sequence than when comparing nucleic acid sequences.

1. There are only 4 letters to compare in the DNA alphabet compared to the 20 letters in the protein one
2. The genetic code is redundant
3. The 3D structure of a protein and hence its function, is determined by the amino acid sequence

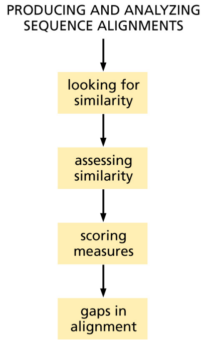
Scoring alignments

The quality of an alignment is measured by giving it a quantitative score

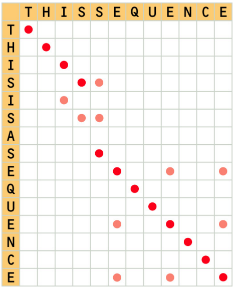
Percent identity: obtained by dividing the number of identical matches by the total length of the aligned region and multiplying by 100.

A good percentage of identity depends on the length of the sequence.

Substitution matrices: the score is assigned to each aligned pair of amino acids by a matrix that defines values for all possible pairs of residues.



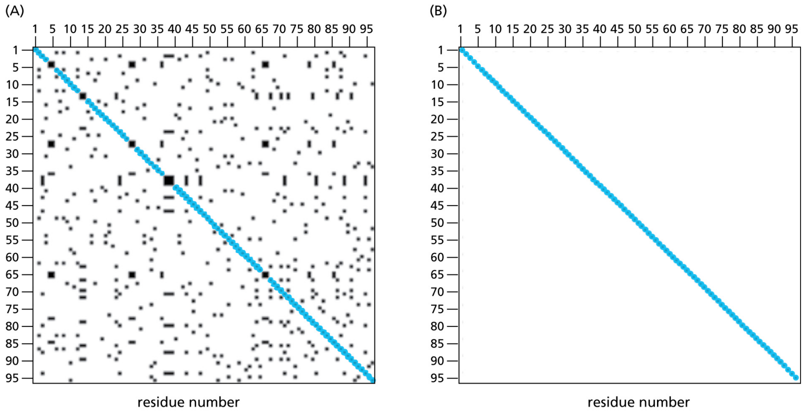
Scoring alignments: identity percentage and similarity percentage



Dot-plots: it is the simplest way to compare sequence similarities.

Use of filters:

- Window size allows to overlap fixed-length windows
- Minimum identity score: it is the minimum identity score fixed for the window previously set.

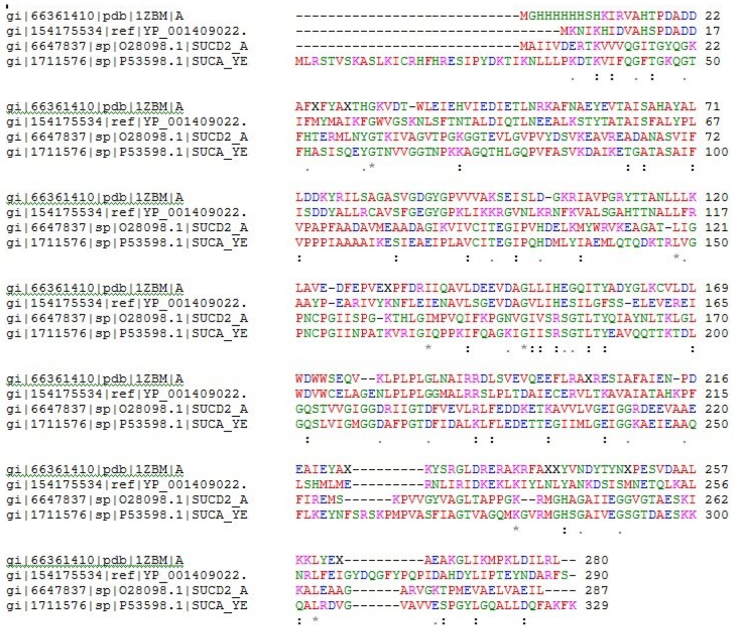


Two views of dot-plot representations of an SH2 sequence compared to itself. A) Unfiltered dot-plot. The identity is shown by the unbroken diagonale. There is some background noise. B) Dot-plot of the same sequence comparison with a window of 10 residues and a minimum identity score within the window set to 3.

Scoring alignments: identity percentage and similarity percentage

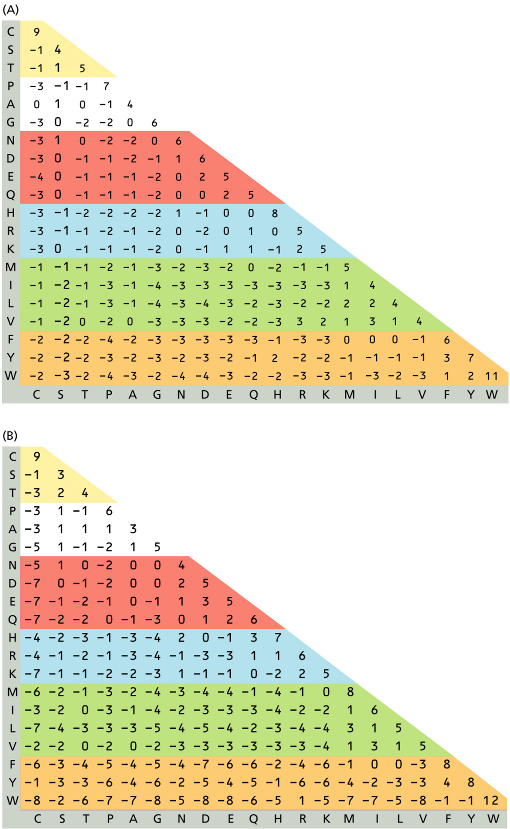
Similarity percentage: it takes into account the so-called conservative substitution

```
T H I S I S A - S E Q U E N C E
T H - - - - A T S E Q U E N C E

T H I S I S A - S E Q U E N C E
T H A T - - - - S E Q U E N C E
```



```
gi|66361410|pdb|1ZBM|A              ----------------------------MGHHHHHHSHKIRVAHTPDADD 22
gi|154175534|ref|YP_001409022.      ---------------------------------MKNIKHIDVAHSPDADD 17
gi|6647837|sp|O28098.1|SUCD2_A      ----------------------------MAIIVDERTKVVVQGITGYQGK 22
gi|1711576|sp|P53598.1|SUCA_YE      MLRSTVSKASLKICRHFHRESIPYDKTIKNLLLPKDTKVIFQGFTGKQGT 50
                                                                      .  : :  . :   .

gi|66361410|pdb|1ZBM|A              AFXFYAXTHGKVDT-WLEIEHVIEDIETLNRKAFNAEYEVTAISAHAYAL 71
gi|154175534|ref|YP_001409022.      IFMYMAIKFGWVGSKNLSFTNTALDIQTLNEEALKSTYTATAISFALYPL 67
gi|6647837|sp|O28098.1|SUCD2_A      FHTERMLNYGTKIVAGVTPGKGGTEVLGVPVYDSVKEAVREADANASVIF 72
gi|1711576|sp|P53598.1|SUCA_YE      FHASISQEYGTNVVGGTNPKKAGQTHLGQPVFASVKDAIKETGATASAIF 100
                                     .      .*          :                    : :     :

gi|66361410|pdb|1ZBM|A              LDDKYRILSAGASVGDGYGPVVVAKSEISLD-GKRIAVPGRYTTANLLLK 120
gi|154175534|ref|YP_001409022.      ISDDYALLRCAVSFGEGYGPKLIKKRGVNLKRNFKVALSGAHTTNALLFR 117
gi|6647837|sp|O28098.1|SUCD2_A      VPAPFAADAVMEAADAGIKVIVCITEGIPVHDELKMYWRVKEAGAT-LIG 121
gi|1711576|sp|P53598.1|SUCA_YE      VPPPIAAAAIKESIEAEIPLAVCITEGIPQHDMLYIAEMLQTQDKTRLVG 150
                                    :           :        :  .  :  .    :           *.

gi|66361410|pdb|1ZBM|A              LAVE-DFEPVEXPFDRIIQAVLDEEVDAGLLIHEGQITYADYGLKCVLDL 169
gi|154175534|ref|YP_001409022.      AAYP-EARIVYKNFLEIENAVLSGEVDAGVLIHESILGFSS-ELEVEREI 165
gi|6647837|sp|O28098.1|SUCD2_A      PNCPGIISPG-KTHLGIMPVQIFKPGNVGIVSRSGTLTYQIAYNLTKLGL 170
gi|1711576|sp|P53598.1|SUCA_YE      PNCPGIINPATKVRIGIQPPKIFQAGKIGIISRSGTLTYEAVQQTTKTDL 200
                                                    *    :    . *:: :.. : :          :

gi|66361410|pdb|1ZBM|A              WDWWSEQV--KLPLPLGLNAIRRDLSVEVQEEFLRAXRESIAFAIEN-PD 216
gi|154175534|ref|YP_001409022.      WDVWCELAGENLPLPLGGMALRRSLPLTDAIECERVLTKAVAIATAHKPF 215
gi|6647837|sp|O28098.1|SUCD2_A      GQSTVVGIGGDRIIGTDFVEVLRLFEDDKETKAVVLVGEIGGRDEEVAAE 220
gi|1711576|sp|P53598.1|SUCA_YE      GQSLVIGMGGDAFPGTDFIDALKLFLEDETTEGIIMLGEIGGKAEIEAAQ 250
                                     :        .     .     : :      :      :  .      .

gi|66361410|pdb|1ZBM|A              EAIEYAX---------KYSRGLDRERAKRFAXXYVNDYTYNXPESVDAAL 257
gi|154175534|ref|YP_001409022.      LSHMLME---------RNLIRIDKEKLKIYLNLYANKDSISMNETQLKAL 256
gi|6647837|sp|O28098.1|SUCD2_A      FIREMS------KPVVGYVAGLTAPPGK--RMGHAGAIIEGGVGTAESKI 262
gi|1711576|sp|P53598.1|SUCA_YE      FLKEYNFSRSKPMPVASFIAGTVAGQMKGVRMGHSGAIVEGSGTDAESKK 300
                                                               *     : .    .

gi|66361410|pdb|1ZBM|A              KKLYEX----------AEAKGLIKMPKLDILRL-- 280
gi|154175534|ref|YP_001409022.      NRLFEIGYDQGFYPQPIDAHDYLIPTEYNDARFS- 290
gi|6647837|sp|O28098.1|SUCD2_A      KALEAAG------ARVGKTPMEVAELVAEIL---- 287
gi|1711576|sp|P53598.1|SUCA_YE      QALRDVG------VAVVESPGYLGQALLDQFAKFK 329
                                    : *              .:   :     :
```

Scoring alignments: substitution matrices



(A)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	3	2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11

(B)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	−1	3																		
T	−3	2	4																	
P	−3	1	−1	6																
A	−3	1	1	1	3															
G	−5	1	−1	−2	1	5														
N	−5	1	0	−2	0	0	4													
D	−7	0	−1	−2	0	0	2	5												
E	−7	−1	−2	−1	0	−1	1	3	5											
Q	−7	−2	−2	0	−1	−3	0	1	2	6										
H	−4	−2	−3	−1	−3	−4	2	0	−1	3	7									
R	−4	−1	−2	−1	−3	−4	−1	−3	−3	1	1	6								
K	−7	−1	−1	−2	−2	−3	1	−1	−1	0	−2	2	5							
M	−6	−2	−1	−3	−2	−4	−3	−4	−4	−1	−4	−1	0	8						
I	−3	−2	0	−3	−1	−4	−2	−3	−3	−3	−4	−2	−2	1	6					
L	−7	−4	−3	−3	−3	−5	−4	−5	−4	−2	−3	−4	−4	3	1	5				
V	−2	−2	0	−2	0	−2	−3	−3	−3	−3	−3	−3	−4	1	3	1	5			
F	−6	−3	−4	−5	−4	−5	−4	−7	−6	−6	−2	−4	−6	−1	0	0	−3	8		
Y	−1	−3	−3	−6	−4	−6	−2	−5	−4	−5	−1	−6	−6	−4	−2	−3	−3	4	8	
W	−8	−2	−6	−7	−7	−8	−5	−8	−8	−6	−5	1	−5	−7	−7	−5	−8	−1	−1	12

Expectation value (E-value): the probability of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

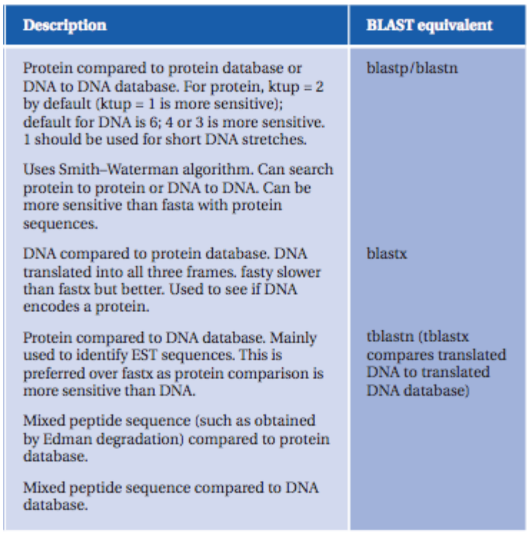
It indicates the number of sequences that would be expected to have that score (or more) if the query sequence were compared against a database containing no sequences related to the query sequence. Thus, a lower E-value indicates that the sequences are more likely to be related than if the comparison had a higher E-value. An E-value of 0.00001 or less (also sometimes written as 1e-5, which is shorthand for 1.0 * 10^{-5}) is often used as good initial evidence that a query and database sequence are related, although further investigation should always be carried out to obtain additional support for such a hypothesis.

Amino acids substitution scoring matrices. A) The BLOSUM-62 matrix and B) the PAM120 matrix. The colored shading indicates different physicochemical properties of the residues.

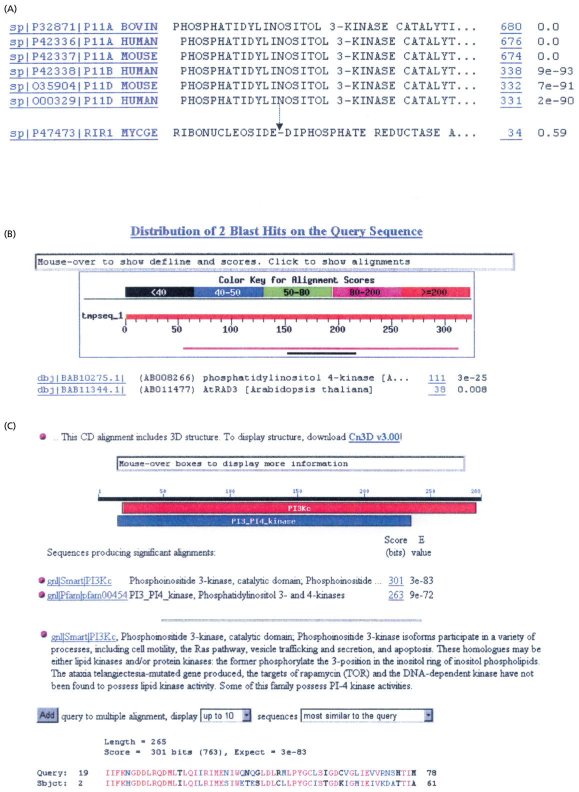
Sequence alignments: BLAST

NCBI

results of BLAST



Description	BLAST equivalent
Protein compared to protein database or DNA to DNA database. For protein, ktup = 2 by default (ktup = 1 is more sensitive); default for DNA is 6; 4 or 3 is more sensitive. 1 should be used for short DNA stretches.	blastp/blastn
Uses Smith–Waterman algorithm. Can search protein to protein or DNA to DNA. Can be more sensitive than fasta with protein sequences.	
DNA compared to protein database. DNA translated into all three frames. fasty slower than fastx but better. Used to see if DNA encodes a protein.	blastx
Protein compared to DNA database. Mainly used to identify EST sequences. This is preferred over fastx as protein comparison is more sensitive than DNA.	tblastn (tblastx compares translated DNA to translated DNA database)
Mixed peptide sequence (such as obtained by Edman degradation) compared to protein database.	
Mixed peptide sequence compared to DNA database.	



(A)

```
sp|P32871|P11A BOVIN  PHOSPHATIDYLINOSITOL 3-KINASE CATALYTI...   680  0.0
sp|P42336|P11A HUMAN   PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...   676  0.0
sp|P42337|P11A MOUSE   PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...   674  0.0
sp|P42338|P11B HUMAN   PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...   338  9e-93
sp|035904|P11D MOUSE   PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...   332  7e-91
sp|000329|P11D HUMAN   PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...   331  2e-90

sp|P47473|RIR1 MYCGE  RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE A...    34  0.59
```

(B)

Distribution of 2 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

```
dbj|BAB10275.1|  (AB008266) phosphatidylinositol 4-kinase [A...   111  3e-25
dbj|BAB11344.1|  (AB011477) AtRAD3 [Arabidopsis thaliana]          38  0.008
```

(C)

● .. This CD alignment includes 3D structure. To display structure, download Cn3D v3.00!

Mouse-over boxes to display more information

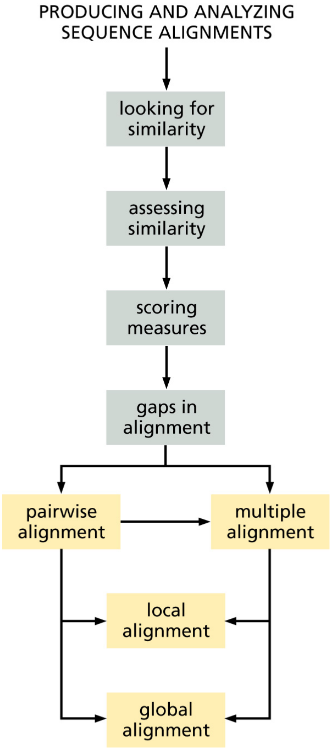
```
                                                              Score   E
Sequences producing significant alignments:                   (bits)  value

● gnl|Smart|PI3Kc       Phosphoinositide 3-kinase, catalytic domain; Phosphoinositide ...   301  3e-83
● gnl|Pfam|pfam00454 PI3_PI4_kinase, Phosphatidylinositol 3- and 4-kinases                  263  9e-72
```

● gnl|Smart|PI3Kc, Phosphoinositide 3-kinase, catalytic domain; Phosphoinositide 3-kinase isoforms participate in a variety of processes, including cell motility, the Ras pathway, vesicle trafficking and secretion, and apoptosis. These homologues may be either lipid kinases and/or protein kinases: the former phosphorylate the 3-position in the inositol ring of inositol phospholipids. The ataxia telangiectesia-mutated gene produced, the targets of rapamycin (TOR) and the DNA-dependent kinase have not been found to possess lipid kinase activity. Some of this family possess PI-4 kinase activities.

Add query to multiple alignment, display up to 10 sequences most similar to the query

```
            Length = 265
            Score =  301 bits (763), Expect = 3e-83

Query:  19    IIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIM  78
Sbjct:  2     IIFKHGDDLRQDMLILQILRIMESIWETESLDLCLLPYGCISTGDKIGMIEIVKDATTIA  61
```

Types of alignments



PRODUCING AND ANALYZING SEQUENCE ALIGNMENTS

looking for similarity

assessing similarity

scoring measures

gaps in alignment

pairwise alignment

multiple alignment

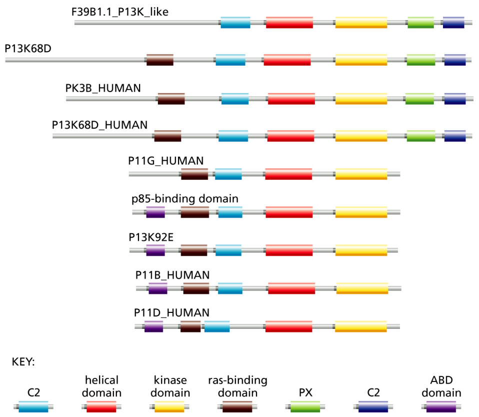
local alignment

global alignment

Global alignment: it is used to find or compare closely related sequences that are similar over their whole sequence.

Local alignment: can reveal that parts of sequences are related.

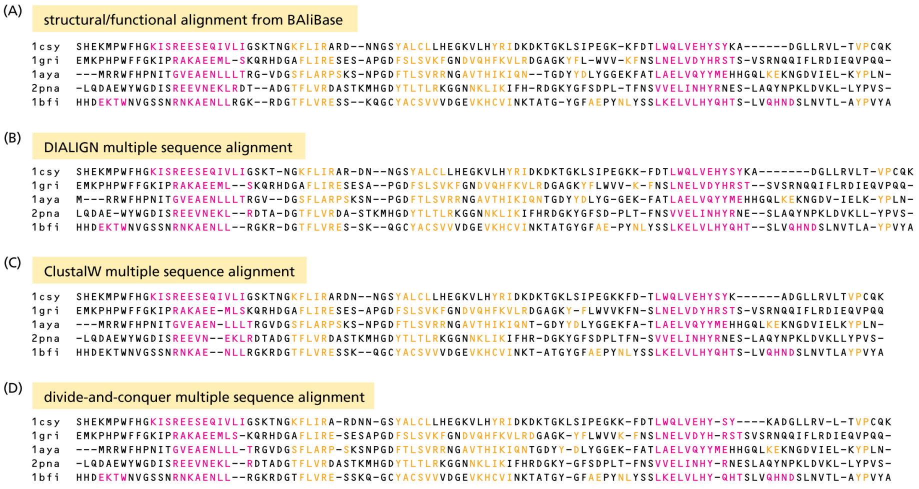
It is useful in multidomain proteins.



PI3-kinase is a multidomain protein. Output from Pfam.

Multiple alignments

They can be constructued by different techniques.



(A) structural/functional alignment from BAliBase

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARD--NNGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGK-KFDTLWQLVEHYSYKA------DGLLRVL-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEML-SKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKYFL-WVV-KFNSLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLLTRG-VDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQN--TGDYYDLYGGEKFATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFH-RDGKYGFSDPL-TFNSVVELINHYRNES-LAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLLRGK--RDGTFLVRESS--KQGCYACSVVVDGEVKHCVINKTATG-YGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTL-AYPVYA
```

(B) DIALIGN multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRAR-DN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHYSYKA-------DGLLRVLT-VPCQK
1gri  EMKPHPWFFGKIPRAKAEEML--SKQRHDGAFLIRESESA--PGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVV-K-FNSLNELVDYHRST--SVSRNQQIFLRDIEQVPQQ-
1aya  M---RRWFHPNITGVEAENLLLTRGV--DGSFLARPSKSN--PGDFTLSVRRNGAVTHIKIQNTGDYYDLYG-GEK-FATLAELVQYYMEHHGQLKEKNGDV-IELK-YPLN-
2pna  LQDAE-WYWGDISREEVNEKL--RDTA-DGTFLVRDA-STKMHGDYTLTLRKGGNNKLIKIFHRDGKYGFSD-PLT-FNSVVELINHYRNE--SLAQYNPKLDVKLL-YPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKR-DGTFLVRES-SK--QGCYACSVVVDGEVKHCVINKTATGYGFAE-PYNLYSSLKELVLHYQHT--SLVQHNDSLNVTLA-YPVYA
```

(C) ClustalW multiple sequence alignment

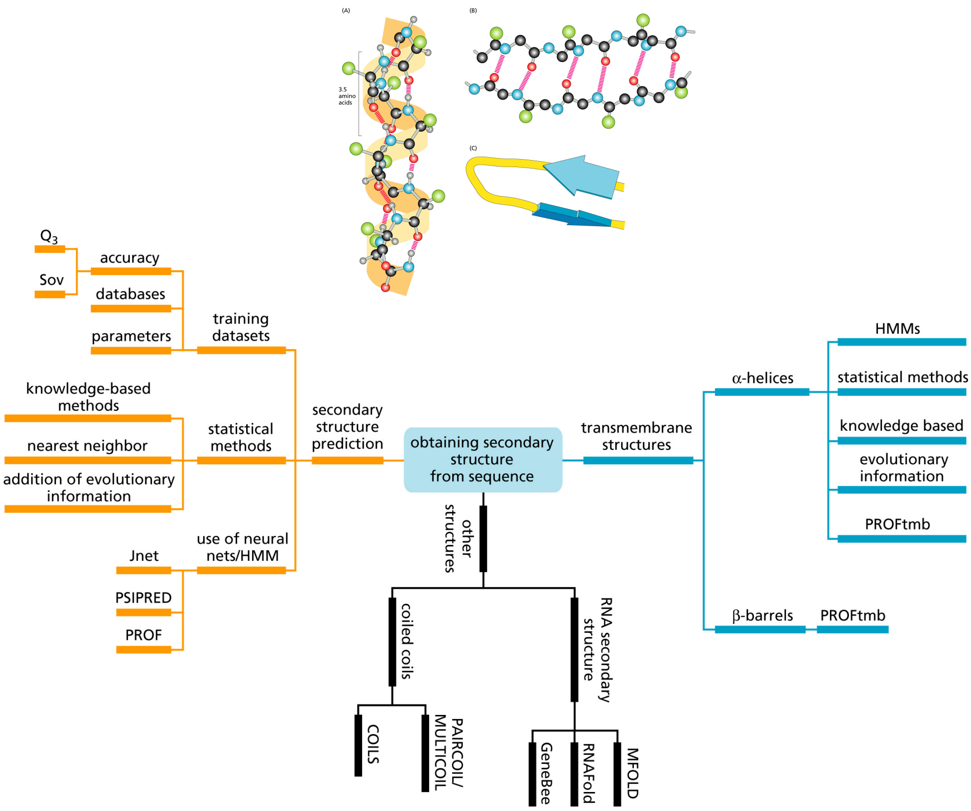
```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARDN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKKFD-TLWQLVEHYSYK------ADGLLRVLTVPCQK
1gri  EMKPHPWFFGKIPRAKAEE-MLSKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKY-FLWVVKFN-SLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ
1aya  ---MRRWFHPNITGVEAEN-LLLTRGVDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQNT-GDYYDLYGGEKFA-TLAELVQYYMEHHGQLKEKNGDVIELKYPLN-
2pna  -LQDAEWYWGDISREEVN--EKLRDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHR-DGKYGFSDPLTFN-SVVELINHYRNES-LAQYNPKLDVKLLYPVS-
1bfi  HHDEKTWNVGSSNRNKAE--NLLRGKRDGTFLVRESSK--QGCYACSVVVDGEVKHCVINKT-ATGYGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTLAYPVYA
```

(D) divide-and-conquer multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRA-RDNN-GSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHY-SY----KADGLLRV-L-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEMLS-KQRHDGAFLIRE-SESAPGDFSLSVKFGNDVQHFKVLRDGAGK-YFLWVVK-FNSLNELVDYH-RSTSVSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLL-TRGVDGSFLARP-SKSNPGDFTLSVRRNGAVTHIKIQNTGDYY-DLYGGEK-FATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKL--RDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHRDGKY-GFSDPLT-FNSVVELINHY-RNESLAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKRDGTFLVRE-SSKQ-GCYACSVVVDGEVKHCVINKTATGY-GFAEPYNLYSSLKELVLHY-QHTSLVQHNDSLNVTL-AYPVYA
```

Structural alignments: if the structure of one of the proteins is known, then the gap penalty can be increased for regions of known secondary structure such as helices and strands, as these regions are less likely to suffer insertions or deletions. This will mean that few or no gaps are introduced into these regions.

Protein secondary structure prediction



obtaining secondary structure from sequence

secondary structure prediction

training datasets

accuracy

Q_3

Sov

databases

parameters

statistical methods

knowledge-based methods

nearest neighbor

addition of evolutionary information

use of neural nets/HMM

Jnet

PSIPRED

PROF

transmembrane structures

α-helices

HMMs

statistical methods

knowledge based

evolutionary information

PROFtmb

β-barrels

PROFtmb

other structures

coiled coils

COILS

PAIRCOIL/ MULTICOIL

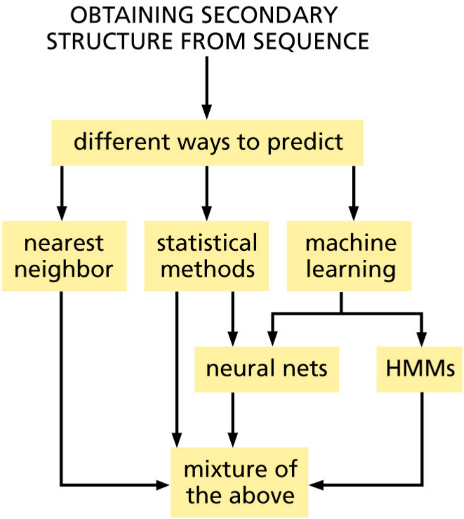
RNA secondary structure

GeneBee

RNAFold

MFOLD

Types of secondary structure prediction



OBTAINING SECONDARY STRUCTURE FROM SEQUENCE

different ways to predict

nearest neighbor

statistical methods

machine learning

neural nets

HMMs

mixture of the above

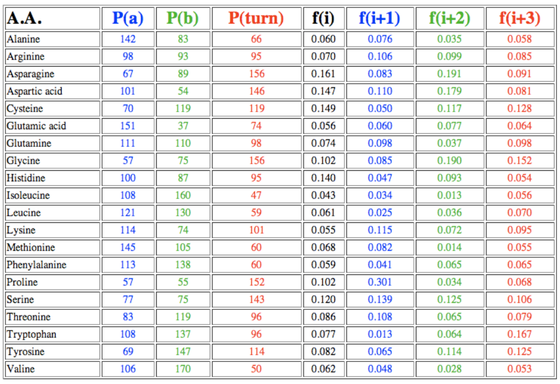
Statistical methods are based on rules that give the probability that a residue will form part of a particular secondary structure.

The probabilities are derived from analysing structure and sequence data from large sets of proteins of known structure.

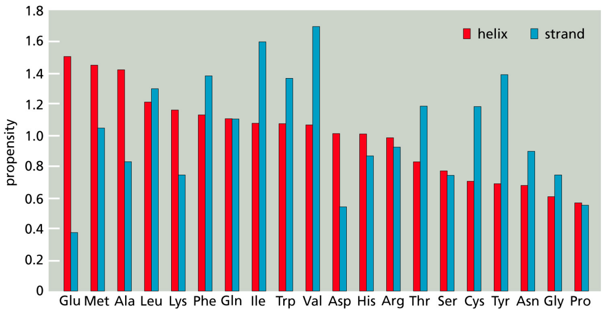
Nearest neighbor methods are statistical methods that incorporate additional information about protein structure (shapes, sizes and physicochemical properties of the different amino acid residues).

Machine learning approaches train a neural net or other learning alghoritms to aquire structure-sequence relationships which can then be applied to predict structure from a protein sequence.

Statistical and knowledge-based methods: Chou and Fasman



A.A.	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.060	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053



Chou-Fasman is one of most commonly used algorithms

- **measured frequencies at which each amino acid appeared in particular types of secondary sequences in a set of proteins of known structure**
- **assigns the amino acids three conformational parameters based on the frequency at which they were observed in alpha helices, beta sheets and beta turns**
  1. **P(a) = propensity to form alpha helices**
  2. **P(b) = propensity to form beta sheets**
  3. **P(turn) = propensity to form beta turns**
- **also assigns 4 turn parameters based on frequency at which they were observed in the first, second, third or fourth position of a beta turn**
  1. **f(i) = probability of being in position 1**
  2. **f(i+1) = probability of being in position 2**
  3. **f(i+2) = probability of being in position 3**
  4. **f(i+3) = probability of being in position 4**

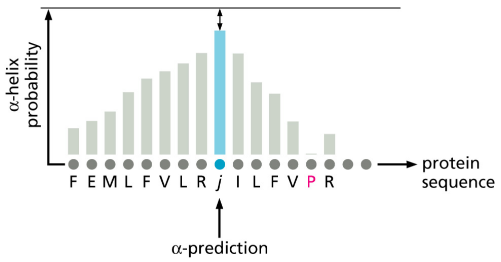
Statistical and knowledge-based methods: Chou and Fasman

identifies helix and sheet"nuclei", then applies a set of heuristic rules to determine if these clusters of amino acids are sufficient to nucleate a region of alpha-helix or beta-sheet.

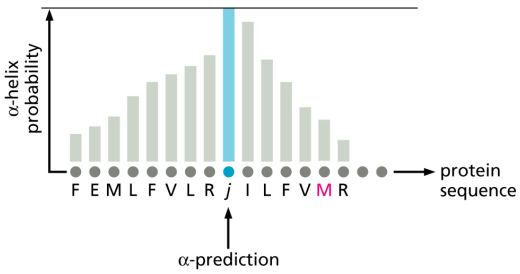
- **helix: 4 out of 6 amino acids with P(a) >100**
  - **extends the nucleus in each direction until reach four amino acids in a row with P(a) <100**
  - **for each of these regions, add up all the P(a) and all the P(b) values.**
  - **If the total P(a) is larger than the total of P(b) and the run is more than 5 amino acids long, then it is predicted to be alpha helix**
- **sheet: 4 out of 6 amino acids with P(b)>100 (some people use 3 out of 5).**
  - **extends the nucleus in each direction until reach four amino acids in a row with P(b) <100**
  - **for each of these regions, add up all the P(a) and all the P(b) values.**
  - **If the total P(b) is larger than the total of P(a), the run is more than 5 amino acids long, and the average P(b) > 100 then it is predicted to be beta sheet.**
- **If helices and sheets overlap then compare the total P(a) and total P(b) for the overlapping region. If the total P(a) is larger than the total of P(b) then it is predicted to be alpha helix (and vice-versa)**
- **beta turn**
  - **calculate the likelihood of a turn P(t)for amino acid at position i as the sum of f(i) + the f(i+1) value for the following amino acid + the f(i+2) value for the next amino acid+ the f(i+3) value for the amino acid at the plus three position.**
  - **Predict a beta- turn at position i if the following criteria are met:**
    - the calculated **P(t) is >0.5**
    - **the average P(turn) for amino acids i to i+3 is > 100**
    - **the sum of the P(turn) values for amino acids i to i+3 is larger than the sum of the P(a) and P(b)values**
- **Accuracy = 50-85%, depending on the protein**

Statistical and knowledge-based methods: GOR

It incorporates the effects of local interactions between amino acids residues by taking successive windows of 17 residues and considering the effect of residues from position j-8 to j+8 on the conformation of the residue at position j.

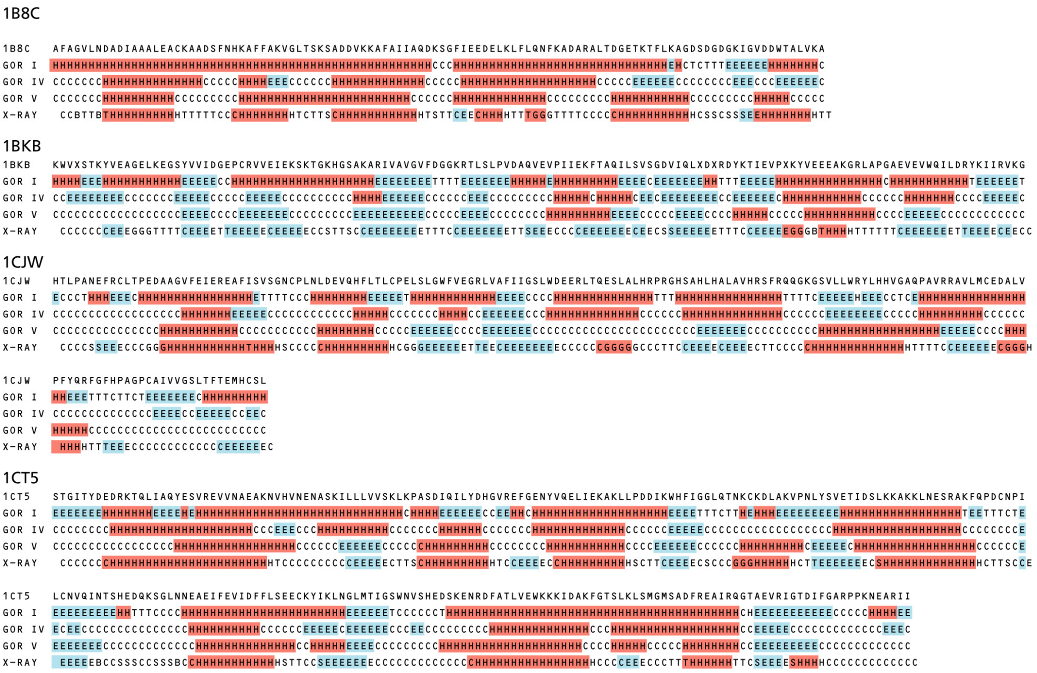


The effect of an helix breaker (Pro) at position j +5. The proline diminishes the overall additive propensity of residue j to form helix



The effect of a non helix breaker (Met) at position j+5. The methionine improves the overall additive propensity of residue j to form helix

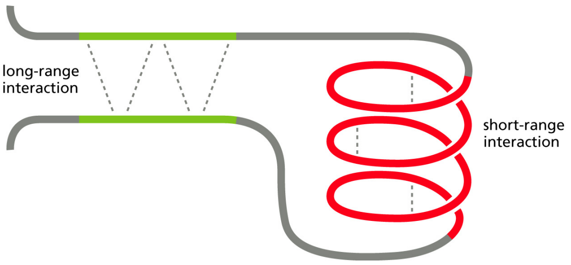
Statistical methods improvements: GOR I to V



```
1B8C

1B8C   AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIAQDKSGFIEEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDDWTALVKA
GOR I  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEHCTCTTTEEEEEEHHHHHHHC
GOR IV CCCCCCCHHHHHHHHHHHHHHCCCCCHHHHEEECCCCCCHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCCCCEEECCCEEEEEEEC
GOR V  CCCCCCCHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHCCCCC
X-RAY   CCBTTBTHHHHHHHHHHTTTTTCCCHHHHHHHHTCTTSCHHHHHHHHHHHHTSTTCEECHHHHTTTGGGTTTTCCCCCHHHHHHHHHHHCSSCSSSEEHHHHHHHHTT

1BKB
1BKB   KWVXSTKYVEAGELKEGSYVVIDGEPCRVVEIEKSKTGKHGSAKARIVAVGVFDGGKRTLSLPVDAQVEVPIIEKFTAQILSVSGDVIQLXDXRDYKTIEVPXKYVEEEAKGRLAPGAEVEVWQILDRYKIIRVKG
GOR I  HHHHEEEHHHHHHHHHHHEEEEECCHHHHHHHHHHHHHHHHHHHHEEEEEEEETTTTEEEEEEEHHHHHEHHHHHHHHHEEEECEEEEEEEHHTTTEEEEEHHHHHHHHHHHHHHHCHHHHHHHHHHHTEEEEEET
GOR IV CCEEEEEEEECCCCCCCEEEEECCCCCEEEEECCCCCCCCCCHHHHEEEEEECCCCCCEEECCCCCCCCCHHHHHCHHHHHCEECEEEEEEEECCEEEEEECHHHHHHHHHHHCCCCCCHHHHHHHCCCCEEEEEC
GOR V  CCCCCCCCCCCCCCCCCCEEEECCCCEEEEEEECCCCCCCCCEEEEEEEEEECCCCCEEEECCCCCCCCHHHHHHHHHEEEECCCCCEEEECCCCHHHHHCCCCCHHHHHHHHHHCCCCEEEEECCCCCCCCCCCC
X-RAY   CCCCCCCEEEGGGTTTTCEEEETTEEEEEECEEEEECCSTTSCCEEEEEEEETTTCCEEEEEEETTSEEECCCCEEEEEEEECEECSSEEEEEETTTCCEEEEEGGGBTHHHHTTTTTTCEEEEEEETTEEEECEECC

1CJW
1CJW   HTLPANEFRCLTPEDAAGVFEIEREAFISVSGNCPLNLDEVQHFLTLCPELSLGWFVEGRLVAFIIGSLWDEERLTQESLALHRPRGHSAHLHALAVHRSFRQQGKGSVLLWRYLHHVGAQPAVRRAVLMCEDALV
GOR I  ECCCTHHHEEECHHHHHHHHHHHHHHHHETTTTCCCHHHHHHHHEEEEETHHHHHHHHHHHHEEEECCCCHHHHHHHHHHHHHHTTTHHHHHHHHHHHHHHHTTTTCEEEEEHEEECCTCEHHHHHHHHHHHHHHH
GOR IV CCCCCCCCCCCCCCCCCCHHHHHHHEEEEECCCCCCCCCCCCHHHHHCCCCCCCHHHHCCEEEEEECCCHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHCCCCCCEEEEEEEECCCCCHHHHHHHHHCCCCCC
GOR V  CCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHCCCCCEEEEEECCCCEEEEEEECCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCCCCCCCHHHHHHHHHHHHHHHHHEEEEECCCCHHH
X-RAY   CCCCSSEEECCCGGGHHHHHHHHHHHTHHHHSCCCCCHHHHHHHHHHCGGGEEEEEETTEECEEEEEEEECCCCCCGGGGGCCCTTCCEEEECEEEECTTCCCCCHHHHHHHHHHHHHHTTTTCCEEEEEECGGGH

1CJW   PFYQRFGFHPAGPCAIVVGSLTFTEMHCSL
GOR I  HHEEETTTCTTCTEEEEEEECHHHHHHHHH
GOR IV CCCCCCCCCCCCCCEEEECCEEEEECCEEC
GOR V  HHHHHCCCCCCCCCCCCCCCCCCCCCCCCC
X-RAY   HHHHTTTEEECCCCCCCCCCCCCEEEEEEC

1CT5
1CT5   STGITYDEDRKTQLIAQYESVREVVNAEAKNVHVNENASKILLLVVSKLKPASDIQILYDHGVREFGENYVQELIEKAKLLPDDIKWHFIGGLQTNKCKDLAKVPNLYSVETIDSLKKAKKLNESRAKFQPDCNPI
GOR I  EEEEEEEHHHHHHHEEEEHEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHEEEEEECCEEHHCHHHHHHHHHHHHHHHHHHHEEEETTTCTTHEHHHEEEEEEEEEHHHHHHHHHHHHHHHHHTEETTTCTE
GOR IV CCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCEEECCCHHHHHHHHHHCCCCCCCHHHHHHCCCCCCCHHHHHHHHHHHHHCCCCCCEEEEECCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHCCCCCCCCE
GOR V  CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCCEEEEEECCCCCCHHHHHHHHHCCCCCCCCHHHHHHHHHHHCCCCEEEEEECCCCCCHHHHHHHHHCEEEEECHHHHHHHHHHHHHHHHHCCCCCCE
X-RAY   CCCCCCCHHHHHHHHHHHHHHHHHHHHHHHTCCCCCCCCCCEEEEECTTSCHHHHHHHHHHTCCEEEECCHHHHHHHHHHSCTTCEEEECSCCCGGGHHHHHHCTTEEEEEEECSHHHHHHHHHHHHHHCTTSCCE

1CT5   LCNVQINTSHEDQKSGLNNEAEIFEVIDFFLSEECKYIKLNGLMTIGSWNVSHEDSKENRDFATLVEWKKKIDAKFGTSLKLSMGMSADFREAIRQGTAEVRIGTDIFGARPPKNEARII
GOR I  EEEEEEEEEHHTTTCCCCHHHHHHHHHHHHHHHHHHHHHHHEEEEEETCCCCCCTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHEEEEEEEEEEECCCCCHHHHEE
GOR IV ECEECCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCEEEEECEEEEEECCCEECCCCCCCCCHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHCCEEEEECCCCCCCCCCCCCEEEC
GOR V  EEEEEEECCCCCCCCCCCCCHHHHHHHHHHHHHHCCHHHHHEEEECCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCCHHHHHCCCCCHHHHHHHHCCEEEEEEEEECCCCCCCCCCCCC
X-RAY   EEEEEBCCSSSCCSSSBCCHHHHHHHHHHHHSTTCCSEEEEEEECCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCEEECCCTTTHHHHHHTTCSEEEESHHHHCCCCCCCCCCCCC
```

Nearest neighbor methods

The formation of secondary structure in proteins does not only depend on local interactions (beta-sheets are made up of beta-strands that are separated from some distance in the poypeptide chain).

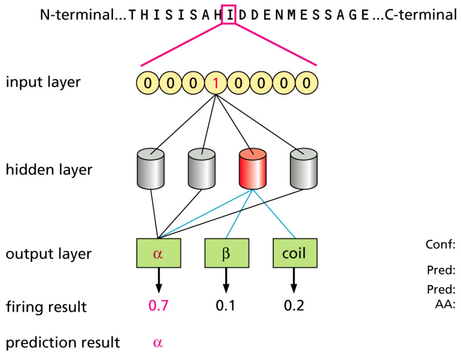


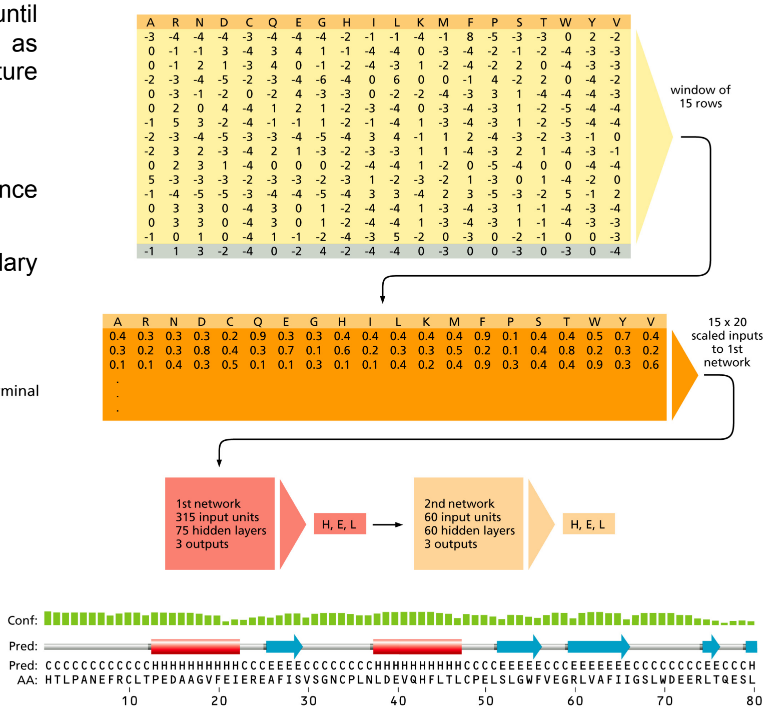
Neural networks methods

The algorithm will learn by iterative changes to its parameters until the predicted structure is as similar to the observed structure as possible.

PSIPRED is a three stage method:

1. It generates a multiple sequence alignment
2. It generates an initial secondary structure
3. It filters the initial prediction





A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3
-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	4	-2	2	0	-4	-2
0	-3	-1	-2	0	-2	4	-3	-3	0	-2	-2	-4	-3	3	1	-4	-4	-4	-3
0	2	0	4	-4	1	2	1	-2	-3	-4	0	-3	-4	-3	1	-2	-5	-4	-4
-1	5	3	-2	-4	-1	-1	1	-2	-1	-4	1	-3	-4	-3	1	-2	-5	-4	-4
-2	-3	-4	-5	-3	-3	-4	-5	-4	3	4	-1	1	2	-4	-3	-2	-3	-1	0
-2	3	2	-3	-4	2	1	-3	-2	-3	-3	1	1	-4	-3	2	1	-4	-3	-1
0	2	3	1	-4	0	0	0	-2	-4	-4	1	-2	0	-5	-4	0	0	-4	-4
5	-3	-3	-3	-2	-3	-3	-2	-3	1	-2	-3	-2	1	-3	0	1	-4	-2	0
-1	-4	-5	-5	-3	-4	-4	-5	-4	3	3	-4	2	3	-5	-3	-2	5	-1	2
0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3	-4
0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3	-3
-1	0	1	0	-4	1	-1	-2	-4	-3	5	-2	0	-3	0	-2	-1	0	0	-3
-1	1	3	-2	-4	0	-2	4	-2	-4	-4	0	-3	0	0	-3	0	-3	0	-4

window of 15 rows

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.3	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6
.																			
.																			
.																			

15 x 20 scaled inputs to 1st network

1st network
315 input units
75 hidden layers
3 outputs

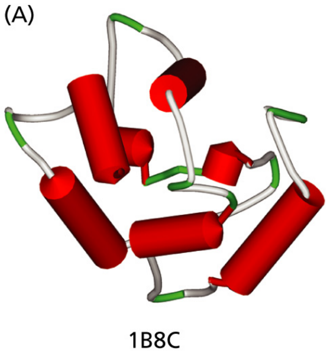
H, E, L

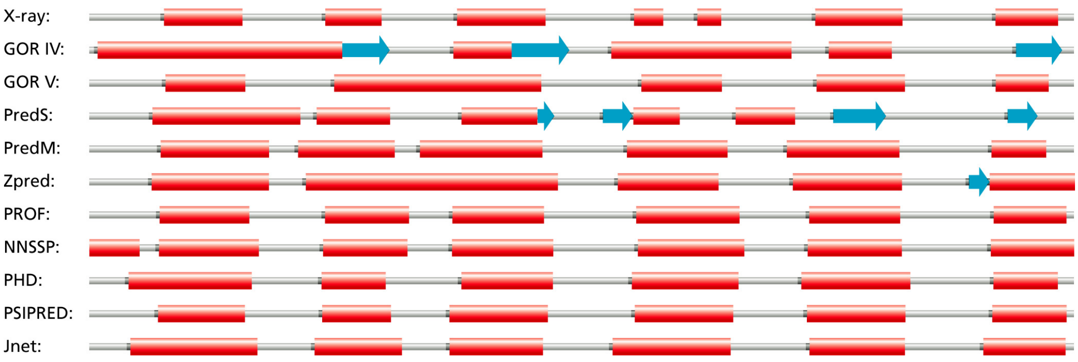
2nd network
60 input units
60 hidden layers
3 outputs

H, E, L

```
Conf:
Pred:
Pred: CCCCCCCCCCCCHHHHHHHHHHCCCEEEECCCCCCCCHHHHHHHHHHCCCCEEEEECCCEEEEEEECCCCCCCCEECCCH
  AA: HTLPANEFRCLTPEDAAGVFEIEREAFISVSGNCPLNLDEVQHFLTLCPELSLGWFVEGRLVAFIIGSLWDEERLTQESL
              10        20        30        40        50        60        70        80
```

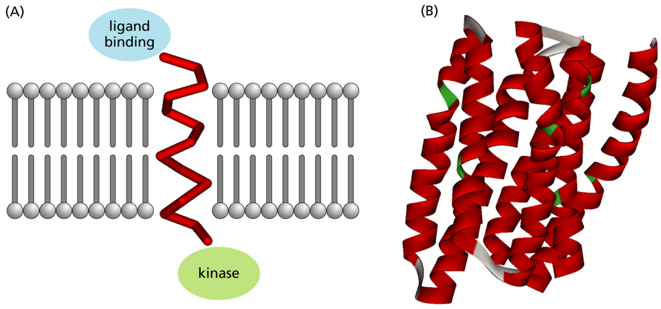
Secondary structure prediction methods

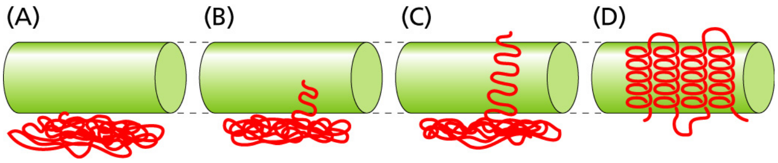




Transmembrane proteins

Membrane proteins are functionally important. For example, the receptors are formed by 1 or more helices spanning the mebrane

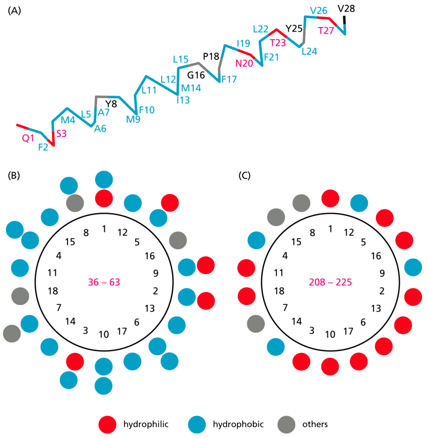




The four main ways in which proteins may be attached to a membrane. A) Attachment by interactions between the protein and the cytosolic face of the lipid bilayer. B) Attachment via an anchor (lipidic or terminals of the protein) that are added post-translationally. C) Transmembrane proteins have part of the protein chain embadded in the lipid bilayer. D) Transmembrane proteins where the protein chain threads back and forth across the mebrane multiple times.

Transmembrane proteins

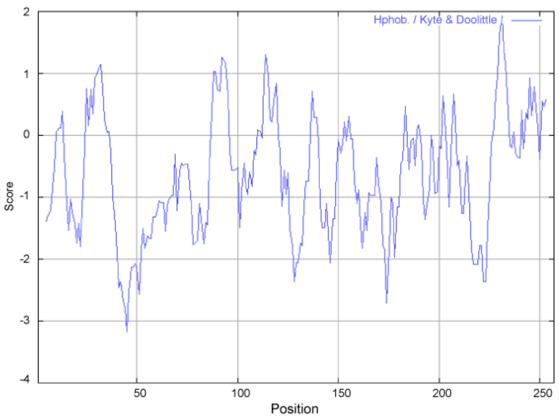
Helix wheel



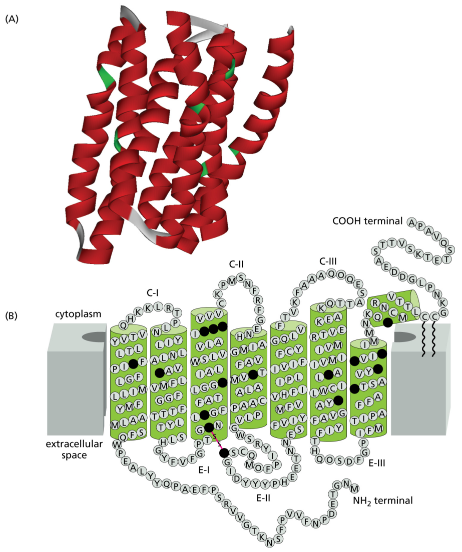
Hydrophobicity diagram

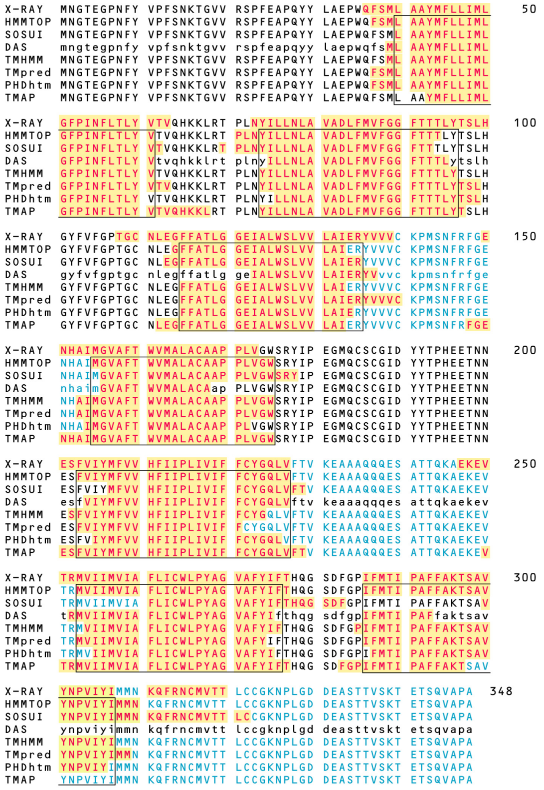
Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:

```
Ala:  1.800  Arg: -4.500  Asn: -3.500  Asp: -3.500  Cys:  2.500  Gln: -3.500
Glu: -3.500  Gly: -0.400  His: -3.200  Ile:  4.500  Leu:  3.800  Lys: -3.900
Met:  1.900  Phe:  2.800  Pro: -1.600  Ser: -0.800  Thr: -0.700  Trp: -0.900
Tyr: -1.300  Val:  4.200  : -3.500  : -3.500  : -0.490
```



Transmembrane proteins





```
X-RAY   MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML    50
HMMTOP  MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML
SOSUI   MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML
DAS     mngtegpnfy vpfsnktgvv rspfeapqyy laepwqfsML AAYMFLLIML
TMHMM   MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML
TMpred  MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML
PHDhtm  MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML
TMAP    MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML

X-RAY   GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH   100
HMMTOP  GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
SOSUI   GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
DAS     GFPINFLTLY Vtvqhkklrt plnyILLNLA VADLFMVFGG FTTTLytslh
TMHMM   GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
TMpred  GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
PHDhtm  GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH
TMAP    GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH

X-RAY   GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE   150
HMMTOP  GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
SOSUI   GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
DAS     gyfvfgptgc nlegffatlg geIALWSLVV LAIERYVvvc kpmsnfrfge
TMHMM   GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
TMpred  GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
PHDhtm  GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
TMAP    GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE

X-RAY   NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN   200
HMMTOP  NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN
SOSUI   NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN
DAS     nhaimGVAFT WVMALACAap PLVGWSRYIP EGMQCSCGID YYTPHEETNN
TMHMM   NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN
TMpred  NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN
PHDhtm  NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN
TMAP    NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN

X-RAY   ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV   250
HMMTOP  ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
SOSUI   ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
DAS     esfVIYMFVV HFIIPLIVIF FCYGQLVftv keaaaqqqes attqkaekev
TMHMM   ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
TMpred  ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
PHDhtm  ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
TMAP    ESFVIYMFVV HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV

X-RAY   TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV   300
HMMTOP  TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV
SOSUI   TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV
DAS     tRMVIIMVIA FLICWLPYAG VAFYIfthqg sdfgpIFMTI PAFfaktsav
TMHMM   TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV
TMpred  TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV
PHDhtm  TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV
TMAP    TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKTSAV

X-RAY   YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA   348
HMMTOP  YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
SOSUI   YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
DAS     ynpviyimmn kqfrncmvtt lccgknplgd deasttvskt etsqvapa
TMHMM   YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
TMpred  YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
PHDhtm  YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
TMAP    YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
```

Protein Sequence Motifs or Patterns

What is required is a method of searching for the occurrence of short sequence patterns, or motifs.

A motif, in general, is any conserved element of a sequence alignment (CONSENSUS), whether composed of a short sequence of contiguous residues or a more distributed pattern. Functionally related sequences will share similar distribution patterns of critical functional residues that are not necessarily contiguous.

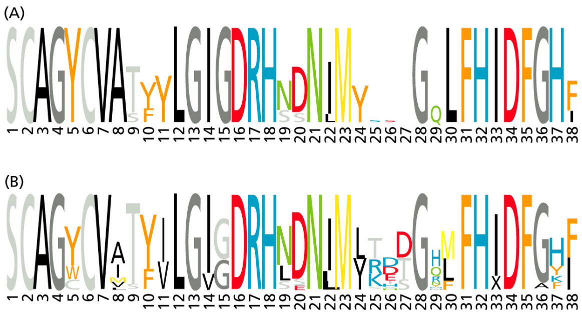
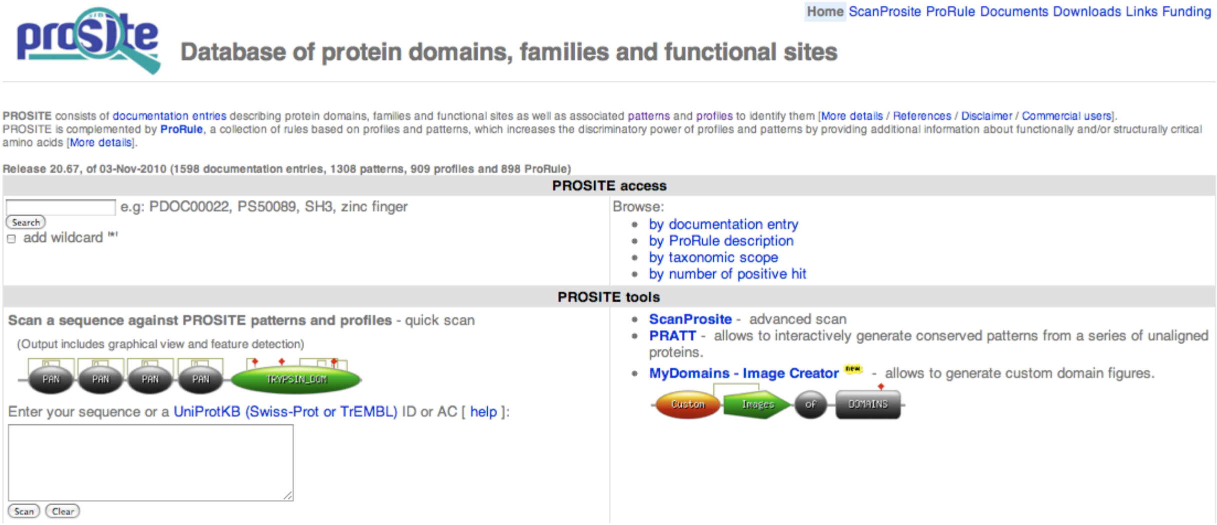


Figure 4.15
Residues that contribute to one of the blocks returned by the BLOCKS database after submission of the PI3-kinase p100α sequence.
(A) A block for four homologous sequences, and (B) for 31 homologous sequences. These representations are called logos, and are computed using a position-specific scoring matrix. This block contains the active-site amino acids and the DFG kinase motif. The size of the letters indicates the level of conservation and the colors indicate physicochemical properties of the residues: acidic, red; basic, blue; small and polar, white; asparagine and glutamine, green; sulfur-containing amino acids, yellow; hydrophobic, black; proline, purple; glycine, gray; aromatic, orange.

Protein Sequence Motifs or Patterns

The PROSITE database is a compilation of motifs and patterns extracted from protein sequences and compiled by inspection of protein families. This database can be searched with an unknown protein sequence to obtain a list of hits to possible patterns or protein signatures.



Database of protein domains, families and functional sites

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].

Release 20.67, of 03-Nov-2010 (1598 documentation entries, 1308 patterns, 909 profiles and 898 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger
Search
□ add wildcard '*'

Browse:
- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

PROSITE tools

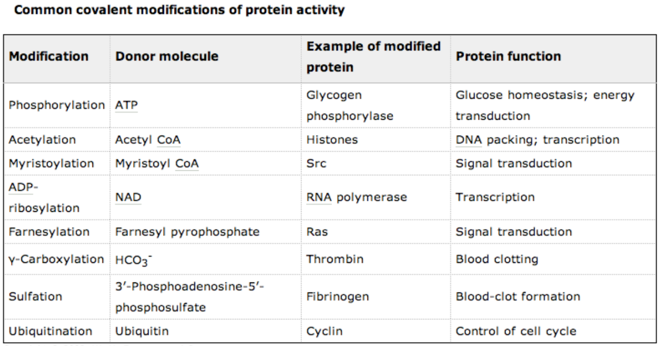
Scan a sequence against PROSITE patterns and profiles - quick scan
(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [help]:

Scan Clear

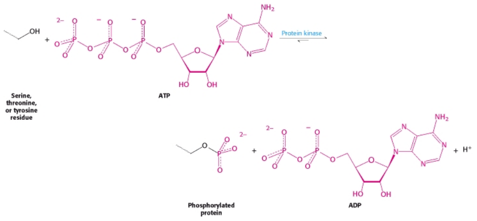
- ScanProsite - advanced scan
- PRATT - allows to interactively generate conserved patterns from a series of unaligned proteins.
- MyDomains - Image Creator new - allows to generate custom domain figures.

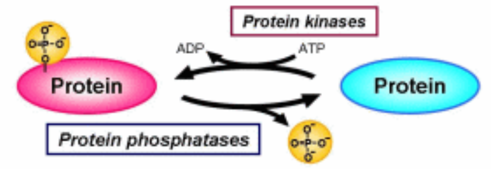
Protein Sequence Motifs or Patterns



Common covalent modifications of protein activity

Modification	Donor molecule	Example of modified protein	Protein function
Phosphorylation	ATP	Glycogen phosphorylase	Glucose homeostasis; energy transduction
Acetylation	Acetyl CoA	Histones	DNA packing; transcription
Myristoylation	Myristoyl CoA	Src	Signal transduction
ADP-ribosylation	NAD	RNA polymerase	Transcription
Farnesylation	Farnesyl pyrophosphate	Ras	Signal transduction
γ-Carboxylation	HCO_3^-	Thrombin	Blood clotting
Sulfation	3′-Phosphoadenosine-5′-phosphosulfate	Fibrinogen	Blood-clot formation
Ubiquitination	Ubiquitin	Cyclin	Control of cell cycle

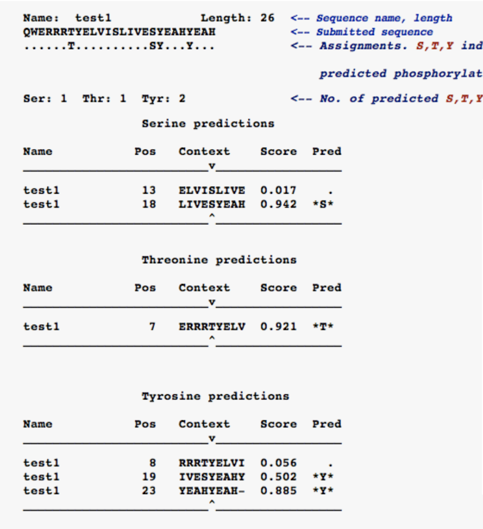




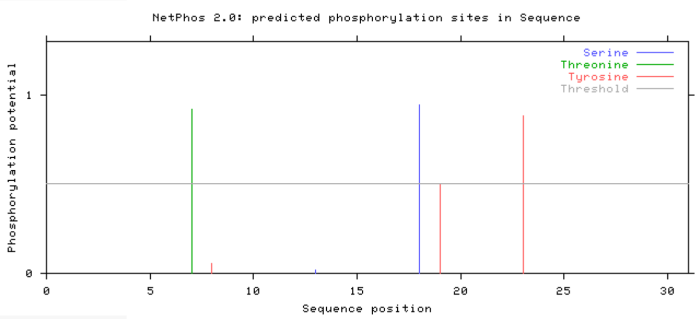
The consensus sequence recognized by protein kinase A is Arg-Arg-X-Ser-Z or Arg-Arg-X-Thr-Z, in which X is a small residue, Z is a large hydrophobic one, and Ser or Thr is the site of phosphorylation. It should be noted that this sequence is not absolutely required.

Protein Sequence Motifs or Patterns

NetPhos predicts phosphorylation sites in a protein sequence due to kinase acting post-translationally.



```
Name:  test1            Length: 26  <-- Sequence name, length
QWERRRTYELVISLIVESYEAHYEAH          <-- Submitted sequence
......T..........SY...Y...          <-- Assignments. S,T,Y indicates

                                        predicted phosphorylation sites

Ser: 1  Thr: 1  Tyr: 2               <-- No. of predicted S,T,Y phosph. sites

                Serine predictions

Name            Pos   Context    Score  Pred
___________________________v_________________

test1            13   ELVISLIVE  0.017    .
test1            18   LIVESYEAH  0.942  *S*
___________________________^_________________


                Threonine predictions

Name            Pos   Context    Score  Pred
___________________________v_________________

test1             7   ERRRTYELV  0.921  *T*
___________________________^_________________



                Tyrosine predictions

Name            Pos   Context    Score  Pred
___________________________v_________________

test1             8   RRRTYELVI  0.056    .
test1            19   IVESYEAHY  0.502  *Y*
test1            23   YEAHYEAH-  0.885  *Y*
___________________________^_________________
```


NetPhos 2.0: predicted phosphorylation sites in Sequence
Serine
Threonine
Tyrosine
Threshold
Phosphorylation potential
Sequence position