Practical course 2

1. Search for protein domains and their structure

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain	2 – 113	WH1 PROSITE-ProRule annotation	Add BLAST		112

Sequence feature

Entry & position(s)	P50552[2 - 113]
Description	
Feature key	Domain
Feature identifier	

```
        10         20         30         40         50
MSETVICSSR ATVMLYDDGN KRWLPAGTGP QAFSRVQIYH NPTANSFRVV
        60         70         80         90        100
GRKMQPDQQV VINCAIVRGV KYNQATPNFH QWRDARQVWG LNFGSKEDAA
       110        120        130        140        150
QFAAGMASAL EALEGGGPPP PPALPTWSVP NGPSPEEVEQ QKRQQPGPSE
```

Detailed signature matches

IPR000697 WH1/EVH1 domain

- SM00461 (WH1)
- PF00568 (WH1)
- PS50229 (WH1)

WH1 domain Edit Wikipedia article

Contents [hide]

1 Function
2 Interactions
3 Examples
4 References
5 External links

WH1 domain	
Identifiers	
Symbol	WH1
Pfam	PF00568
InterPro	IPR000697
SMART	WH1
SCOP	1evh
SUPERFAMILY	1evh
Available protein structures:	[show]

Function

WH1 domain is an evolutionary conserved protein domain.[1] Therefore, it has an important function. WH1 domains are found on WASP proteins, which are often involved in actin polymerization. Hence, WH1 is important for all cellular processes involving actin, this includes cell motility, cell trafficking, cell division and cytokinesis, cell signalling, and the establishment and maintenance of cell junctions and cell shape.[2]

Interactions

The WASP protein family control actin polymerization by activating the Arp2/3 complex. WASP is defective in Wiskott-Aldrich syndrome (WAS) whereby in most patient cases, the majority of point mutations occur within the N-terminal WH1 domain. The metabotropic glutamate receptors mGluR1alpha and mGluR5 bind a protein called homer, which is a WH1 domain homologue.[3][4]

A subset of WH1 domains has been termed the EVH1 domain and appear to bind a polyproline motif. The EVH1 (WH1, RanBP1-WASP) domain is found in multi-domain proteins implicated in a diverse range of signalling, nuclear transport and cytoskeletal events. This domain of around 115 amino acids is present in species ranging from yeast to mammals. Many EVH1-containing proteins associate with actin-based structures and play a role in cytoskeletal organisation. EVH1 domains recognise and bind the proline-rich motif FPPPP with low-affinity, further interactions then form between flanking residues.[4][5]

WASP family proteins contain an EVH1 (WH1) in their N-terminals which bind proline-rich sequences in the WASP interacting protein. Proteins of the RanBP1 family contain a WH1 domain in their N-terminal region, which seems to bind a different sequence motif present in the C-terminal part of RanGTP protein.[6][7]

Tertiary structure of the WH1 domain of the Mena protein revealed structure similarities with the pleckstrin homology (PH) domain. The overall fold consists of a compact parallel beta-sandwich, closed along one edge by a long alpha-helix. A highly conserved cluster of three surface-exposed aromatic side-chains forms the recognition site for the molecules target ligands.[8]

the PDB codes of the structures of the EVH1 (WH1) domain : **1EGX**

2. Classification database

A] fichier dans dossier

B]

Domain Annotation: CATH

CATH Database (version 4.0.0) Homepage

Chains	Domain	Class	Architecture	Topology	Homology
A	1rgwA00	Mainly Beta	Roll	Pdz3 Domain	

C]

PDZ domain

Edit Wikipedia article

The **PDZ domain** is a common structural domain of 80-90 amino-acids found in the signaling proteins of bacteria, yeast, plants, viruses[1] and animals.[2] Proteins containing PDZ domains play a key role in anchoring receptor proteins in the membrane to cytoskeletal components. Proteins with these domains help hold together and organize signaling complexes at cellular membranes. These domains play a key role in the formation and function of signal transduction complexes.[3] PDZ domains also play a highly significant role in the anchoring of cell surface receptors (such as Cftr[disambiguation needed] and FZD7) to the actin cytoskeleton via mediators like NHERF and ezrin.[4]

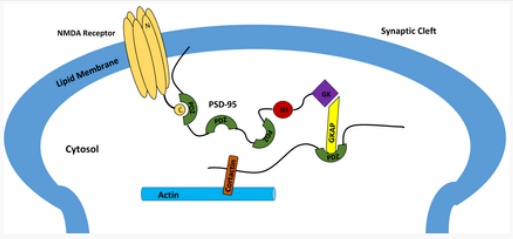
PDZ is an initialism combining the first letters of the first three proteins discovered to share the domain — post synaptic density protein (PSD95), Drosophila disc large tumor suppressor (Dlg1), and zonula occludens-1 protein (zo-1).[5] PDZ domains have previously been referred to as DHR (Dlg homologous region)[6] or GLGF (glycine-leucine-glycine-phenylalanine) domains.[7]

In general PDZ domains bind to a short region of the C-terminus of other specific proteins. These short regions bind to the PDZ domain by beta sheet augmentation. This means that the beta sheet in the PDZ domain is extended by the addition of a further beta strand from the tail of the binding partner protein.[8]. The C-terminal carboxylate group is bound by a nest (protein structural motif) in the PDZ domain.

PDZ proteins

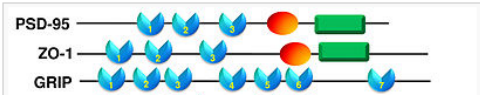
PDZ proteins are a family of proteins that contain the PDZ domain. This sequence of amino-acids is found in many thousands of known proteins. PDZ domain proteins are widespread in eukaryotes and eubacteria,[2] whereas there are very few examples of the protein in archaea. PDZ domains are often associated with other protein domains and these combinations allow them to carry out their specific functions. Three of the most well documented PDZ proteins are PSD-95, GRIP, and HOMER.

PSD-95 is a brain synaptic protein with three PDZ domains, each with unique properties and structures that allow PSD-95 to function in many ways. In general, the first two PDZ domains interact with receptors and the third interacts with cytoskeleton-related proteins. The main receptors associated with PSD-95 are NMDA receptors. The first two PDZ domains of PSD-95 bind to the C-terminus of NMDA receptors and anchor them in the membrane at the point of neurotransmitter release.[29] The first two PDZ domains can also interact in a similar fashion with Shaker-type K+ channels.[29] A PDZ interaction between PSD-95, nNOS and syntrophin is mediated by the second PDZ domain. The third and final PDZ domain links to cysteine-rich PDZ-binding protein (CRIPT), which allows PSD-95 to associate with the cytoskeleton.[29]



Basic functioning of PSD-95 in forming a complex between NMDA Receptor and Actin.

Glutamate receptor interacting protein (GRIP) is a post-synaptic protein with that interacts with AMPA receptors in a fashion analogous to PSD-95 interactions with NMDA receptors. When researchers noticed apparent structural homology between the C-termini of AMPA receptors and NMDA receptors, they attempted to determine if a similar PDZ interaction was occurring.[30] A yeast two-hybrid system helped them discover that out of GRIP's seven PDZ domains, two (domains four and five) were essential for binding of GRIP to the AMPA subunit called GluR2.[15] This interaction is vital for proper localization of AMPA receptors, which play a large part in memory storage. Other researchers discovered that domains six and seven of GRIP are responsible for connecting GRIP to a family of receptor tyrosine kinases called ephrin receptors, which are important signaling proteins.[31] A clinical study concluded that Fraser syndrome, an autosomal recessive syndrome that can cause severe deformations, can be caused by a simple mutation in GRIP.[32]



3. Structure superimposition

A]

Results for job clustalo-I20181024-081640-0163-38140832-p2m

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Show Colors | View result with Jalview | Send to Simple Phylogeny | Send to MView

```
CLUSTAL O(1.2.4) multiple sequence alignment


1RGW:A|PDBID|CHAIN|SEQUENCE      ------MAYSVTLTGPGFWGFRLQGGKDFNM---PLTISRITPGSKAA-QSQLSQGDLVV   50
3PDZ:A|PDBID|CHAIN|SEQUENCE      PKPGDIFEVEL-AKNDNSLGISVTGGVNTSVRHGGIYVKAVIPQGAAESDGRIHKGDRVL   59
1Z86:A|PDBID|CHAIN|SEQUENCE      -----RRRVTVRKADAGGLGISIKGGRENKM---PILISKIFKGLAADQTEALFVGDAIL   52
                                           :   . .   *: : ** : .:    : :. :     *     :  ** ::

1RGW:A|PDBID|CHAIN|SEQUENCE      AIDGVNTDTMTHLEAQNKIKSASYNLSLTLQKSKR--  85
3PDZ:A|PDBID|CHAIN|SEQUENCE      AVNGVSLEGATHKQAVETLRNTGQVVHLLLEKGQSPT  96
1Z86:A|PDBID|CHAIN|SEQUENCE      SVNGEDLSSATHDEAVQALKKTGKEVVLEVKYMKE--  87
                                 ::: * . .   ** :* : ::.:.  : * ::  :
```

11 aa are conserved. No Highly conservation.

B] Global alignement

```
# bin/ggsearch36 -E 10.0 -f -12 -g -2 24915.1.seq 24915.2.seq
GGSEARCH performs a global/global database searches
 version 36.3.5e Nov, 2012(preload8)
Query: 24915.1.seq
  1>>>unknown 96 bp - 96 aa
Library: 24915.2.seq
       87 residues in     1 sequences

Statistics: (shuffled [500]) Unscaled normal statistics: mu= -35.6820  var=195.0590 Ztrim: 0
 statistics sampled from 1 (1) to 500 sequences
Algorithm: Global/Global affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2
 Scan time:  0.000

The best scores are:                                      n-w bits E(1)
unknown 87 bp                                     (  87)   97 31.4   1e-21

>>unknown 87 bp                                            (87 aa)
 n-w opt:  97  Z-score: 145.0  bits: 31.4 E(1): 1e-21
global/global (N-W) score: 97; 30.9% identity (59.8% similar) in 97 aa overlap (1-96:1-87)

                10         20        30        40        50
unknow PKPGDIFEVELAKND-NSLGISVTGGVNTSVRHGGIYVKAVIPQGAAESDGRIHKGDRVL
        .     .: . : : ..:::: :: ....    : .. ..   ::..   .  :: .:
unknow RR-----RVTVRKADAGGLGISIKGGRENKM---PILISKIFKGLAADQTEALFVGDAIL
                   10        20          30        40        50

      60        70        80        90
unknow AVNGVSLEGATHKQAVETLRNTGQVVHLLLEKGQSPT
       .::: .: .::: .::..:..::. :  .::
unknow SVNGEDLSSATHDEAVQALKKTGKEV--VLEVKYMKE
             60        70         80



96 residues in 1 query   sequences
87 residues in 1 library sequences
 Tcomplib [36.3.5e Nov, 2012(preload8)] (4 proc in memory [0G])
 start: Wed Oct 24 08:10:00 2018 done: Wed Oct 24 08:10:00 2018
 Total Scan time:  0.000 Total Display time:  0.000

Function used was GGSEARCH [36.3.5e Nov, 2012(preload8)]
```

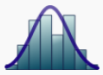
##	Scoring			RMSD	N_{align}	N_g	$\%_{seq}$	Query	Target structure				
	Q	P	Z					$\%_{sse}$	Match	$\%_{sse}$	N_{res}	×	Title
1	0.62	6.8	7.6	1.83	84	3	32	100	1z86.pdb:A	86	87	☐	SOLUTION STRUCTURE OF THE PDZ DOMAIN OF ALPHA-SYNTROPHIN

Root-mean-square deviation

From Wikipedia, the free encyclopedia

For the bioinformatics concept, see Root-mean-square deviation of atomic positions.

Statistics



Outline · Statisticians · Glossary · Notation · Journals · Lists of topics · Articles · Portal · Category

V · T · E

The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** (or sometimes **root-mean-squared error**) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called *residuals* when the calculations are performed over the data sample that was used for estimation and are called *errors* (or prediction errors) when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.[1]

RMSD is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSD is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used.

RMSD is the square root of the average of squared errors. The effect of each error on RMSD is proportional to the size of the squared error; thus larger errors have a disproportionately large effect on RMSD. Consequently, RMSD is sensitive to outliers.[2][3]

Standard score

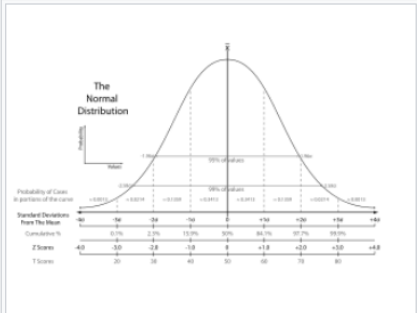
From Wikipedia, the free encyclopedia

> "Standardize" redirects here. For industrial and technical standards, see Standardization.
> For Fisher z-transformation in statistics, see Fisher transformation. For Z-values in ecology, see Z-value. For z-transformation to complex number domain, see Z-transform. For Z-factor in high-throughput screening, see Z-factor. For Z-score financial analysis tool, see Altman Z-score.

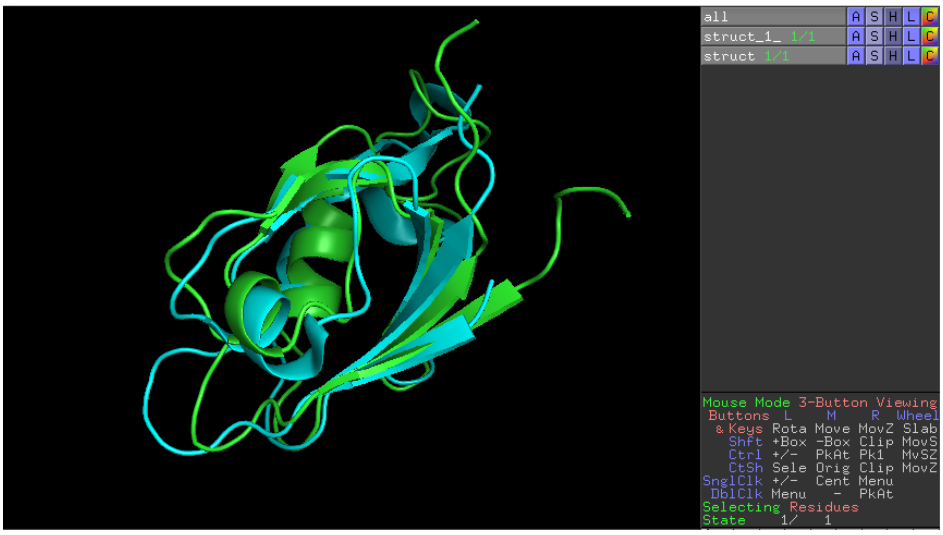
In statistics, the **standard score** is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. Observed values above the mean have positive standard scores, while values below the mean have negative standard scores. The standard score is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called **standardizing** or **normalizing** (however, "normalizing" can refer to many types of ratios; see normalization for more).

Standard scores are also called **z-values, *z*-scores, normal scores,** and **standardized variables**. They are most frequently used to compare an observation to a standard normal deviate, though they can be defined without assumptions of normality.

Computing a z-score requires knowing the mean and standard deviation of the complete population to which a data point belongs; if one only has a sample of observations from the population, then the analogous computation with sample mean and sample standard deviation yields the *t*-statistic.



Compares the various grading methods in a normal distribution. Includes: Standard deviations, cumulative percentages, percentile equivalents, Z-scores, T-scores

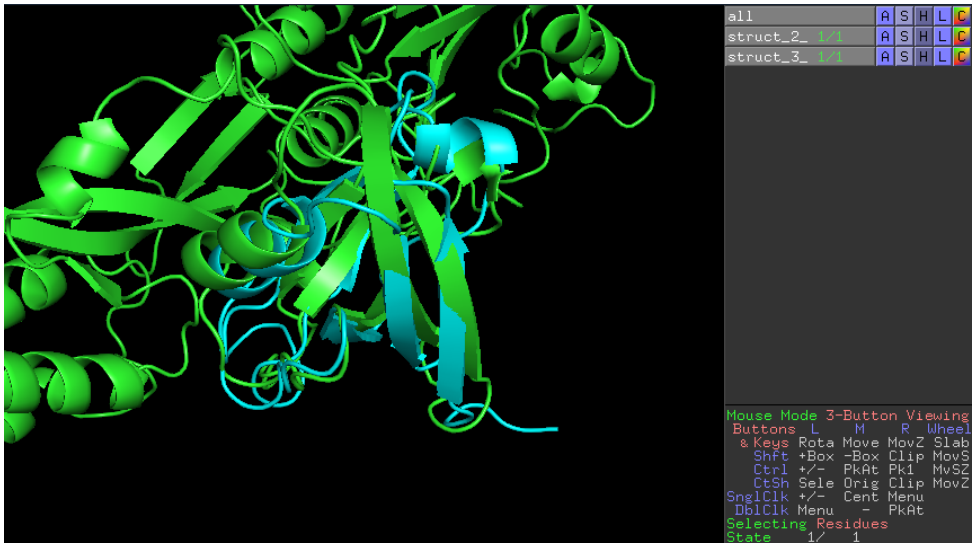


D]

No global alignement found.

```
Query: 13177.1.seq
  1>>>unknown 388 bp - 388 aa
Library: 13177.2.seq
       96 residues in     1 sequences

Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1716;  K=0.03016
 statistics sampled from 1 (1) to 500 sequences
Threshold: E() < 10 score: 28
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2010)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2
 Scan time:  0.000

>>unknown 96 bp                                          (96 aa)
 Waterman-Eggert score: 68;  21.9 bits; E(1) <  0.0096
26.8% identity (61.0% similar) in 82 aa overlap (83-162:21-96)

               90       100       110       120       130       140
unknow SVTG-VGLEITYDGGSGKDVVVLTPAPGGPAEKAG-ARAGDVIVTVDGTAVKGLSLYDVS
       :::: :.  . . :   : :.     : : ::. :  . :: ...:.....: .  ..
unknow SVTGGVNTSVRHGGIYVKAVI-----PQGAAESDGRIHKGDRVLAVNGVSLEGATHKQAV
                30        40            50        60        70

               150       160
unknow DLLQGEADSQVEVVLHAPGAPS
       . :.. ..  :....:   .:.
unknow ETLRNTGQV-VHLLLEKGQSPT
           80         90

>--
 Waterman-Eggert score: 44;  15.9 bits; E(1) <  0.45
29.2% identity (54.2% similar) in 48 aa overlap (106-149:1-48)

          110       120           130       140
unknow PAPGGPAEKAGARAGDVI-VTVDG---TAVKGLSLYDVSDLLQGEADS
       : ::   :   :.  . . ..: :   :.:.  ..:  . . :: :.:
unknow PKPGDIFEVELAKNDNSLGISVTGGVNTSVRHGGIYVKAVIPQGAAES
                10        20        30        40

>--
 Waterman-Eggert score: 42;  15.4 bits; E(1) <  0.57
22.5% identity (62.5% similar) in 40 aa overlap (229-265:23-62)

      230       240           250       260
unknow VAGLVLDIRNNG---GGLFPAGVNVARMLVDRGDLVLIAD
       ...:  ..:..:     ...: :.  .    . .:: :: ..
unknow TGGVNTSVRHGGIYVKAVIPQGAAESDGRIHKGDRVLAVN
                30        40        50        60


388 residues in 1 query   sequences
96 residues in 1 library sequences
 Scomplib [36.3.5e Nov, 2012(preload8)]
 start: Wed Oct 24 08:22:26 2018 done: Wed Oct 24 08:22:26 2018
 Total Scan time:  0.000 Total Display time:  0.000

Function used was LALIGN [36.3.5e Nov, 2012(preload8)]
```

##	Scoring			RMSD	N_{align}	N_g	%seq	Query	Target structure				
	Q	P	Z					%sse	Match	%sse	N_{res}	×	Title
1	0.084	0.0	6.1	2.36	71	3	23	14	3pdz.pdb:A	67	96	☐	SOLUTION STRUCTURE OF THE PDZ2 DOMAIN FROM HUMAN PHOSPHATASE HPTP1E



On a comparé des séquences peu conservées mais qui ont des structures similaires. Évidemment, pas d'alignement global possible puisqu'une des 2 protéines n'est alignée structurellement qu'a une partie de l'autre protéine.