

# Identification and analysis of a new family of bacterial serine proteinases

Shashi B. Pandit and N. Srinivasan\*

Molecular Biophysics Unit  
Indian Institute of Science  
Bangalore 560 012, India

\* Corresponding author

Phone: +91-80-2293 2837; Fax: +91-80-2360 0535; Email:  
[ns@mbu.iisc.ernet.in](mailto:ns@mbu.iisc.ernet.in)

Edited by E. Wingender; received July 30, 2004; revised and accepted October 22, 2004; published October 25, 2004

---

## Abstract

A family of hypothetical proteins, identified predominantly from archaeal genomes, has been analyzed in order to understand its functional characteristics. Using extensive sequence similarity searches it is inferred that this family is remotely related (best sequence identity is 19%) to ClpP proteinases that belongs to serine proteinase class. This family of hypothetical proteins is referred to as SDH proteinase family based on conserved sequential order of Ser, Asp and His residues and predicted serine proteinase activity. Results of fold recognition of SDH family sequences confirmed the remote relationship between SDH proteinases and Clp proteinases and revealed similar tertiary location of putative catalytic triad residues critical for serine proteinase function. However, the best sequence alignment we could obtain suggests that while catalytic Ser is conserved across Clp and SDH proteinases the location of the other catalytic triad residues, namely, His and Asp are swapped in their amino acid alignment

positions and hence in 3-D structure. The evidence of conserved catalytic triad suggests that SDH could be a new family of serine proteinases with the fold of Clp proteinase, however sharing the catalytic triad order of carboxypeptidase clan. Signal peptide sequence identified at the N-terminus of some of the homologues suggests that these might be secretory serine proteinases involved in cleavage of extracellular proteins while the remote homologues, ClpP proteinases, are known to work in intracellular environment.

*Key words:* ClpP proteinase, genome analysis, hypothetical proteins, serine proteinase, structure prediction

---

## Introduction

The completely sequenced genomes of a number of organisms provide an opportunity to understand the molecular basis of their physiology, metabolism, regulation and evolution [e. g., [Aravind et al., 2003](#)]. This is mainly inferred from the functional characterization of the gene products encoded in the genome. Clues about the functions of the proteins encoded in genomes can be obtained by their similarity to experimentally well characterized proteins whose sequences are available in the sequence databases. However, there are a large number of proteins encoded in genomes that do not exhibit obvious sequence similarity to proteins of known function. Moreover, most of these hypothetical proteins have not been experimentally explored for possible functions. Clues about the functions of these hypothetical proteins can be obtained by exploring their relationships to proteins of known function by means of sophisticated sequence analysis and fold recognition methods.

We have investigated such a collection of homologous hypothetical proteins of unknown function, from prokaryotes, by associating them to a protein of known function. This set of homologous proteins constitutes a new family with constituent members primarily from archaea such as *Pyrococcus furiosus* and *Methanocaldococcus jannaschii*. A few are from eubacteria such as *Corynebacterium glutamicum*. Most of the members of this new family are predicted to have an N-terminus signal peptide sequence. In most of the public sequence databases, the members of this family are annotated as "hypothetical proteins" except for one of the members, gi19553024 from *C. glutamicum*, which has an automated annotation as "probable periplasmic serine proteinase" in the Non-redundant database (NRDB) at NCBI. In sequence domain databases this family has no predicted function. For instance, in Pfam database [[Bateman et al., 2002](#)] this family is classified as "Domain of unknown function" (DUF114 family). However, in Cluster of Orthologous Group (COG) database [[Tatusov et al., 2001](#)] it has been classified under COG0616, which has predicted function as periplasmic serine proteinases. But, there is no evidence of experimental or detailed sequence analysis for the key residues involved in the serine proteinase functionality or attempt to classify this family among serine proteinase clans.

Using sensitive profile-based sequence similarity search tools such as PSI-BLAST [[Altschul \*et al.\*, 1997](#)] and threading of the SDH sequences on known 3-D folds we have explored the relatedness of SDH family to serine proteinase members. We also performed detailed investigation and predict key residues required for serine proteinase function. The present analysis showed that the identified family is remotely related to ClpP proteinases. However, the catalytic triad residues showed interesting variation with positions of His and Asp residues swapped in the amino acid sequence compared to ClpP proteinase members.

---

## Methods

### Databases

The sequences of members belonging to various families were retrieved from Pfam and non-redundant sequence databases (NRDB). The NRDB has been obtained from National Center for Biotechnology Information (<ftp://ftp.ncbi.nlm.nih.gov/blast/db>).

### Sequence analysis

The gapped BLAST and PSI-BLAST [[Altschul \*et al.\*, 1997](#)] was used to search against NRDB to retrieve the homologues. The IMPALA [[Schaffer \*et al.\*, 1999](#)] searchable profiles for PALI [[Balaji \*et al.\*, 2001](#); [Gowri \*et al.\*, 2003](#)] families have been generated as described in methods in [Pandit \*et al.\*, 2002](#). The sensitive profile matching method IMPALA [[Schaffer \*et al.\*, 1999](#)] was used to match each sequence against PALI profiles with e-value cutoff of  $10^{-5}$ . Multiple sequence alignment was done using T-Coffee available at [[Notredame \*et al.\*, 2000](#)] (<http://www.ch.embnet.org/software/TCoffee.html>). The dendrogram was generated using PHYLIP [[Felsenstein, 1993](#)] package by considering alignment of sequences obtained from ClustalW [[Thompson \*et al.\*, 1994](#)].

---

## Results and discussion

### Recognition of conserved residues

Sequences of the members of the new family were retrieved by searching of non-redundant sequence database (NRDB) at NCBI, using gapped BLAST [[Altschul \*et al.\*, 1997](#)] program with one of the members of this new family as a query. Both e-value and h-value cut-off used was  $10^{-4}$ . The homologues identified, with sequence identity (greater than 30%), in the first round are considered as the

member of the family. The homologous members of this new family were aligned using the multiple sequence alignment tool T-Coffee. A representative set of sequences of the new family is shown aligned in [Figure 1](#). As evident from the multiple sequence alignment of the family members, there is conservation of Ser, Asp and His, in specific positions of the alignment, apart from other conserved hydrophobic residues. It is important to note that Ser, Asp and His in specific locations are alone conserved. Ser, Asp and His in other alignment positions are not completely conserved. This observation raises the possibility of serine proteinase function for these proteins as also proposed by the automated annotation of one of the members of the family in a public database.



**Figure 1:** Multiple sequence alignment of representative members of SDH family with ClpP proteinase and Protease IV representative sequences. The alignment was generated using T-coffee. The final multiple sequence alignment was manually refined by considering high scoring sequence pairs given by PSI-BLAST searches and secondary structure predictions. In the alignment shown each block has first two sequences (CLPP\_ECOLI and CLP1\_MYXXA) of ClpP proteinase family, next six of SDH family and last two are Protease IV sequences (gi15964138 and gi17934547). The sequences are indicated either by SWISSPROT [[Boeckmann et al., 2003](#)] accession number or by Gene identification (GI) number from GenBank [[Benson et al., 2003](#)], followed by abbreviated species name. The start and end residues in every sequence are also indicated. The observed secondary structures from the crystal structure of ClpP protease (PDB: 1tyf) [[Wang et al., 1997](#)] are indicated above each block of the alignment. The orange rectangular blocks and arrows represent the observed  $\alpha$ -helices and  $\beta$ -stands, respectively. The consensus secondary structures predicted for SDH members are denoted in green color below each alignment block. One of the catalytic triad residues Ser that is conserved across the family is shown in bold red type against yellow background. The other catalytic residue Asp and His conserved in ClpP proteinase is shown in bold blue letters and His and Asp conserved in SDH family is shown in green bold type against yellow background showing the swapped nature of these residues. The residues that are conserved in all the

sequences are shown in pink bold letters. The other residues that are substituted by similar residues are shown in brown bold type. The species abbreviations are: M.xan (*Myxococcus xanthus*), E.col (*Escherichia coli*), M.maz (*Methanosarcina mazei* Goe1), T.mar (*Thermotoga maritima*), M.jan (*Methanocaldococcus jannaschii*), P.fur (*Pyrococcus furiosus* DSM 3638), C.aur (*Chloroflexus aurantiacus*), M.bar (*Methanosarcina barkeri*), S.mel (*Sinorhizobium meliloti*), A.tum (*Agrobacterium tumefaciens* str. C58). The other members of SDH family that are not included in the alignment are: gi14601916 (*Aeropyrum pernix*); gi14520474 (*Pyrococcus abyssi*); gi11498462 (*Archaeoglobus fulgidus*); gi14590206 (*Pyrococcus horikoshii*); gi17231317 (*Nostoc* sp. PCC 7120); gi23127726 (*Nostoc punctiforme*); gi15606180 (*Aquifex aeolicus*); gi20091240 (*Methanosarcina acetivorans* str. C2A); gi19553024 (*Corynebacterium glutamicum*), gi15669689 (*Methanococcus jannaschii*).

The serine proteinases are classified into six clans (families) based on their distinct three-dimensional (3-D) folds and biological roles, though a highly similar spatial orientation of catalytic triad residues is shared among them [[Rawlings and Barrett, 1994](#)]. These clans also have different order of catalytic triad residues Asp, His and Ser, in their amino acid sequences [[Rawlings and Barrett, 1994](#)]. The proposed new family has the conservation order of Ser, Asp and His and hence these homologues are collectively referred as "SDH proteinase family".

### **Similarities between SDH and ClpP proteinase families**

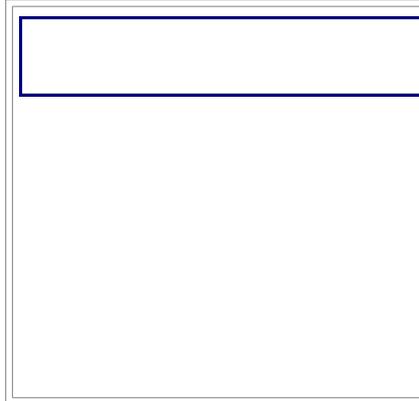
We have explored the possibility of SDH proteinase family being distantly related to any of the other protein families, using profile-based search procedures and 3-D fold recognition methods. When we used SDH family members as queries in PSI-BLAST searches against NRDB, in the second round of iterations we identified homologues of Protease IV (E-value:  $5 \times 10^{-7}$  to  $1 \times 10^{-4}$ ) family and a few viral members of ClpP proteinase (E-values:  $2 \times 10^{-14}$  to  $1 \times 10^{-6}$ ) as distant homologues of the SDH family. The sequence identities of SDH proteinases with Protease IV and ClpP proteinases are of the order of 17% and 19% respectively. Interestingly, the region of alignment involved conserved Ser of SDH family members and a conserved Ser residue in Protease IV. The Protease IV is an endopeptidase that utilizes this conserved serine for its catalytic activity and is involved in secretion of proteins [[Ichihara et al., 1986](#); [Palmer and St John, 1987](#)]. It has no conserved Asp and His residues. In the subsequent iterations of PSI-BLAST, ClpP proteinase members were detected with E-values ranging from

$2 \times 10^{-5}$  to  $3 \times 10^{-46}$ . Using IMPALA [Schaffer *et al.*, 1999], searches have been made with SDH sequences as queries against structure-based sequence profile database of protein domain families of known structure available in PALI database [Balaji *et al.*, 2001; Gowri *et al.*, 2003]. SDH members could be related to ClpP proteinase with E-values from  $3 \times 10^{-21}$  to  $4 \times 10^{-5}$ . We have included a mention of this potential relationship between SDH proteinase family (referred to as "Domain of unknown function", DUF114 in Pfam) and the superfamily of ClpP proteinases in our in-house SUPFAM database [Pandit *et al.*, 2002], which relates sequence-based (Pfam) and structure-based (PALI) families. Interestingly, ClpP proteinases possess a serine proteinase activity [Maurizi *et al.*, 1990a; 1990b] by utilizing the characteristic conserved catalytic triad [Wang *et al.*, 1997] and is known to work intracellularly [Maurizi *et al.*, 1990a]. Representative sequences of Protease IV and ClpP proteinases are also shown aligned to SDH proteinases in [Figure 1](#).

### **Two of the catalytic triad residues are swapped in the amino acid sequences of SDH proteinases compared to ClpP proteinases**

The SDH family and ClpP proteinase family members have the best pairwise sequence identity of 19%. However, fold recognition for the sequences of the SDH family showed that the SDH sequences could be comfortably accommodated into the fold of ClpP proteinase that is ClpP/crotonase fold. This was done by using fold recognition methods, GenThreader [Jones, 1999] (<http://bioinf.cs.ucl.ac.uk/psiform.html>), 3D-PSSM [Kelley *et al.*, 2000] (<http://www.sbg.bio.ic.ac.uk/~3dpssm>) and FUGUE [Shi *et al.*, 2001] (<http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html>), with sequences of SDH family as queries. All the above mentioned methods reliably associated the SDH proteinase with the 3-D fold of ClpP proteinase. The E-values for the various queries from SDH family all resulting in the top most hit of ClpP proteinases in GenThreader and 3D-PSSM are less than 0.01 and 0.005 respectively. The Z-scores in FUGUE are greater than 6. The best sequence alignment we could obtain between SDH and ClpP proteinase members as shown in [Figure 1](#) suggests that, while catalytic Ser is conserved across SDH and ClpP proteinases the location of other catalytic residues, His and Asp, are swapped in their amino acid alignment positions in SDH. The fold of ClpP proteinase, which is predicted for SDH family, is shown in [Figure 2](#) with SDH sequence threaded on the ClpP proteinase fold. Threading of the sequence of a SDH proteinase on the 3-D fold of ClpP proteinase shows that catalytic Asp and His of SDH proteinases are topologically equivalent respectively to the catalytic His and Asp of ClpP proteinases ([Figure 2](#)). Thus the spatial geometric relationship among the catalytic triad residues is preserved in SDH proteinases and hence it is suggested that SDH could be a new family of serine proteinases with adopting a ClpP proteinase fold, but sharing the catalytic triad order of carboxypeptidase clan. The Protease IV members show only one of the catalytic residues (Ser) conserved. The prediction of the secondary structure of SDH proteinases by employing the PHD protocol [Rost and Sander, 1993; Rost, 1996] (<http://www.cubic.bioc.columbia.edu/predictprotein>) shows predicted  $\alpha$ -helical and  $\beta$ -strand regions similar to those observed in the corresponding regions of

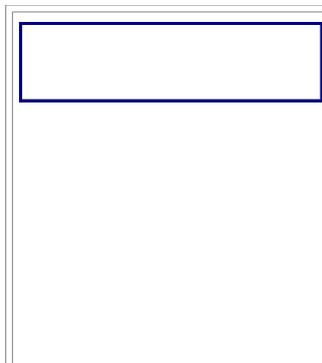
the crystal structure of ClpP proteinases ([Figure 1](#)).



**Figure 2:** An overlay of SDH proteinase sequence over ClpP proteinase fold, showing that two of the conserved catalytic triad residues of ClpP and SDH are swapped in the primary structures, but, proximal in 3-D space. The side chain in red color is that of conserved Ser. The ClpP His and Asp side chains are shown in blue, and Asp and His of SDH family are highlighted in green. This figure has been produced using Setor software [[Evans, 1993](#)].

### Evolutionary relationships among SDH, Protease IV and ClpP Proteinase families

The dendrogram ([Figure 3](#)) is generated by considering representative sequences from various related Pfam family members that belongs to the same superfamily as ClpP protease in order to explore evolutionary relationship of SDH family of proteins. According to SCOP [[Murzin et al., 1995](#)] Peptidase S41B and Enoyl-CoA hydratase/isomerase corresponds, respectively, to the families of Tail specific protease and Crotonase-like. Protease IV members are classified in Peptidase S49 Pfam family. The dendrogram shows distinct clusters of various families with ClpP protease, Peptidase S49 or Protease IV and SDH proteinase present in three distinct family clusters. The sequence similarities between the members of SDH family, ClpP proteinase family and Protease IV family are very low (less than 20%) suggesting that these clusters correspond to distinct protein families although they may be evolutionarily related. These three distinct families have a conserved catalytic Ser residue. However, ClpP and SDH proteinase families have two other catalytic triad residues Asp and His residues conserved although they are swapped in their amino acid sequential positions between two proteinase families. In case of other families they have only "Ser" residue conserved for their catalytic function.



**Figure 3:** Dendrogram showing distinct cluster of members belonging to SDH family with respect to other related Pfam family members that belongs to same superfamily. The Pfam family name is indicated for of each cluster. Pfam families Peptidase S41B and Enoyl-CoA hydratase/isomerase corresponds to Tail specific protease and Crotonase-like family in SCOP [[Murzin et al., 1995](#)] respectively. Protease IV members are present in Peptidase S49 Pfam family. The representative sets of

proteins for each family are taken from Pfam database. The dendrogram is generated using ClustalW [Thompson *et al.*, 1994] alignment. Each family members share high sequence identity among themselves. However, the sequence identity across family members is poor. The alignment generated using such sets of protein with low sequence identity is not good enough to provide a reliable dendrogram to infer the evolution of SDH family. However, such dendrogram provides clusters of proteins sharing similar functional features. The members used in the alignment are: gi15605158 (*Chlamydia trachomatis*); gi14194529 (*Chlamydia muridarum*); gi29839992 (*Chlamydophila caviae*); gi15618431 (*Chlamydophila pneumoniae* CWL029); gi23004911 (*Magnetospirillum magnetotacticum*); gi15615680 (*Bacillus halodurans*); gi7387634 (*Sus scrofa*); gi33859811 (*Mus musculus*); gi18677763 (*Rattus norvegicus*); gi23102708 (*Azotobacter vinelandii*); gi32043639, gi15595648 (*Pseudomonas aeruginosa*); gi26986944 (*Pseudomonas putida* KT2440); gi33601768 (*Bordetella bronchiseptica*); gi29654774 (*Coxiella burnetii* RSA 493); gi24112837 (*Shigella flexneri* 2a str. 301); gi33861736 (*Prochlorococcus marinus*); gi28898926 (*Vibrio parahaemolyticus*); gi6723247 (*Wolbachia* sp.); gi23000247 (*Magnetococcus* sp.); gi24216571 (*Leptospira interrogans*); gi15838452 (*Xylella fastidiosa*); gi15645963 (*Helicobacter pylori* 26695); gi15643510 (*Thermotoga maritima*) gi20127408, gi543064 and gi595267 are from *Homo sapiens*; gi16331391 and gi16329323 are from *Synechocystis* sp. PCC 6803. The GI codes with name of organism of SDH family members are provided in [Figure 1](#).

### **Geometry of catalytic triad residues in serine proteinases, variations and flexibility**

The spatial proximity of Asp, His and Ser in the catalytic triad of serine proteinases is characterized by specific geometry and interactions among the catalytic triad residues. Structural parameters characterizing the geometry of serine proteinases, flexibility and their influence in understanding of the mechanism of action have been analysed extensively by various groups [e. g., [Matthews \*et al.\*, 1977](#); [Kossiakoff and Spencer, 1981](#); [Huber and Bennett, 1983](#); [Blow, 1991](#); [Iengar and Ramakrishnan, 1999](#)]. The inter-residue distances of catalytic triad residues are well conserved within the members of a family of serine proteinases. For example, in the homologous members of the family of

trypsin like serine proteinases the  $C_{\alpha}$ - $C_{\alpha}$  distances between Asp-His, Asp-Ser and His-Ser are observed to be close to 6.4Å, 10.2Å and 8.3Å respectively [[Iengar and Ramakrishnan, 1999](#)]. On the other hand, geometry of the catalytic triad residues across different families of serine proteinases shows some marked variation. For example, inter  $C_{\alpha}$  distances of Asp-His, Asp-Ser and His-Ser in the structure of Clp proteinase are 4.8Å, 9Å and 6Å respectively. Differences of the order of 2Å can be noticed compared to the corresponding distances in homologous members of the trypsin family. Despite these differences relative positioning of the functional groups in the sidechains of Asp, His and Ser are reasonably well preserved. For example, both in trypsin and Clp proteinases His in the triad acts as hydrogen bonding donor (with Asp) and acceptor (with Ser).

An analysis of the potential structural features of SDH family members have been made in order to investigate the possibility of proper geometry and hydrogen bonding at the putative catalytic triad. As mentioned before, in the structure of Clp proteinase, inter- $C_{\alpha}$  Asp-Ser and Asp-His distances are longest and shortest respectively. For the reasons mentioned later in this paper it appears that the backbone conformations of the regions containing Asp and His in the SDH proteinases are not expected to be very similar to the aligned regions containing His and Asp, respectively, of Clp proteinases (see [Figure 1](#)). However if backbone conformation of SDH proteinase at the putative catalytic triad region is imagined to be identical to that of Clp proteinases then His-Ser distance could turn out to be largest in SDH proteinases as position of Asp in Clp proteinases corresponds to His in SDH proteinases ([Figure 1](#)) and Asp-Ser distance is the largest in Clp proteinases.

However a careful analysis of the sequence alignment in the catalytic triad region ([Figure 1](#)) shows that the loop containing His in SDH proteinases has more number of residues compared to the corresponding loop containing Asp of Clp proteinases. This strongly suggests that the conformation of this loop in SDH proteinases would be different compared to the corresponding (shorter) loop in Clp proteinases. Due to more number of residues this loop in SDH proteinases can be expected to be bulkier possibly resulting in closer positioning of His residue to Ser compared to positioning of Asp in Clp proteinases.

It can also be noted in [Figure 1](#) that next to Asp in the putative catalytic triad of SDH proteinases there is deletion of a residue in the alignment. This could possibly result in less protruding conformation of the loop in SDH proteinases compared to the corresponding loop in Clp proteinases. Thus Asp in SDH proteinases might be located at a farther distance from Ser compared to His-Ser distance of Clp proteinases.

Thus, it appears that Asp-Ser distance in SDH proteinase is longer than His-Ser distance of Clp proteinases and His-Ser distance in SDH proteinases is shorter than Asp-Ser distance of Clp proteinases. Thus the geometry of the catalytic triad of SDH proteinases is likely to be consistent with the requirements of the serine proteinase activity with relative disposition of the three catalytic triad residues in SDH proteinases closely resembling that of the Clp proteinases although there is a swapping between Asp and His positions in the sequence alignment. It can also be noted from [Figure 1](#) that regions closely flanking catalytic Ser residue have

no insertions or deletions and hence position of Ser in SDH and Clp proteinases are likely to be similar.

The crystal structure of Clp proteinase shows that there are hydrogen bonds between catalytic His and Asp as well as between His and Ser. As mentioned already His and Asp positions in SDH proteinases are swapped compared to Clp proteinases. However, for the reasons discussed already, due to predicted small, but, critical alterations in the catalytic loop conformations involving Asp and His of SDH proteinases it is suggested that hydrogen bonds between His and Asp as well as between His and Ser are likely to be present in SDH proteinases as well.

---

## Conclusions

The in-depth analysis showed that SDH family is remotely related to ClpP proteinase and has catalytic triad residues conserved required for serine proteinase function. Hence, we propose that SDH family is a probable serine proteinase and is involved in proteolysis in the prokaryotes. Some of these are predicted to have signal peptide at the N-terminus, suggesting their role as secretory serine proteinases while ClpP proteinases are known to be intracellular. The SDH family presents an interesting variation in the serine proteinase class wherein it shares fold of ClpP proteinase clan and has catalytic order of carboxypeptidase clan. We are unaware of such a variation in among the families of serine proteinases with two proteinase families having the same fold and different order of catalytic triad residues in their amino acid sequences.

It is well known that critical functional residues in enzymatic families are well conserved even in the cases of extensive divergent evolution. In the exceptional cases with functional residues not preserved in a homologue of an enzyme, either the homologue is usually an enzyme with completely different function [[Bartlett et al., 2003](#)] or, it is inactive as an enzyme [[Pils and Schultz, 2004](#)]. Retention of function with altered residue arrangement, such as the example of SDH proteinases and Clp proteinases described in this paper, critically depends on geometry and other conformational aspects of the catalytic residues. A grossly simple "triangular" geometry of catalytic triads in Clp and SDH proteinases appears to have enabled swapping of Asp and His positions in these distantly related families. The retention of the geometry of catalytic triad residues is likely to be achieved by the flexibility in the loop regions containing catalytic residues and these regions are characterized by insertions and deletions in the alignment. We are currently unaware of any other enzyme superfamily of this kind with a common fold and function, but, with altered sequential order of catalytic residues.

---

## Acknowledgements

S.B.P is supported by Council of Scientific and Industrial Research, New Delhi. This research is supported by the award of International Senior Fellowship to N.S. by the Wellcome Trust, London and by the computational genomics project funded by the Department of Biotechnology, New Delhi.

---

## References

---

- [Aravind, L., Iyer, L. M., Wellems, T. E. and Miller, L. H. \(2003\). Plasmodium Biology. Genomic Gleanings. Cell 115, 771-785.](#)
- [Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. \(1997\). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.](#)
- [Balaji, S., Sujatha, S., Kumar, S. S. C. and Srinivasan, N. \(2001\). PALI-a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res. 29, 61-65.](#)
- [Bartlett, G. J., Borkakoti, N. and Thornton, J. M. \(2003\). Catalysing new reactions during evolution: economy of residues and mechanism. J. Mol. Biol. 331, 829-860.](#)
- [Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. \(2002\). The Pfam protein families database. Nucleic Acids Res. 30, 276-280.](#)
- [Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. \(2003\). GenBank. Nucleic Acids Res. 31, 23-27.](#)
- [Blow, D. \(1991\). Enzymology. Lipases reach the surface. Nature 351, 444-445.](#)
- [Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A.,](#)

Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365-370.

---

- Evans, S. V. (1993). SETOR: hardware-lighted three-dimensional solid model representations of macromolecules. J. Mol. Graph. 11, 134-138.
- 

- Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.
- 

- Gowri, V. S., Pandit, S. B., Karthik, P. S., Srinivasan, N. and Balaji, S. (2003). Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res. 31, 486-488.
- 

- Huber, R. and Bennett, W. S. jr. (1983). Functional significance of flexibility in proteins. Biopolymers 22, 261-279.
- 

- Ichihara, S., Suzuki, T., Suzuki, M. and Mizushima, S. (1986). Molecular cloning and sequencing of the *sppA* gene and characterization of the encoded protease IV, a signal peptide peptidase, of *Escherichia coli*. J. Biol. Chem. 261, 9405-9411.
- 

- Iengar, P. and Ramakrishnan, C. (1999). Knowledge-based modeling of the serine protease triad into non-proteases. Protein Eng. 12, 649-656.
- 

- Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797-815.
- 

- Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. 299, 499-520.
- 

- Kossiakoff, A. A and Spencer, S. A. (1981). Direct determination of the protonation states of aspartic acid-102 and histidine-57 in the tetrahedral intermediate of the serine proteases: neutron structure of trypsin. Biochemistry 20, 6462-6474.
- 

- Matthews, D. A., Alden, R. A., Birktoft, J. J., Freer, T. and Kraut, J. (1977). Re-examination of the charge relay system in subtilisin comparison with other serine proteases. J. Biol. Chem. 252, 8875-8883.
-

- [Maurizi, M. R., Clark, W. P., Katayama, Y., Rudikoff, S., Pumphrey, J., Bowers, B. and Gottesman, S. \(1990a\). Sequence and structure of Clp P, the proteolytic component of the ATP-dependent Clp protease of \*Escherichia coli\*. J. Biol. Chem. 265, 12536-12545.](#)

---
- [Maurizi, M. R., Clark, W. P., Kim, S. H. and Gottesman, S. \(1990b\). Clp P represents a unique family of serine proteases. J. Biol. Chem. 265, 12546-12552.](#)

---
- [Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. \(1995\). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.](#)

---
- [Notredame, C., Higgins, D. G. and Heringa, J. \(2000\). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205-217.](#)

---
- [Palmer, S. M and St. John, A. C. \(1987\). Characterization of a membrane-associated serine proteinase in \*Escherichia coli\*. J. Bacteriol. 169, 1474-1479.](#)

---
- [Pandit, S. B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S. S., Mhatre, N. S., Sowdhamini, R. and Srinivasan, N. \(2002\). SUPFAM - a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res. 30, 289-293.](#)

---
- [Pils, B. and Schultz, J. \(2004\). Inactive enzyme-homologues find new function in regulatory processes. J. Mol. Biol. 340, 399-404.](#)

---
- [Rawlings, N. D. and Barrett, A. J. \(1994\). Families of serine peptidases. Methods Enzymol. 244, 19-61.](#)

---
- [Rost, B. and Sander, C. \(1993\). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599.](#)

---
- [Rost, B. \(1996\). PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol. 266, 525-539.](#)

---

- [Schäffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. and Altschul, S. F. \(1999\). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. \*Bioinformatics\* 15, 1000-1011.](#)

---
- [Shi, J., Blundell, T. L. and Mizuguchi, K. \(2001\). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. \*J. Mol. Biol.\* 310, 243-257.](#)

---
- [Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E. V. \(2001\). The COG database: new developments in phylogenetic classification of proteins from complete genomes. \*Nucleic Acids Res.\* 29, 22-28.](#)

---
- [Thompson, J. D., Higgins, D. G. and Gibson, T. J. \(1994\). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. \*Nucleic Acids Res.\* 22, 4673-4680.](#)

---
- [Wang, J., Hartling, J. A. and Flanagan, J. M. \(1997\). The structure of ClpP at 2.3 Å resolution suggests a model for ATP-dependent proteolysis. \*Cell\* 91, 447-456.](#)