Protein Science (1998), 7:2499–2510. Cambridge University Press. Printed in the USA.

A homology identification method that combines protein sequence and structure information

LIHUA YU, JAMES V. WHITE,[1] AND TEMPLE F. SMITH

BioMolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street,
Boston, Massachusetts 02215

(RECEIVED January 7, 1998; ACCEPTED June 23, 1998)

Abstract

A new method is presented for identifying distantly related homologous proteins that are unrecognizable by conventional sequence comparison methods. The method combines information about functionally conserved sequence patterns with information about structure context. This information is encoded in stochastic discrete state-space models (DSMs) that comprise a new family of hidden Markov models. The new models are called sequence-pattern-embedded DSMs (pDSMs). This method can identify distantly related protein family members with a high sensitivity and specificity. The method is illustrated with trypsin-like serine proteases and globins. The strategy for building pDSMs is presented. The method has been validated using carefully constructed positive and negative control sets. In addition to the ability to recognize remote homologs, pDSM sequence analysis predicts secondary structures with higher sensitivity, specificity, and Q3 accuracy than DSM analysis, which omits information about conserved sequence patterns. The identification of trypsin-like serine proteases in new genomes is discussed.

Keywords: globin; Hidden Markov Model (HMM); homology identification; serine protease; sequence comparison; sequence-pattern-embedded Discrete State-Space Models (pDSMs); structure prediction

The sequencing of the whole genomes of different organisms continues to supply an enormous amount of new protein sequence information. The inferring of homology to the proteins with known structure and/or function is the most common method of annotating these new sequences. Sequence comparison (for review, see Hubbard, 1997) is the direct approach for inferring homology. However, homologous proteins having little sequence similarities are often statistically undetectable by conventional sequence comparison methods (Gibrat et al., 1996; Holm & Sander, 1996). Homology or common ancestry in such cases can be inferred from their common three-dimensional structure and function (Murzin et al., 1995; Michie et al., 1996).

A few examples (Holm & Sander, 1995, 1997) have shown that structural similarity, detected by comparing three-dimensional structures of proteins, can successfully find remote, but highly probable, homologs when combined with the identification of functionally conserved residues. Since direct homology identification based on the comparison of three-dimensional structures is limited to proteins with known structures, it has very limited application to recently determined sequences. Moreover, even when protein function information is not available from direct experimental work, the identification of a functionally conserved sequence pattern indicates the possible functional role of proteins (we use the term sequence pattern in a broad sense, it can be as simple as one amino acid in each conserved position or be as complicated as a complete profile), because the key residues in the active sites of enzymatic proteins are always functionally conserved (Bairoch, 1991) in a conserved three-dimensional structural context (Lesk & Chothia, 1980; Greer, 1990; Stocker & Bode, 1995).

We describe a new method of sequence analysis that combines structure-fold recognition with the identification of functionally conserved sequence patterns for identifying remote homologs. There have been several reports that incorporate structural information to varying degrees, either in the sequence pattern descriptions, such as ARIADNE (Lathrop et al., 1987) and Scrutineer (Sibbald & Argos, 1990), or to restrict the consensus sequence template in the structurally conserved regions of the sequence (Taylor, 1986). However, those methods rely primarily on sequence information. Our method uses new probabilistic structural models (discrete state-space models (DSMs)) (Stultz et al., 1993, 1997; White et al., 1994) that contain the functionally conserved sequence pattern elements. The functionally conserved sequence pattern is embedded in a DSM by replacing the amino acid probability distributions

Reprint requests to: Temple F. Smith, BioMolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street, Boston, Massachusetts 02215; e-mail: tsmith@darwin.bu.edu.

[1]Present address: TASC, Inc., 55 Walkers Brook Drive, Reading, Massachusetts 01867.

associated with particular secondary structural states, with the distributions of the conserved sequence pattern elements. We refer to these new DSMs as sequence-pattern-embedded DSMs (pDSMs).

In this paper, the pDSM method is validated on the recognition of homologs in two protein families. The first is the trypsin-like serine protease family (we use the names serine proteases and trypsin-like serine proteases interchangeably) (Greer, 1990), which is diverse in both structure and sequence. The second is the well-studied globin family (Lesk & Chothia, 1980), which also contains very divergent sequences. In addition to recognizing distant homologs, the pDSM method provides improved secondary structure prediction. We discuss the application of the method in genome research, as illustrated by the identification of new trypsin-like serine proteases in recently sequenced genomes.

Results

Trypsin-like serine proteases

Homology identification

The results of homology identification for serine proteases are listed in Table 1. The conserved sequence pattern we chose (see Materials and methods: pDSM for trypsin-like serine proteases) covers all of the sequences in the two positive sets (see Materials and methods: pDSM for trypsin-like serine proteases) from the PDB and Genbank. Therefore, the sensitivity of the conserved sequence pattern searches is 100% for both positive sets. The average sensitivity of BLAST searches, with a reasonably high cutoff score (10^{-6}), was 65% for PDB and 78% for Genbank. Searches using only the DSM's structure model recognize the serine protease fold with 84% sensitivity on PDB and 60% on Genbank. In contrast, the pDSM searches, which use both the conserved sequence pattern and the structure model, achieved a sensitivity of 100% on both positive sets.

Since all of the sequences in the negative control set have the minimal conserved pattern (see Materials and methods: pDSM for trypsin-like serine proteases), the specificity of the conserved sequence pattern searches is 0% by definition. In contrast, BLAST achieves an average specificity of 100%. The DSM searches for serine proteases have a specificity of 88%. By combining the structure information and the minimal conserved sequence pattern in our pDSMs, we obtain a specificity of 93%.

Table 1. *Sensitivity and specificity of serine protease homology identification by different methods*[a]

	Sensitivity		Specificity
Search method	PDB(32) (%)	Genbank(111) (%)	PDB(206) (%)
Conserved sequence Pattern	100	100	0
BLAST	65	78	100
DSM	84	60	88
pDSM	100	100	93

[a] Number of sequences in each dataset are shown in parentheses.

Half (7 out of 14) of the false positives (a false positive is a protein in the negative control set that is incorrectly classified as a serine protease) actually have all β-folds, and they have the His-Asp-Ser pattern in their primary sequences, but their secondary structural elements are packed differently than serine proteases. Consequently, the His, Asp, and Ser residues that are recognized are not close to each other in the three-dimensional structure. One of these false positives, 1HAVA (Allaire et al., 1994), is a cysteine proteinase, but has a similar fold to chymotrypsin-like serine proteases. Residues His44 and Asp84 of 1HAVA can be aligned with the two conserved His57 and Asp102 in three chymotrypsins. (We

Table 2. *The potential trypsin-like serine proteases in genomes identified by pDSM sequence analysis*

	***B. subtilis*: MPR_PBS**
Prediction	(1) Probability: 0.85 (2) Serine protease domain: 104–313 (3) Catalytic triad: His146, Asp191, Ser267
Comment	(1) Annotation[a]: extracellular metalloprotease (Rufo et al., 1996) (2) Weakly similar to 1TRY and 1ELT (PDB), similar to GSEP_BACLI[b] (SWISS-PROT) (3) Signature[c]: TRYPSIN_HIS (4) Alignment with known serine protease: Figure 1
	***E. coli*: b1598**
Prediction	(1) Probability: 0.86 (2) Serine protease domain: entire sequence (3) Catalytic triad: His84, Asp145, Ser223
Comment	(1) Annotation: 24% identical to MPR_BACSU (2) Weakly similar to MPR_PBS and GSEP_BACLI (SWISS-PROT) (3) Signatures: TRYPSIN_HIS and TRYPSIN_SER (4) Alignment with known serine protease: Figure 1
	***S. cerevisiae*: YNL123W**
Prediction	(1) Probability: 0.85 (2) Serine protease domain: 76–286 (3) Catalytic triad: His121, Asp152, Ser236
Comment	(1) Annotation: weak similarity to *C. jejuni* serine protease (2) Similar to HTRA_ECOLI (SWISS-PROT) (3) Signature: none (4) Alignment with known serine protease: Figure 2
	***C. elegans*: CEIV000158**
Prediction	(1) Probability: 0.95 (2) Serine protease domain: entire sequence (3) Catalytic triad: His69, Asp117, Ser212
Comment	(1) Annotation: similar to peptidase family S1 (trypsin) (2) Similar to 1PFX, etc.[d] (3) Signature: TRYPSIN_HIS and TRYPSIN_SER (4) Alignment with known serine proteases: Figure 3

[a] The annotations are obtained from the original genome databases.

[b] GSEP_BACLI has recently been identified as a remote homolog of trypsin-like serine proteases by sequence analysis (Alexandre et al., 1996; Pearson, 1997).

[c] The signatures of serine proteases are defined in PROSITE (Bairoch, 1991).

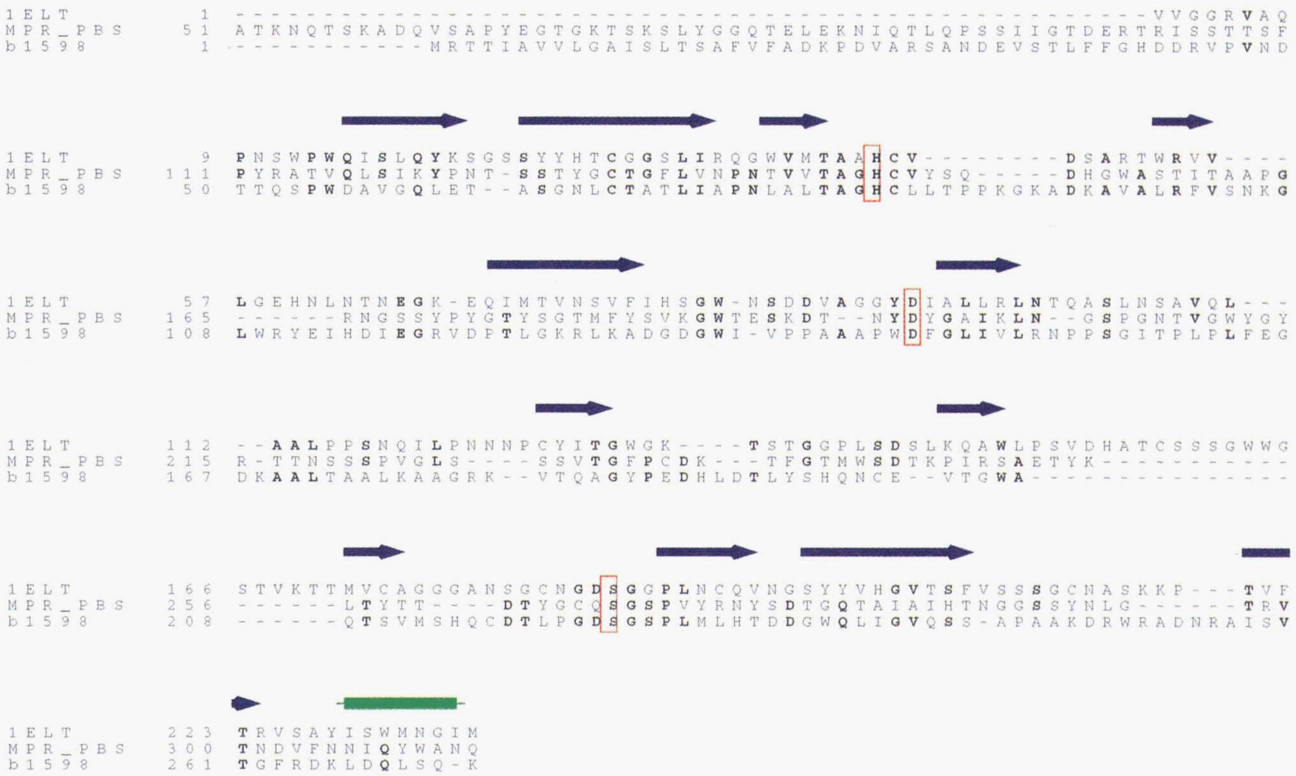
[d] CEIV000158 matches many serine proteases by BLAST search.

refer to the residue numbers according to the sequence alignment based on the chymotrypsinogen (Hartley & Kauffman, 1966). The catalytic triad is numbered as His57, Asp102, and Ser195.) The conserved Ser195 of chymotrypsins was replaced by Cys172 in 1HAVA, but in that neighborhood there is a Ser178. So not surprisingly, 1HAVA has been predicted as the serine protease by the pDSM search. This false positive demonstrates the ability of the pDSM search to find remote homologs. Although this cysteine proteinase is considered to be a remote homolog of trypsin-like serine proteases (Alexandre et al., 1996), we assigned it to the negative control set for the study, because the presence of the complete catalytic triad is required for all positives.

In summary, the pDSMs identify serine protease homologs with a high sensitivity comparable with that of the conserved sequence pattern or the DSM alone and the BLAST searches. In particular, the higher sensitivity of pDSM vs. DSM searches shows that the embedding of the sequence pattern helps to improve the fold recognition significantly. The BLAST search with a cutoff of 10^{-6} is the most specific method for identifying homologs, but it has low sensitivity. The conserved sequence pattern searches are the least specific. The structure information encoded in the pDSMs apparently accounts for the high specificity of the searches.

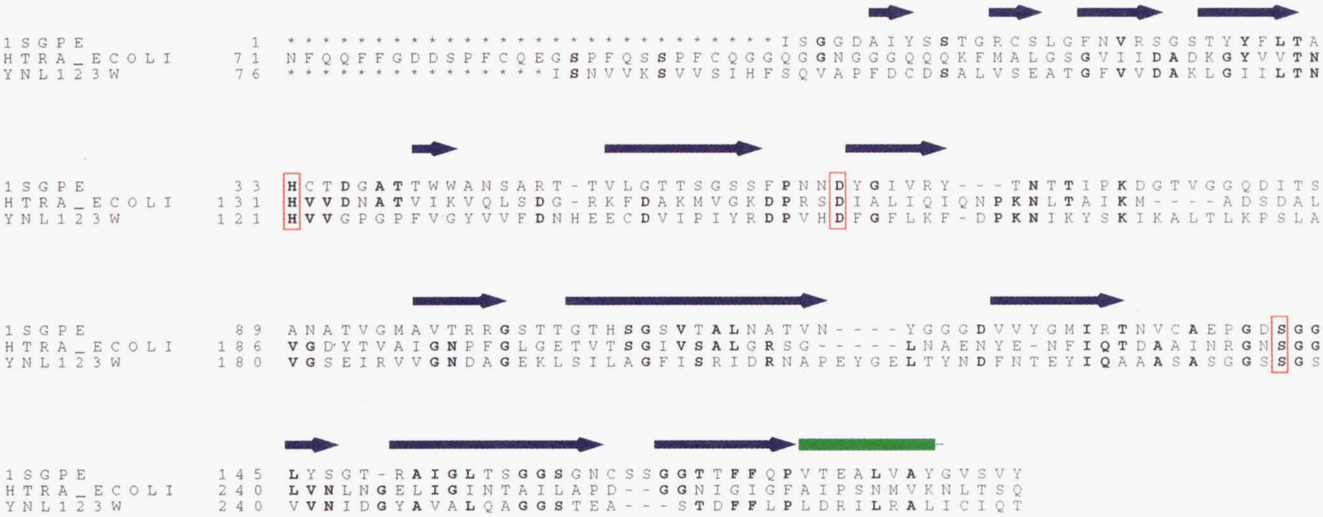
More sensitive sequence search methods have been published recently. A new iterative version of BLAST called PSI–BLAST (Altschul et al., 1997) has been shown to achieve higher sensitivity than BLAST by recruiting a position-specific score matrix. The PROBE program (Neuwald et al., 1997) exploits similar logic as PSI-BLAST but in addition generates multiple sequence alignment models. We compared our method with the publicly available PSI-BLAST. Since PSI-BLAST is available only over the internet (http://www.ncbi.nlm.nih.gov/cgi-gin/BLAST/nph-psi_blast), the method cannot be applied to our positive and negative control sets. We therefore summarize the PSI-BLAST sensitivity and specificity separately rather than in the same table. We submitted our PDB positive controls to the PSI-BLAST server using the default parameters ($E = 0.01$). The results show that by starting from a single query sequence in one of the two serine protease clusters, all the sequences within the same cluster are returned along with a varying number of sequences from the other cluster. If we calculate the sensitivity in the same way as for BLAST (Equation 1 in Materials and methods) and only count those returned sequences that are in our positive control sets, PSI-BLAST achieved the 85% sensitivity for serine proteases.

We get a similar result for the globins with a higher sensitivity of 92%. These numbers are significantly higher than those from the original BLAST search. In addition, PSI-BLAST has very high specificity under the default parameter setting. The only false positives were streptavidin for the serine protease case, and this protein does have a similar beta barrel fold. Only one false positive was found in the globin case, a parvalbumin, an all alpha protein like the globins. These results support the use of an iterative search starting with a single sequence and using each of its matched sequences to initiate additional searches until no new matches are found. It is a very efficient way to find remote homologs. The method does, however, require that the entire range of taxonomic variation is well represented in the database searched. This is best



```
1ELT       1  ----------------------------------------------------VVGGRVAQ
MPR_PBS   51  ATKNQTSKADQVSAPYEGTGKTSKSLYGGQTELEKNIQTLQPSSIIGTDERTRISSTTSF
b1598      1  -----------MRTTIAVVLGAISLTSAFVFADKPDVARSANDEVSTLFFGHDDRVPVND

1ELT       9  PNSWPWQISLQYKSGSSYYHTCGGSLIRQGWVMTAAHCV--------DSARTWRVV----
MPR_PBS  111  PYRATVQLSIKYPNT-SSTYGCTGFLVNPNTVVTAGHCVYSQ-----DHGWASTITAAPG
b1598     50  TTQSPWDAVGQLET--ASGNLCTATLIAPNLALTAGHCLLTPPKGKADKAVALRFVSNKG

1ELT      57  LGEHNLNTNEGK-EQIMTVNSVFIHSGW-NSDDVAGGYDIALLRLNTQASLNSAVQL---
MPR_PBS  165  ------RNGSSYPYGTYSGTMFYSVKGWTESKDT--NYDYGAIKLN--GSPGNTVGWYGY
b1598    108  LWRYEIHDIEGRVDPTLGKRLKADGDGWI-VPPAAAPWDFGLIVLRNPPSGITPLPLFEG

1ELT     112  --AALPPSNQILPNNNPCYITGWGK----TSTGGPLSDSLKQAWLPSVDHATCSSSGWWG
MPR_PBS  215  R-TTNSSSPVGLS----SSVTGFPCDK---TFGTMWSDTKPIRSAETYK-----------
b1598    167  DKAALTAALKAAGRK--VTQAGYPEDHLDTLYSHQNCE--VTGWA---------------

1ELT     166  STVKTTMVCAGGGANSGCNGDSGGPLNCQVNGSYYVHGVTSFVSSSGCNASKKP---TVF
MPR_PBS  256  ------LTYTT----DTYGCQSGSPVYRNYSDTGQTAIAIHTNGGSSYNLG------TRV
b1598    208  ------QTSVMSHQCDTLPGDSGSPLMLHTDDGWQLIGVQSS-APAAKDRWRADNRAISV

1ELT     223  TRVSAYISWMNGIM
MPR_PBS  300  TNDVFNNIQYWANQ
b1598    261  TGFRDKLDQLSQ-K
```

Fig. 1. The multiple sequence alignment of MPR_PBS from *B. subtilis* and b1598 from *E. coli* aligned with 1ELT in the PDB, a known trypsin-like serine protease with a solved crystal structure. The secondary structure of 1ELT is indicated above the sequences. β Strands are shown as arrows, and the ending helix is shown as a bar. The catalytic triad of His, Asp, and Ser is highlighted in boxes. The highly conserved residues are indicated in boldface.



```
1SGPE          1 ****************************ISGGDAIYSSTGRCSLGFNVRSGSTYYFLTA
HTRA_ECOLI    71 NFQQFFGDDSPFCQEGSPFQSSPFCQGGQGGNGGGQQQKFMALGSGVIIDADKGYVVTN
YNL123W       76 ***************ISNVVKSVVSIHFSQVAPFDCDSALVSEATGFVVDAKLGIILTN

1SGPE         33 HCTDGATTWWANSART-TVLGTTSGSSFPNNDYGIVRY---TNTTIPKDGTVGGQDITS
HTRA_ECOLI   131 HVVDNATVIKVQLSDG-RKFDAKMVGKDPRSDIALIQIQNPKNLTAIKM----ADSDAL
YNL123W      121 HVVGPGPFVGYVVFDNHEECDVIPIYRDPVHDFGFLKF-DPKNIKYSKIKALTLKPSLA

1SGPE         89 ANATVGMAVTRRGSTTGTHSGSVTALNATVN----YGGGDVVYGMIRTNVCAEPGDSGG
HTRA_ECOLI   186 VGDYTVAIGNPFGLGETVTSGIVSALGRSG-----LNAENYE-NFIQTDAAINRGNSGG
YNL123W      180 VGSEIRVVGNDAGEKLSILAGFISRIDRNAPEYGELTYNDFNTEYIQAAASASGGSSGS

1SGPE        145 LYSGT-RAIGLTSGGSGNCSSGGTTFFQPVTEALVAYGVSVY
HTRA_ECOLI   240 LVNLNGELIGINTAILAPD--GGNIGIGFAIPSNMVKNLTSQ
YNL123W      240 VVNIDGYAVALQAGGSTEA---STDFFLPLDRILRALICIQT
```

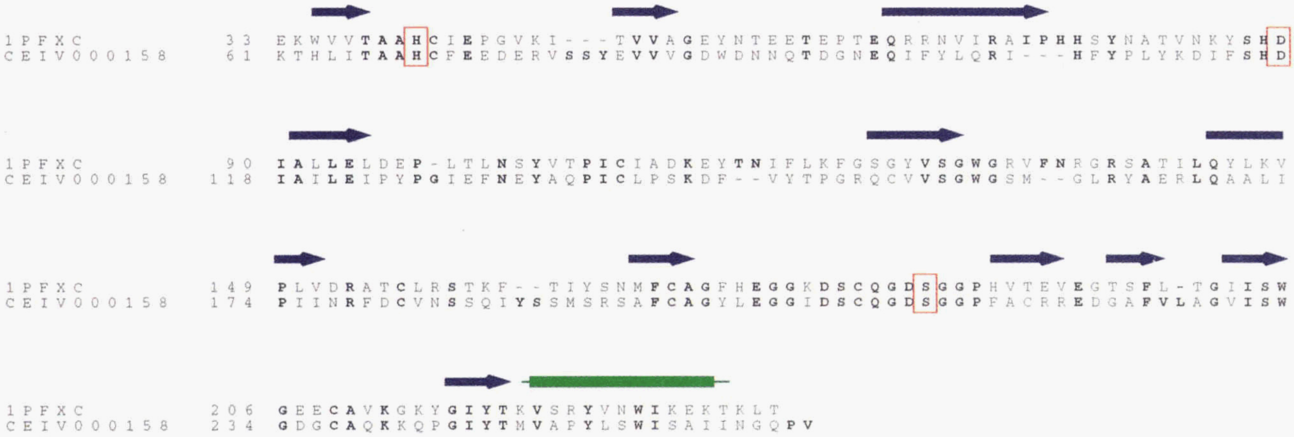
Fig. 2. The multiple sequence alignment of YNL123W from *S. cerevisiae* and HTRA_ECOLI from *E. coli* aligned with 1SGPE in the PDB, a known trypsin-family serine protease with a solved crystal structure. The secondary structure of 1SGPE is indicated above the sequences. β Strands are shown as arrows, and the ending helix is shown as a bar. The catalytic triad of His, Asp, and Ser is highlighted in boxes. The highly conserved residues are indicated in boldface.

illustrated in the globin case, for which the sensitivity is only 42% rather than the above noted 92% if each of our positive control sequences is used to initiate single BLAST search.

Searching the genome databases

To illustrate the application of the pDSM method in genome research, we searched five recently sequenced genomes for trypsin-like serine proteases: *Bacillus subtilis* (http://www.pasteur.fr/Bio/SubtiList.html) (Kunst et al., 1997), *Caenorhabditis elegans* (http://www.sanger.ac.uk/Projects/C_elegans/), *Escherichia coli* (http://www.genetics.wisc.edu/) (Blattner et al., 1997), *Methanococcus jannaschii* (http://www.tigr.org/tdb/mdb/mdb.html) (Bult et al., 1996), and *Saccharomyces cerevisiae* (http://speedy.mips.biochem.mpg.de/) (Mewes et al., 1997). The results are presented in Table 2. To our surprise, we found only one trypsin-like serine protease but several subtilisin-like serine proteases in all genomes except for *M. jannaschii*.

The four serine proteases were predicted by the pDSM method with high probabilities. Additional tests were made to support the pDSM prediction. These include: (1) BLAST search against PDB, SWISS-PROT, (2) consensus pattern search against PROSITE (Bairoch, 1991), and (3) sequence aligned with experimentally verified serine proteases (Figs. 1–3). The aligned positions of predicted catalytic triads with the experimentally identified triads are highlighted. Among the four examples, only CEIV000158 shows significant sequence similarity to known serine proteases. By a method of transitive search similar to that used by PSI-BLAST, YNL123W is related to 1SGPE through an htrA-family serine protease (Lip-



```
1PFXC         33 EKWVVTAAHCIEPGVKI---TVVAGEYNTEETEPTEQRRNVIRAIPHHSYNATVNKYSHD
CEIV000158    61 KTHLITAAHCFEEDERVSSYEVVVGDWDNNQTDGNEQIFYLQRI---HFYPLYKDIFSHD

1PFXC         90 IALLELDEP-LTLNSYVTPICIADKEYTNIFLKFGSGYVSGWGRVFNRGRSATILQYLKV
CEIV000158   118 IAILEIPYPGIEFNEYAQPICLPSKDF--VYTPGRQCVVSGWGSM--GLRYAERLQAALI

1PFXC        149 PLVDRATCLRSTKF--TIYSNMFCAGFHEGGKDSCQGDSGGPHVTEVEGTSFL-TGIISW
CEIV000158   174 PIINRFDCVNSSQIYSSMSRSAFCAGYLEGGIDSCQGDSGGPFACRREDGAFVLAGVISW

1PFXC        206 GEECAVKGKYGIYTKVSRYVNWIKEKTKLT
CEIV000158   234 GDGCAQKKQPGIYTMVAPYLSWISAIINGQPV
```

Fig. 3. The sequence alignment of CEIV000158 from *C. elegans* with 1PFXC in the PDB, a known trypsin-family serine protease with a solved crystal structure. The secondary structure of 1PFXC is indicated above the sequences. β Strands are shown as arrows, and the ending helix is shown as a bar. The catalytic triad of His, Asp, and Ser is highlighted in boxes. The highly conserved residues are indicated in boldface.

inska et al., 1989), HTRA_ECOLI from SWISS-PROT. MPR_PBS was originally classified experimentally as metalloprotease (Rufo et al., 1990).

Why does the MPR_PBS classification based on sequence analysis differ from its classification from biochemical experiments? The experiments (Rufo et al., 1990) involved PMSF, a serine protease inhibitor, but the inhibition experiment as described in the paper had no positive controls. Moreover, there are serine proteases, such as v8 protease, that are known to be resistant to PMSF inhibitor. Also, the evidence for the involvement of metal ion is not very convincing given the high EDTA concentration (25 mM). An alternative explanation of the involvement of metal ion was given in Alexandre et al. (1996) based on homologous modeling of the active site. In conclusion, we classify MPR_PBS as a trypsin-like serine protease.

Secondary structure prediction

In Table 3, we present fully cross-validated secondary structure predictions for all of the 32 serine proteases in the positive set from the PDB (see Materials and methods: Model validation). Both DSM and pDSM results are reported.

A residue is predicted to be in a strand when the probability of being in a strand for that residue is higher than 0.5. The probability is computed using the optimal smoothing algorithm (Stultz et al., 1993). The secondary structure prediction was compared with the secondary structure assignment by DSSP (the region assigned as E in DSSP and more than two residues long is considered to be a strand) (Kabsch & Sander, 1983) (Fig. 4). The agreement with DSSP is measured by sensitivity and specificity of strand prediction and the Q3 accuracy (the percentage of all residues that are correctly predicted to belong to one of the three types of secondary structures: helix, strand, and coil). As expected, the pDSM prediction reliability is higher than the DSM prediction reliability because of the additional sequence pattern information. The average sensitivity increases from 61 to 69%, average specificity increases from 83 to 86%, and average Q3 accuracy increases from 69 to 76%. This is not surprising given that improved fold recognition (Table 1) allows us to choose the proper model more often, as demonstrated by the significantly improved secondary structure prediction of 1ABIH, 1DST, 1PCU, 1PFXC, and 3RP2A. Moreover, the pDSMs provide more accurate secondary structure prediction over DSMs, even when they both represent the proper structures. The decreased standard deviations of all three quantities demonstrate that the secondary structure prediction by pDSM analysis is more stable.

Globins

Homology identification

The sensitivities and specificities of homology identification for globins by four different methods are summarized in Table 4. Like the sensitivity results for serine proteases in Table 1, the highest sensitivity of 100% is achieved by both conserved sequence pattern searches and the pDSM searches. In contrast, the average sensitivity of BLAST searches, at 10^{-6}, is only 42%. This is attributed to the fact that globins fall into six clusters according to sequence similarity (see Materials and methods: pDSM for globins). The low BLAST sensitivity reflects the low sequence similarity between any two of these clusters. In contrast, the DSM and pDSM searches are influenced much less by the low sequence similarity.

Table 3. *The Q3 accuracy, sensitivity, and specificity of strand prediction for positive serine proteases from the PDB by DSMs and pDSMs*[a]

	Q3 (%)		Sensitivity (%)		Specificity (%)	
Loci	DSM	pDSM	DSM	pDSM	DSM	pDSM
1ABIH	46	76	28	69	88	87
1ACBE	66	70	54	57	79	83
1ARB	63	67	41	59	87	82
1BIT	73	78	60	68	86	91
1BMAA	78	82	69	78	85	88
1BRA	77	79	60	71	88	93
1BTP	77	80	64	71	91	92
1DST	37	76	23	70	87	89
1ELT	83	83	76	82	91	90
1FUJA	78	76	62	70	89	83
1GBAA	65	63	43	42	86	86
1HCGA	75	77	72	75	85	88
1HF1	76	81	65	74	89	95
1HNEE	83	84	75	79	94	92
1HPGA	68	73	66	75	75	75
1HYLA	77	77	61	71	89	84
1LMWB	79	81	72	81	85	86
1PCU	39	79	35	68	91	93
1PFA	72	75	68	71	85	87
1PFXC	43	73	91	66	24	85
1PYTC	74	72	50	56	85	83
1PYTD	62	70	54	55	77	83
1RTFB	80	82	67	75	90	88
1SGC	69	71	73	77	71	68
1SGPE	67	68	68	72	72	68
1SGT	80	80	76	79	92	89
1TON	72	82	59	64	83	95
1TRY	76	82	68	78	91	93
2CP1	75	76	62	65	85	93
2KAI	80	82	59	68	92	94
2SFA	71	70	69	72	78	72
3RP2A	58	74	75	63	54	85
Average	69	76(64)	61	69(56)	83	86(81)
STD	12	5	15	9	13	7

[a] The DSSP secondary structure assignments are taken as the true secondary structure. The numbers in parentheses, listed in the row Average, were obtained using the GOR algorithm (Garnier et al., 1978) and are for comparison only. The numbers in row STD are the standard deviations of each column.

The sensitivity of DSM searches is 58%. In these searches, 88% (23 out of 26) of the positives were correctly predicted as α class proteins, but they were misclassified into nonglobin folds. The 100% sensitivity of the pDSM searches is achieved because the pDSM has the conserved sequence pattern embedded in its structural model.

The conserved sequence pattern searches have 0% specificity, as expected, while the average specificity of BLAST searches is 100%. The DSM specificity is 90%, and the pDSM specificity is 97%, both higher than the corresponding specificities for serine proteases in Table 1. The probable reason for this improvement is that globins have less structural variability than serine proteases. Only one DSM and one pDSM are needed to model the globin family, while the serine proteases are modeled by eight DSMs and eight pDSMs (two secondary structure topologies, each having four ranges of sequence length).

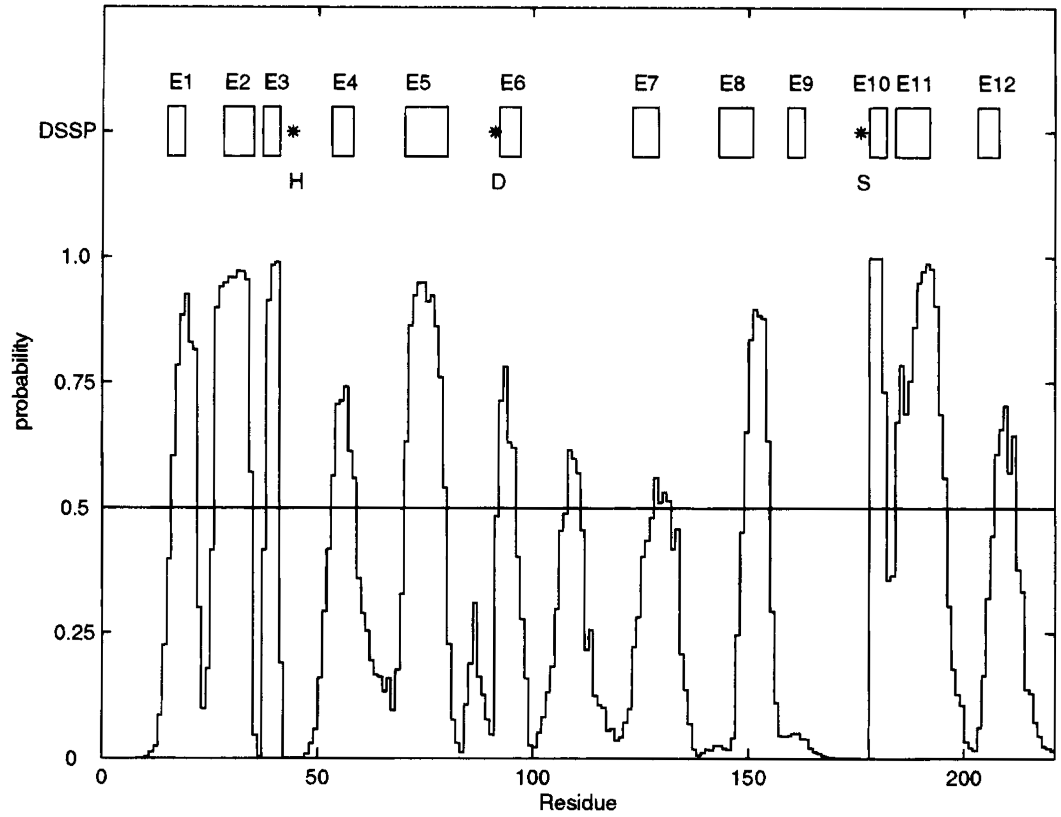


Fig. 4. An example of secondary structure prediction of hydrolase (PDB locus: 1FUJA) by pDSM. The probability of being in a strand is calculated by a smoothing algorithm. The prediction of being in a strand is made when the probability is higher than 0.5. The DSSP strand assignments are indicated by boxes. The catalytic sites are denoted by asterisks. Strands E3, E4, E5, and E10 are predicted with zero or one residue difference from DSSP assignments. Strand E7, E8, and E12 are predicted with positional shift. E1, E2, and E11 are lengthened, and E6 is shortened. Strand E9 is missed by pDSM prediction. There is also a false strand prediction between E6 and E7.

Secondary structure prediction

A residue is predicted to be in a helix if the probability of the residue being in a helix is larger than 0.5. As shown in the fully cross-validated results of Table 5, the pDSM prediction of secondary structure for globins is better than the DSM prediction alone. The average sensitivity of helix prediction is significantly higher (92% for pDSM compared to 73% for DSM), the specificity is a little higher (52% for pDSM compared to 48% for DSM), and the Q3 accuracy is also significantly higher (80% for pDSM compared to 66% for DSM). Compared to the secondary structure prediction of serine proteases in Table 3, the sensitivity of helix prediction for globins is higher than the sensitivity of strand prediction for serine proteases, and the specificity is lower.

Table 4. *Sensitivity and specificity of homology identification for globins by different methods*

Search method	Sensitivity (26 proteins) (%)	Specificity (77 proteins) (%)
Conserved sequence pattern	100	0
BLAST	42	100
DSM	58	90
pDSM	100	97

The higher sensitivity can be attributed to two factors. First, DSM analysis and most other secondary structure prediction methods are generally better at predicting helices than predicting strands. Second, the structure of globins is much less variable than serine proteases. Therefore, the DSM analysis for globins predicted the helices with a fairly high accuracy. Actually, helices have been slightly overpredicted, as we can tell from the relatively low specificities of helix prediction. In addition to the slight overprediction of helices, the relatively low specificity is also due to the low loop content of the globins. The helices in globins are usually connected by very short loops or turns. The number of residues in nonhelix regions of a globin is usually 30–50 (we count helix C as a normal helix, although DSSP sometimes assigns it as a 3_10 helix), while globins are 150 residues long on average. Therefore, the specificity is very sensitive to the wrong prediction of helices.

Discussion

The pDSM method has the following limitations. First, at least one determined structure is required for building a model. However, since no statistical training procedure is needed for model building,

Table 5. *The Q3 accuracy, sensitivity, and specificity of helix prediction for globins by DSMs and pDSMs*[a]

	Q3 (%)		Sensitivity (%)		Specificity (%)	
Loci	DSM	pDSM	DSM	pDSM	DSM	pDSM
1ASH	76	82	89	92	34	51
1BBBA	77	82	91	92	36	53
1BVC	88	87	96	94	64	67
1CMYB	25	77	19	91	92	39
1ECA	47	71	45	90	28	26
1FDHG	55	80	45	91	57	49
1FLP	82	85	95	92	44	67
1FSLA	76	83	91	92	32	59
1GDI	83	85	97	92	41	65
1HBG	78	80	94	92	32	49
1HBHA	51	83	53	94	50	58
1HBHB	61	81	61	93	54	46
1HBIA	51	84	53	92	41	62
1HDSA	58	72	61	100	54	42
1HDSB	43	60	49	88	45	28
1HLB	76	75	99	97	41	43
1HLM	72	72	99	95	39	43
1ITHA	76	81	91	92	40	57
1LHS	86	88	96	96	59	64
1MBA	80	79	91	90	46	49
1MYT	76	82	94	95	34	52
1OUTA	55	82	45	88	56	67
1OUTB	25	79	19	89	88	55
1SCTA	58	79	53	88	60	47
1SCTB	80	84	91	91	37	60
2LHB	80	81	95	94	42	51
Average	66	80(62)	73	92(62)	48	52(74)
STD	18	6	26	3	16	11

[a] The DSSP secondary structure assignments are taken as the true secondary structure. The numbers in parentheses, listed in the row Average, were obtained using the GOR algorithm (Garnier et al., 1978) and are for comparison only. The numbers in row STD are the standard deviations of each column.

the pDSM method does not require a large set of known homologous sequences or structures. Second, our current pDSMs are designed for single-domain globular proteins and are currently not suited for multiple-domain proteins. Third, the mapping of structure information from three to one dimension loses some information, such as the three-dimensional arrangement of the secondary structure elements and/or residue contacts. This limitation may explain why half (7 out of 14) of the false positives predicted by the pDSM as serine proteases have β-folds with the nonprotease packing.

The currently most successful means of inferring a protein's biochemical function is by homology, which is normally established through recognizable sequence similarity. However, when the query sequence is not sufficiently similar in sequence, a method is needed to detect remote homologs. With the complete sequencing of several genomes, there is a pressing need for such methods. To meet this need, the pDSM method exploits both sequence information (functionally conserved sequence pattern) and structure information that is encoded in DSMs. Our pDSM method uses the entire sequence to estimate the overall structural context in which a very limited amount of conserved sequence similarity may still be recognized.

The conditional probabilities of observing each amino acid, given a residue's structural environment (or state), are obtained statistically from a large representative set of known protein structures. This representative set is not limited to any one particular protein family, which gives us the freedom to build pDSMs for a protein family for which we have only one determined structure.

We intentionally used minimal sequence patterns for serine proteases and for globins as a very stringent test. However, the real conserved sequence patterns from multiply aligned sequences are generally more informative. We wanted to show that even with very limited patterns structure information could help improve specificity while significantly increasing sensitivity. A more general pattern could be used, such as a standard profile (Gribskov et al., 1987; Henikoff & Henikoff, 1991), which would be expected to further improve the specificity. It may also be possible, using mutational information such as that arising from alanine scanning (Cunningham & Wells, 1989), to suggest key residues for inclusion in a pDSM when insufficient sequence examples are available for conserved profile construction.

The pDSM analysis is both more sensitive and more specific than DSM analysis, because pDSMs put three constraints on homology identification: (1) the functionally conserved sequence pattern must be present, (2) the estimated secondary structure topology must be consistent with the fold, and (3) the conserved sequence pattern must occur in the correct structural context. In contrast, two-step methods, which look first for the conserved pattern and then check for the fold separately, do not impose the third constraint.

We are not suggesting that pDSM is superior in performance to all other threading methods. Instead, our results show that by essentially using the same threading method (DSMs in this paper) and embedding minimum sequence information, the fold recognition method can be used to recognize remote homologs with improved performance on fold recognition and secondary structure prediction. Earlier, Lathrop and Smith (1996) demonstrated in a threading analysis that the leghemoglobin can be more accurately aligned with the myoglobin structural model with two histidine residue positions fixed. We present in this paper a more systematic and thorough study of a similar method.

Materials and methods

Discrete state-space models

We have developed an extension of discrete state-space models (DSMs) (White, 1988; Stultz et al., 1993; White et al., 1994) to recognize a probable protein homolog that is unrecognizable by sequence comparison. It has been demonstrated that DSMs can be used to recognize proteins with similar folds by relying on similar patterns of implied structural states (Stultz et al., 1993) rather than sequence similarity. The DSMs are convenient mathematical structures for encoding information about patterns of secondary structures associated with different folds. The DSMs have the same mathematical structure as hidden Markov models (HMMs) (Rabiner, 1989). The DSMs, however, are built from the physical interpretation of a given structural fold to encompass all possible members. In contrast, most applications of HMMs in protein analysis have been based primarily on statistical analysis of training datasets. Such HMMs are trained to represent either sequence profiles of protein families with amino acid sequences as the training data (Krogh et al., 1994; Eddy, 1996; Sonnhammer et al., 1997) or structure profiles of protein folds with the secondary structure sequences as the training data (Di Francesco et al., 1997; Karplus et al., 1997). Goldman et al. (1996) described a simple HMM with three fully connected states for structure prediction while illustrating the significance of explicitly using evolutionary trees.

The DSM structural state of a residue position is defined by the secondary structure in which it occurs and its degree of exposure to the solvent. A DSM for a specific fold represents all possible sequences of structural states that are compatible with that fold. An example is shown in Figure 5.

The conditional probabilities of observing different amino acids, given the structural state of the residue position, are obtained from statistics on a large representative set of known protein structures (unpubl. data). These conditional probabilities are independent of the particular fold being modeled. The transition probabilities of moving from one structural state to another, in contrast, are selected to model the structural topology of a specific fold. The structural states and transition probabilities are selected so that the lengths and patterns of exposure and secondary structures conform to the specific fold being modeled. The detailed description of how to build DSMs has been published in two papers (White et al., 1994; Stultz et al., 1997).

We have built a library (http://bmerc-www.bu.edu/psa/) of DSMs for different protein folds in an attempt to cover the space of protein three-dimensional structures. An optimal filtering algorithm (White, 1988) is used to calculate the posterior probability of each model in our library, given the query sequence. The most probable folds for the query sequence are the folds that are modeled by the DSMs having the highest probabilities. Given the three most probable DSMs, the secondary structure for each residue is predicted using an optimal smoothing algorithm (Stultz et al., 1993). This smoothing algorithm computes the posterior probabilities of residues being in each structural state, given the entire protein sequence.

pDSM for trypsin-like serine proteases

Protein families with an abundance of both known structures and sequences provide the best validation of the proposed method, especially if some of the sequence similarities fall below the "twilight zone" (Doolittle, 1981) of 15 to 20% identities. The trypsin-like serine proteases, which are involved in the hydrolyzation of peptide bonds, are a well-studied and diverse protein family with more than 200 structures in the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977) and more than 400 sequences in Genbank (Benson et al., 1994). The pairwise sequence identities of two serine proteases can be less than 10%.

Besides the wealth of structures and sequences, serine proteases are also a well-studied protein family from the biochemical point of view. A review of the catalytic mechanism and the structure-function relationship of serine proteases can be found in Stroud et al. (1975). The catalytic site consists of a His-Asp-Ser triad. Several homology modeling studies based on sequence analysis of serine proteases have been presented (Greer, 1990; Alexandre et al., 1996). Pearson (1997) reviewed the problem of identifying distantly related proteins by sequence comparison with serine protease as the example.

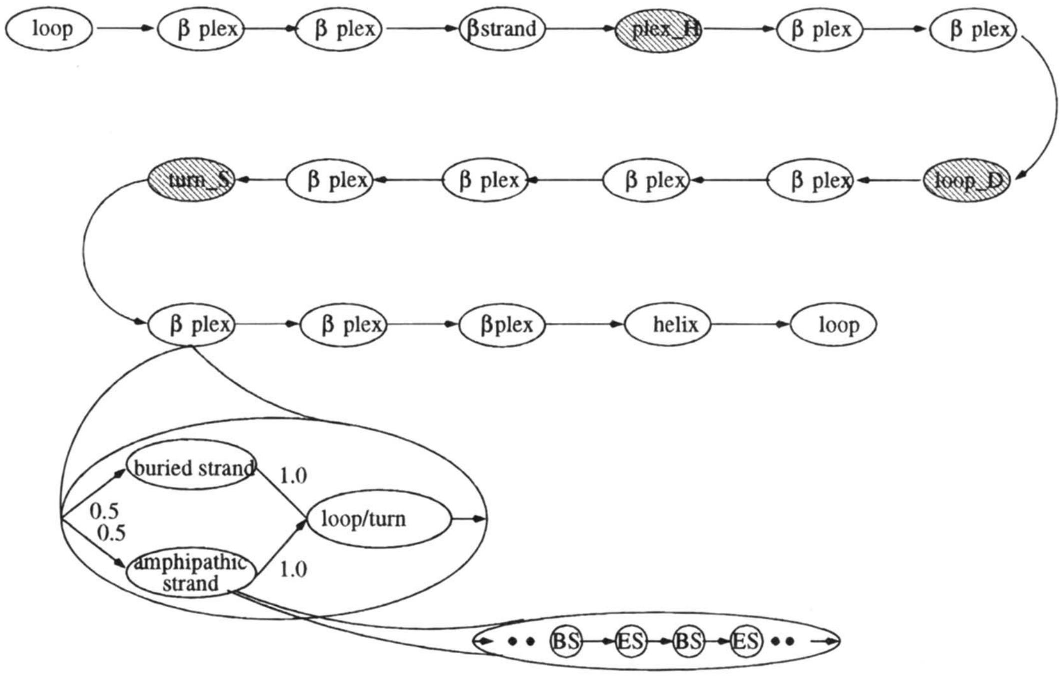


Fig. 5. Schematic of the pDSM for cluster 1 serine proteases. Each oval circle represents a structural building block. Each *β plex* consists of a *β* strand (buried or amphipathic) and the connecting region, which is formed by loops, turns, and sometimes, short helices. At the bottom of the figure, the environmental states of residues in an amphipathic strand are highlighted; they alternate between exposed strand (ES) and buried strand (BS). The paths are labeled with the transition probabilities of moving from one structural element to another. The 0.5 probabilities leading to the buried and amphipathic strands indicate that both types of strand are equally probable.

Functionally conserved sequence pattern

The conserved catalytic sequence pattern is $X_{24-69}HX_{18-86}DX_{40-109}$ SX_{44-141}, where X denotes any amino acid, and the subscript denotes the range of residue positions. This minimal pattern covers all of the serine proteases in our positive sets. While there are more complex patterns of recognizable sequence similarity among the majority of serine proteases, the use of this minimal pattern is sufficient to demonstrate the significance of structural context in identifying very distant homologs.

Construction of test data sets

Positive control sets

We selected our initial trypsin-like serine proteases from the PDB, which provides both three-dimensional structure coordinates and a functional description of each protein. The available structure information allows us to validate the pDSM performance on secondary structure prediction.

In the PDB release 80, there are 253 entries for trypsin-like serine proteases. Among these are identical sequences, or sequences differing only by a few mutated residues. We chose 32 sequences as the positive data set (Table 6). All other sequences have more than 80% sequence identities with at least one of these 32. The 32 positives were grouped, according to sequence similarities, into two clusters by the algorithm PIMAII (Adams et al., 1996). The proteins within each cluster are significantly more similar to each other, both in sequence and structure, than the proteins in the other cluster. The equivalent pairwise sequence identities of any pair, one from each cluster, are fewer than 25%, and the structures differ from each other with a wide range of loop sizes, lengths, and number of strands. This reflects the fact that proteins in different clusters are more remotely related than those within the same cluster. The remoteness is confirmed in the functional annotation and in the organisms in which the proteins are found.

This positive set is rather small. To find more serine proteases, we chose an additional 111 sequences (Table 6) from Genbank that are annotated as serine proteases, single domain, with fewer than 80% pairwise equivalent sequence identities to each other or the original 32. All fragments that are not complete serine protease sequences were removed along with subtilisin serine proteases, which adopt the different three-dimensional fold from that of trypsin-like serine proteases and contain a different active site signature (Asp32–His64–Ser221) (Bode et al., 1987).

Negative control set

We selected 206 sequences from the PDB (Table 6) that satisfy the following criteria: (1) They are between 154–317 amino acids long, which is the length range compatible with the pDSMs for serine proteases and the serine proteases in the positive sets. (2) The pairwise equivalent amino acid identities of all sequences in this dataset are fewer than 25% to eliminate all obvious homologs. (3) They cover a wide variety of species (the sequences in our data set come from 93 organisms ranging from virus to human). (4) They have a wide variety of identified nonserine-protease functions. (The sequences in our data set have as many as 55 enzymatic functions by assigned EC number (Bairoch, 1994).) (5) Their structures distribute widely across the structure space. (We have 30 α proteins, 47 β proteins, and 129 $\alpha\beta$ proteins. The structure class of each protein was obtained from CATH (Michie et al., 1996) or SCOP (Murzin et al., 1995), or in a few cases manually.) (6) Finally, they contain the trypsin-like sequence pattern. This last criterion is essential because every sequence not containing that pattern would automatically be identified as a nontrypsin protein by our method.

Table 6. *Test data sets of trypsin-like serine proteases*[a]

	Positive control sets
PDB	Cluster 1: 1ABIH 1ACBE, 1BIT, 1BMAA, 1BRA, 1BTP, 1DST, 1ELT, 1FUJA, 1HCGA, 1HF1, 1HNEE, 1HYLA, 1LMWB, 1PCU, 1PFA, 1PFXC, 1PYTC, 1PYTD, 1RTFB, 1SGT, 1TON, 1TRY 2CP1, 2KAI(A+B), 3RP2A Cluster 2: 1ARB, 1GBAA, 1HPGA, 1SGC, 1SGPE, 2SFA
Genbank	AB003670 D16687 D30760 D45173 D45417 D63858 D67078 D67079 D67080 D67083 D67084 J04071 J05177 K01173 L04749 L08428 L10038 L16805 L19694 L24914 L24915 L33404 L76741 M11590 M17103 M17104 M18608 M18700 M19647 M24379 M24664 M24665 M33109 M36902 M54900 M57401 M72150 M77814 M81392 M81395 S44609 U03760 U04962 U05203 U13770 U15155 U15157 U21917 U25648 U28641 U32937 U35237 U38463 U39500 U40653 U41476 U43525 U44951 U49931 U56423 U56956 U57055 U57062 U57063 U58945 U62801 U65411 U65412 U66472 U66473 U67907 U67908 U67909 U67910 U67911 U67915 U72330 X15679 X17351 X56744 X59012 X64362 X64363 X66415 X70074 X71438 X75016 X75363 X76886 X78490 X78545 X78875 X83221 X86369 X94691 X94982 X95078 X96387 X97635 Y08133 Y11878 Y11879 Z12296 Z18890 Z22930 Z27239 Z32645 Z49813 Z49815 Z49833 Z69978
	Negative control set
PDB	102L 12CA 1ABMA 1ABN 1ABRA 1ABRB 1ACMA 1ACYH 1ACYL 1AERA 1AHA 1AHHA 1AIN 1AK2 1AMM 1AMP 1APA 1APNA 1APXA 1ARL 1AST 1ATLA 1ATND 1BBT1 1BBT2 1BBT3 1BCFA 1BCRA 1BEC 1BERA 1BLC 1BMC 1BMFG 1BMTA 1BPB 1BPLA 1BPLB 1BROA 1BTMA 1CAUA 1CBY 1CDDA 1CFB 1CGE 1CHD 1CHKA 1CKIA 1CME 1CNE 1CNSA 1CNV 1COLA 1COV1 1COV2 1CPJA 1CPM 1CRVA 1CSEE 1CSMA 1CTT 1CYDA 1DAAA 1DBP 1DBQA 1DEAA 1DHR 1DHY 1DIH 1DKXA 1DLHA 1DPB 1DPRA 1DR1 1DSBA 1ECPA 1EDB 1EDHA 1EDT 1EFUB 1EMA 1ENO 1ENY 1EPAA 1ERIA 1ESFA 1ESP 1EXP 1EZM 1FATA 1FC1A 1FINA 1FINB 1FLV 1FNB 1FRVA 1FUA 1FVPA 1GCA 1GDOA 1GFF2 1GHR 1GLPA 1GLV 1GNE 1GPC 1GRIA 1GSQ 1GTPA 1GYM 1HAR 1HAVA 1HDCA 1HHGA 1HLPA 1HMPA 1HRI3 1HTLA 1HXN 1HYHA 1IAF 1IGNA 1ILLG 1ILLR 1ILMB 1IMAA 1IRC 1IRK 1JUD 1KANA 1KXA 1LAFE 1LAUE 1LBD 1LCA 1LGYA LIMA 1LTPL 1LXA 1LYAB 1MASA 1MAT 1MEC2 1MLA 1MML 1NAL1 1NAR 1NBAA 1NIPA OCCB 1OCCC 1OPR 1ORB 1PBN 1PCRH 1PCRL 1PCRM 1PCZA 1PDA 1PEX 1PGS 1PHK 1PLQ 1PMAB 1PRTA 1PRTB 1PYAB 1PYP 1RGS 1RVAA 1SBP 1SCHA 1SCUA 1SE2 1SFE 1SMNA 1SMVA 1TDE 1TFD 1TFR 1THJA 1THTA 1TIA 1TLK TML 1TSRA 1TTPA 1UKY 1VCAA 1VHRA 1VIRC 1VMOA 1XVAA 1YAL 1YPTA 1ZYMA 2AT2A 2DLN 2HIE 2PCDA 2PCDM 2POR 2SCPA 2STV 2TCT 2TMAA 3PGM

[a] The sequences from the PDB are listed by their locus names and chain numbers. The sequences from the Genbank are listed by accession number.

Model building

All serine proteases fold into two antiparallel β barrels. Besides the variations in the length and exposure pattern of strands and the connecting regions between strands, there are two distinct types of structures, each corresponding to one sequence cluster. The primary differences occur in the region between catalytic Asp and Ser. The structure of cluster 1 has four strands in that region, connected usually by long loops. The structure of cluster 2 has six strands connected by relatively short loops or turns. We thus built two types of DSMs for these two types of structures. Within each DSM, the lengths and other secondary structural parameters were probabilistically specified to cover the range of observed and anticipated variation. As discussed in Stultz et al. (1993), the underlying state transition probabilities are assigned rather than being obtained via a statistical training procedure.

Next, the conserved sequence pattern was embedded in these DSMs as shown schematically in Figure 5. Each β *plex* Markov chain state consists of a β strand state and a connecting set of states modeled as loops, turns, and/or short helices. Each of these states is in turn modeled as a sequence of residue position states. The embedding of the conserved sequence pattern was done by replacing particular position states in the DSM by the proper sequence pattern element states (see Fig. 6). In serine proteases, where only one amino acid is observed in each conserved position of the conserved sequence pattern, the probability of observing that one amino acid is one, and the probability of observing any other amino acid is zero. The shaded circles of Figure 5 show the secondary structures containing the embedded sequence pattern elements. The *plex_H* is a connecting region starting with a small 3_10 helix. The state in the middle of the 3_10 helix is replaced by a conserved His. The *loop_D* is a short loop followed by a conserved Asp. The *turn_S* has a conserved Ser.

pDSM for globins

Globins are a family of proteins with nearly identical three-dimensional structures, but greatly differing sequences. Some pairwise equivalent sequence identities in this family are only 16%, well below statistical significance. Their sequence-structure relationship has been studied extensively and reported in Lesk and Chothia (1980) and Bashford et al. (1987).

Functionally conserved sequence pattern

The globins are heme-binding proteins with only two universally conserved residues: a Phe in helix C packing against the heme cofactor, and the proximal His in helix F, which coordinates the central heme iron atom. The conserved sequence pattern for globins is thus $X_{41\text{-}60}FX_{38}HX_{43\text{-}68}$. This minimal pattern covers all the globins.

Construction of test data sets

Positive control set

Among 234 globin entries in the PDB, we selected 26 sequences with pairwise sequence identities less than 80% for our positive set (Table 7). These 26 globins were grouped into six sequence clusters by PIMAII (Adams et al., 1996).

Negative control set

The negative set for globins was constructed using the same criteria we used for trypsin-like serine proteases, except that the length range was 126–173 residues and the conserved sequence pattern was for globins. Since there are only a few sequences in the PDB that meet all of these criteria, we used the SWISS-PROT database (Bairoch & Boeckmann, 1994). Seventy-seven sequences (Table 7) were selected as negatives from the SWISS-PROT release 33. They are from 53 different organisms and have 30 enzymatic functions (EC numbers): 18 are all α proteins, 22 are all β proteins, and 37 are $\alpha\beta$ proteins. This structure information was inferred by sequence similarity to a given PDB entry.

Model building

Globins fold into an α box fold with eight α helices, A through H. The globin fold varies in the lengths of the helices and, in some structures, helix D does not exist. The globin fold is in general more constrained than the serine protease fold, even in its surface

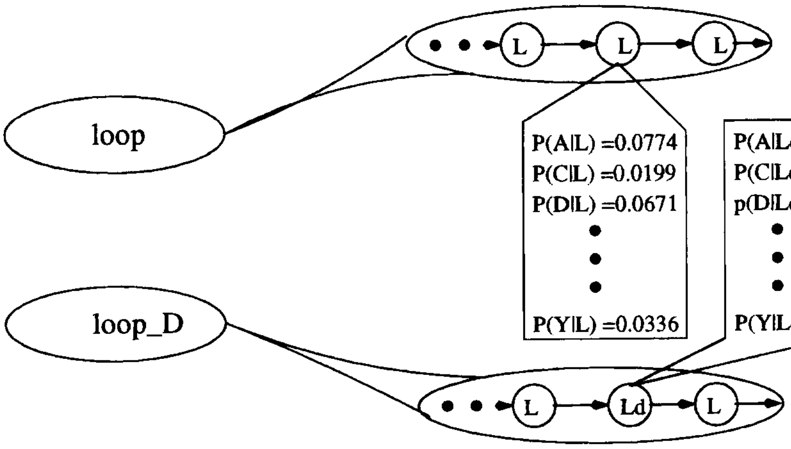


Fig. 6. Illustration of embedding sequence pattern elements into DSMs. The loop structural state (L) is replaced by the conserved sequence pattern element distribution. As shown in the *loop_D* case for the serine proteases, this is a conserved Asp only.

Table 7. *Test data sets of globins*[a]

	Positive control set
PDB	Cluster 1: 1BBBA 1CMYB 1FDHG 1HBHA 1HBHB 1HDSA 1HDSB 1OUTA 1OUTB
	Cluster 2: 1ECA 1FLP 1ITHA
	Cluster 3: 1FSLA 1GDI 1HBG
	Cluster 4: 1HBIA 1HLB 1HLM 1SCTA 1SCTB
	Cluster 5: 1BVC 1LHS 1MBA 1MYT 2LHB
	Cluster 6: ASH
	Negative control set
SWISS-PROT	AMPM_CLOPE AMY4_HORVU ARF3_DROME ARG1_YEAST ARGR_ECOLI AZUR_ALCDE BCCP_ECOLI BCRF_EBV BFR_ECOLI CALM_CHLRE CC21_PEA COAG_CARRO COX1_GEOSD CPC_CUCSA CRBA_RAT CYB5_BOVIN CYP1_ARATH CYPB_BACSU CYPB_ECOLI CYS3_OSTOS D112_ARATH DYRA_STAAU DYR_HALVO E13J_TOBAC FLAV_CHOCR FLAV_DESDE FRH1_SCHMA HXA9_AMBME IAA_HORVU IDE3_ERYCA IF3_BACST IGF1_ONCKI IIK_SOLTU IL1X_HUMAN IL2_FELCA IL4_HUMAN ING_MOUSE IPYR_BARBA KAPA_PIG LIT1_MOUSE LSHB_EQUAS LYC3_PIG MEMG_METCA MEP1_SOYBN MGF1_MOUSE MLEL_DROME MUP8_MOUSE NEU2_HUMAN NIA_LOTTE OBP_BOVIN OGT_MYCTU PA21_HUMAN PER2_HORVU PHEA_ANASP PLAS_ANASP PPAL_YEAST PTGA_ECOLI PTPS_DROME QOR_SALTY RB13_RAT RBS1_ORYSA RET1_ONCMY RR5_CYAPA RUVC_ECOLI SFA2_STRFR SOD1_ORYSA SODM_CORDI TELO_RABIT TGFA_HUMAN TOX5_BORPE TPIS_MYCPI UBC7_ARATH UCRI_BRAJA URE2_STAXY VGG_BPG4 YBP2_DESAM YHLB_VIBCH

[a] The sequences from the PDB are listed by their locus names and chain numbers. The sequences from SWISS-SPROT are listed by their Ids.

loops. Following the method of building models for serine proteases described above, we built the DSM for the globin fold and the pDSM by embedding the two conserved residues in their respective positions.

Model validation

Our study is fully cross-validated: the serine proteases and globins were excluded from the set of proteins used to establish the conditional probabilities of amino acids, given the structural state of the residue.

Sensitivity and specificity provide a good measure of how reliably the pDSMs recognize homologs and predict the associated secondary structures. Sensitivity is the percentage of all positives that are correctly recognized as positives. Specificity is the percentage of all negatives that are correctly recognized as negatives. High reliability corresponds to both high sensitivity and high specificity.

These two measures are used to compare the reliability of pDSM analysis with direct sequence similarity searches using BLAST (Altschul et al., 1990), which is the current standard fast sequence comparison method for inferring homology. In addition, these measures are also used separately to compare the reliability of homology predictions based on the conserved sequence pattern and the DSM structure-fold information.

The sensitivity and the specificity of direct sequence similarity searches vary for each query sequence. Thus, we define average sensitivity sen_{ave} and average specificity spe_{ave} by the equations

$$sen_{ave} = \frac{1}{n} \sum_{i=1}^{n} sen_i, \tag{1}$$

$$spe_{ave} = \frac{1}{n} \sum_{i=1}^{n} spe_i, \tag{2}$$

where sen_i and spe_i are the sensitivity and specificity for the query sequence i, and n is the number of queries.

Acknowledgments

This research was sponsored in part by TASC, Inc. and by grant P41 LM05205-14 from the National Library of Medicine. We acknowledge the assistance of our colleagues Xiaolin Zhang, Jim Freeman, and Bob Rogers. We thank Nancy Sands for critical reading of the manuscript.

References

Adams RM, Das S, Smith TF. 1996. Multiple domain protein diagnostic patterns. *Protein Sci 5*:1240–1249.

Alexandre J, Barbosa RG, Saldanha JW, Garratt RC. 1996. Novel features of serine protease active sites and specificity pockets: Sequence analysis and modeling studies of glutamate-specific endopeptidases and epidermolytic toxins. *Protein Eng 9*:591–601.

Allaire M, Chernaia MM, Malcolm BA, James MN. 1994. Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteases. *Nature 369*:72–76.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res 25*:3389–3402.

Bairoch A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res 19 Suppl*:2241–2245.

Bairoch A. 1994. The ENZYME database. *Nucl Acids Res 22*:3626–3627.

Bairoch A, Boeckmann B. 1994. The SWISS-PROT protein sequence data bank: Current status. *Nucl Acids Res 22*:3578–3580.

Bashford D, Chothia C, Lesk AM. 1987. Determinants of a protein fold unique features of the globin amino acid sequences. *J Mol Biol 196*:199–216.

Benson D, Boguski M, Lipman DJ, Ostell J. 1994. Genbank. *Nucl Acids Res* 22:3441–3444.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science 277*:1453–1462.

Bode W, Papamokos E, Musil D. 1987. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. *Eur J Biochem 166*:673–692.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne D, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science 273*:1058–1073.

Cunningham BC, Wells JA. 1989. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science 244*:1081–1085.

Di Francesco V, Garnier J, Munson PJ. 1997. Protein topology recognition from secondary structure sequences: Application of the hidden Markov models to the α class proteins. *J Mol Biol 267*:446–463.

Doolittle RF. 1981. Similar amino acid sequences: Chance or common ancestry? *Science 214*:149–159.

Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol 6*:361–365.

Garnier JR, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol 120*:97–129.

Gibrat J-F, Madej T, Bryant SH. 1996. Surprising similarities in structure comparison. *Curr Opin Struct Biol 6*:377–385.

Goldman N, Thorne JL, Jones DT. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol 263*:196–208.

Greer J. 1990. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins 7*:317–334.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Nat Acad Sci USA 84*:4355–4358.

Hartley BS, Kauffman DL. 1966. Corrections to the amino acid sequence of bovine chymotrypsinogen A. *Biochem J 101*:229–231.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucl Acids Res 19*:6565–6572.

Holm L, Sander C. 1995. DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem Sci 20*:345–347.

Holm L, Sander C. 1996. Mapping the protein universe. *Science 273*:595–602.

Holm L, Sander C. 1997. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. *Proteins 28*:72–82.

Hubbard T. 1997. New horizons in sequence analysis. *Curr Opin Struct Biol 7*:190–193.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*:2577–2637.

Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. 1997. Predicting protein structure using hidden Markov models. *Proteins 1*:134–139.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology application to protein modeling. *J Mol Biol 235*:1501–1531.

Kunst F, Ogasarawa N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero V, Bessieres P, Bolotin A, Borchert S, et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature 390*:249–256.

Lathrop RH, Smith TF. 1996. Global optimum protein threading with gapped alignment and empirical pair score functions. *J Mol Biol 255*:641–665.

Lathrop RH, Webster TA, Smith TF. 1987. Ariadne: Pattern-directed inference and hierarchical abstraction in protein structure recognition. *Commun ACM 30*:909–921.

Lesk AM, Chothia C. 1980. How different amino acids sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol 136*:225–270.

Lipinska B, Fayet O, Baird L, Georgopoulos C. 1989. Identification, characterization and mapping of the *Escherichia coli* htrA gene, whose product is essential for bacterial growth only at elevated temperatures. *J Bacteriol 171*:1574–1584.

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A. 1997. Overview of the yeast genome. *Nature 387(6632 Suppl)*:7–65.

Michie AD, Orego CA, Thornton JM. 1996. Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol 262*:168–185.

Murzin A, Brenner SE, 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.

Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. 1997. Extracting protein alignment models from the sequence database. *Nucl Acids Res 25*:1665–1677.

Pearson WR. 1997. Identifying distantly related protein sequences. *CABIOS 13*:325–332.

Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE 77*:257–286.

Rufo GA Jr, Sullivan BJ, Sloma A, Pero J. 1990. Isolation and characterization of a novel extracellular metalloprotease from *Bacillus subtilis*. *J Bacteriol 172*:1019–1023.

Sibbald PR, Argos P. 1990. Scrutineer: A computer program that flexibly seeks and describes motifs and profiles in protein sequence databases. *CABIOS 6*:279–288.

Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins 28*:405–420.

Stocker W, Bode W. 1995. Structural features of a superfamily of zinc-endopeptidases: The metzincins. *Curr Opin Struct Biol 5*:383–390.

Stroud RM, Krieger M, Koeppe RE II, Kossiakoff AA, Chamber JL. 1975. Structure-function relationships in the serine proteases. In: Reich E, Rifkin DB, Shaw E, eds. *Proteases and biological control*. New York: Cold Spring Harbor Lab. pp 13–32.

Stultz CM, Nambudripad R, Lathrop RH, White JV. 1997. Predicting protein structure with probabilistic models. In: Allewell N, Woodward C, eds. *Protein structural biology in bio-medical research*, Vol. 22B, in Bittar EE, ed. *Advances in molecular and cell biology*. Greenwich: JAI Press. pp 447–495.

Stultz CM, White JV, Smith TF. 1993. Structural analysis based on state-space modeling. *Protein Sci 2*:305–314.

Taylor WR. 1986. Identification of protein sequence homology by consensus template alignment. *J Mol Biol 188*:233–258.

White JV. 1988. Modeling and filtering for discretely valued time series. In: Spall JC, ed. *Bayesian analysis of time series and dynamic models*. New York: Marcel Dekker. pp 255–283.

White JV, Stultz CM, Smith TF. 1994. Protein classification by stochastic modeling and optimal filtering of amino acid sequences. *Math Biosci 119*:35–75.