Accelerated Publication

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 277, No. 22, Issue of May 31, pp. 19243–19246, 2002
© 2002 by The American Society for Biochemistry and Molecular Biology, Inc.
Printed in U.S.A.

Substrate Recognition Drives the Evolution of Serine Proteases*

Received for publication, March 6, 2002,
and in revised form, March 28, 2002
Published, JBC Papers in Press, March 29, 2002,
DOI 10.1074/jbc.C200132200

Thierry Rose and Enrico Di Cera‡
From the Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110

A method is introduced to identify amino acid residues that dictate the functional diversity acquired during evolution in a protein family. Using over 80 enzymes of the chymotrypsin family, we demonstrate that the general organization of the phylogenetic tree and its functional branch points are fully accounted for by a limited number of residues that cluster around the active site of the protein and define the contact region with the P1–P4 residues of substrate.

Phylogenetic trees of proteins are useful to map evolutionary developments and functional diversity within a given family. Progress made in the cloning and expression of a large number of autologous proteins has generated large databases that can now be used to draw quantitative conclusions on how different specificities and functions emerged during evolution. Residues that are conserved throughout the family provide important determinants of function or structural stability, but cannot be linked to the onset of functional differences or specificities because they are evolutionarily non-informative. On the other hand, residues subject to a high rate of coding nucleotide mutation are often inconsequential on function or structural stability. However, some residues must have changed during evolution to provide the branching patterns observed in phylogenetic trees. Methods that identify these residues can potentially unravel how proteins evolved in general and assist in the identification of function in newly discovered genes.

Several methods have been proposed to identify residues potentially involved in function. Some of these methods explore the functional divergence from sequence multi-alignment and phylogenetic trees (1). Others combine phylogenetic and structural data (2–4). Evolutionary trace (2), ConSurf (3), and three-dimensional cluster analysis (4) methods scan for residue conservation within sub-group of proteins, pooled together within the same tree branch or filtered over identity threshold, and define surface patterns representative of the subgroup. Here we introduce an alternative method that does not require subdivision of the protein family and that takes into account not only conserved residues, but also residue replacements fixed along the transition from one branch of the evolutionary tree to other daughter branches. The method evaluates how every position of the sequence multi-alignment contributes to the architecture of the tree and provides a three-dimensional map of the regions of the protein that played a determining role during evolution.

MATERIALS AND METHODS

700 protein sequences were pulled out from the NCBI non-redundant data base (National Library of Medicine, National Institutes of Health, Bethesda, MD) by several iterations of PSI-BLAST (5) using trypsin homologues as queries. All sequences were aligned with Clustal W1.8 (6). The best alignment obtained after different circular permutations of sequences and incorporation of the alignment from three-dimensional structure superposition was retained. Sequences were clustered by branches into 80 groups using a maximum threshold of 50% for sequence pair identity from a neighbor-junction tree performed with Clustal and accounting for 500 bootstraps (7). The global alignment was checked manually first among groups, then group by group, to improve gap extensions. Eighty protein sequences, one per cluster, were selected keeping human sequences whenever possible (42 out of 115 human serine proteases) or sequences from the closest species. Four sequences were also retained to cover the selection from a previous study (8). The minimum sequence pair identity was 6% with an average of 27% and a median of 28%. Only two residues (Cys^{182}, Gly^{196}) are absolutely conserved over all sequences.

Bootstrap sampling was carried out from sequence alignment using Seqboot. Pairwise sequence relationships were quantified on 200 bootstraps using a distance matrix calculated with Protdist and based on the PAM1 substitution matrix (7). The average pair distance was 1.82. Trees were built from Fitch on the 200 bootstraps using the global rearrangement option (9, 10). The final tree was drawn using Consense (7). The average root distance was 1.46, the average sum of squares between sequence pair distances calculated from a tree and distances from its corresponding matrix was 43.00 and the average percent S.D. was 7.88. Seqboot, Protdist, Fitch, and Consense are components of the Phylip package, Version 3.57c (7).

Impact values were calculated for 230 positions corresponding to residues 16–245 of chymotrypsin using Equations 1–3 in the text. Three-dimensional models of protein structures not available in the RCSB Protein Data Bank (11) were built by homology modeling with Modeler (12) and optimized by minimization with Discover (Accelrys, San Diego, CA). Substrate-protease models were built from the x-ray structure 1PPB and 1NRS of thrombin bound at the active site with H-D-Phe-Pro-Arg-CH_2Cl (13) or Leu-Asp-Pro-Arg (14) extended to Leu-Asp-Pro-Arg-Ser-Phe-Asn. The active site region was partitioned into specificity sub-sites from S4 through S3′. Accessible surface areas were calculated as described previously (15).

RESULTS AND DISCUSSION

A phylogenetic tree associated with serine proteases of the chymotrypsin family is shown in Fig. 1. Sequences are grouped according to main functional properties, and their separation into branches is independent of how the root of the tree is chosen (8). The contribution of each residue in the sequence to the definition of the tree was quantified in terms of an "impact" value. A key parameter associated with the tree is the sum of squares of the deviations between the distances of two sequences a and b in the tree, d_{ab}, and in the matrix of aligned sequences, m_{ab}, normalized by the sum of squares of the d_{ab} values, *i.e.*

$$\sigma^2 = \frac{\sum_{a=1}^{N}\sum_{b=1}^{N}(d_{ab} - m_{ab})^2}{\sum_{a=1}^{N}\sum_{b=1}^{N} d_{ab}^2}, \qquad \text{(Eq. 1)}$$

where N is the number of positions in the sequence. For each

* This work was supported in part by National Institutes of Health Research Grants HL49413 and HL58141. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Dept. of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Box 8231, 660 S. Euclid Ave., St. Louis, MO 63110. Tel.: 314-362-4185; Fax: 314-747-5354; E-mail: enrico@biochem.wustl.edu.

This is an Open Access article under the CC BY license.

Seikagaku Corporation Exhibit 2080 - Page 1 of 4
Charles River Laboratories, Inc. v. Seikagaku Corporation
IPR2025-00440

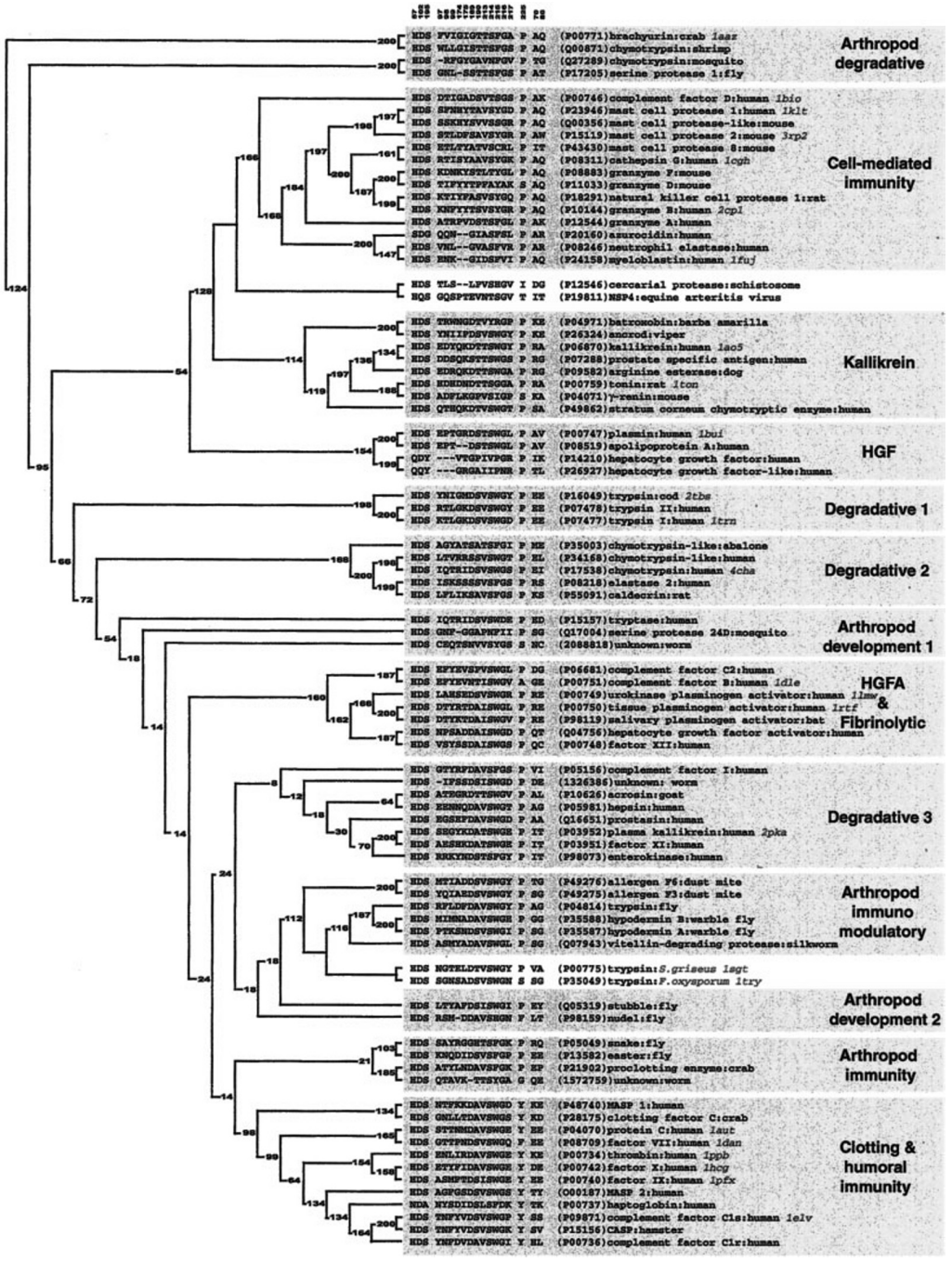

FIG. 1. **Phylogenetic tree of 84 serine proteases of the chymotrypsin family built from the multi-alignment of the proteolytic domains comprising residues 16–245.** The tree is the most frequent outcome from 200 bootstrap samples, is unrooted, and does not have an evolutionary clock. Bootstrap values are reported at the branch nodes, except for cercarial, NSP4, *Streptomyces griseus* and *Fusarium oxysporum* trypsin sequences. Sequence accession codes from Swiss-Prot and GenBank™ are reported in *parentheses* and RCSB Protein Data Bank ID of available x-ray three-dimensional structure files are noted in *italic*. Haptoglobin and azurocidin were included in the tree, although they have no catalytic activity. Several key residues are indicated in front of each sequence, with their chymotrypsin numbering noted at the top: residues of the catalytic triad (57, 102, 195), the S1–S4 sub-sites (97–99, 174–175, 189–190, 213–217), residues defining Na^+ binding (Tyr at position 225), or Ca^{2+} binding (Glu at positions 70 and 80).

position k in the sequence, the residue at that position was replaced by all 20 amino acids and a gap for all sequences in the alignment, and the resulting value of σ^2 derived from the new set of m_{ab} values was calculated in each case using a PAM matrix. The d_{ab} values from the original parent tree generated by the alignment of unperturbed sequences were kept constant. The average value of σ_k^2 was expressed as the linear combination,

$$\sigma_k^2 = f_{k,A}\sigma_{k,A}^2 + f_{k,C}\sigma_{k,C}^2 + \cdots + f_{k,Y}\sigma_{k,Y}^2 + f_{k,gap}\sigma_{k,\mathrm{gap}}^2 \qquad \text{(Eq. 2)}$$

where $\sigma_{k,x}^2$ is the value of σ^2 obtained when residue at position k is replaced with the amino acid x = Ala, Cys, . . . Tyr, or a gap, and $f_{k,x}$ is the correct weighting factor associated with the substitution (7). The impact value of residue k is defined as follows

$$I_k = \frac{exp(-|\sigma^2 - \sigma_k^2|)}{\sum_{k=1}^{N} exp(-|\sigma^2 - \sigma_k^2|)} \qquad \text{(Eq. 3)}$$

Seikagaku Corporation Exhibit 2080 - Page 2 of 4
Charles River Laboratories, Inc. v. Seikagaku Corporation
IPR2025-00440

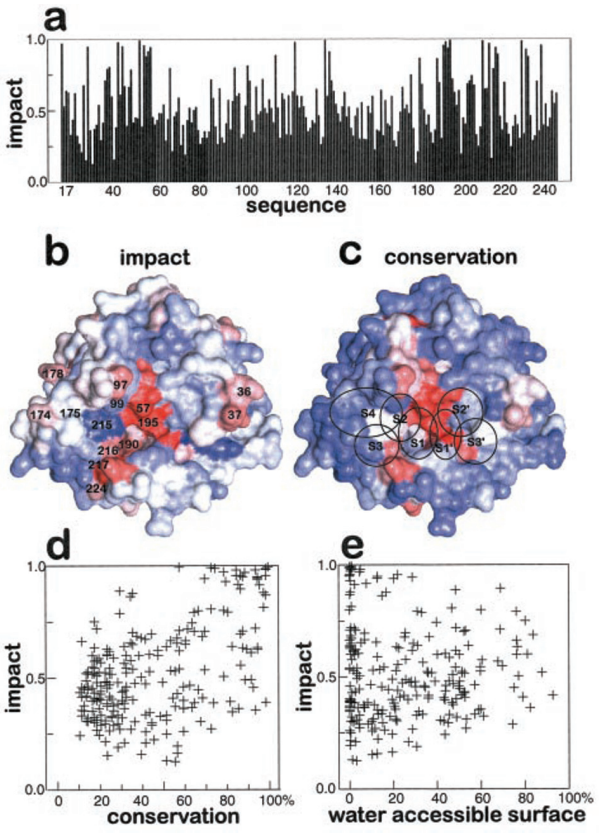

FIG. 2. *a*, impact values of 230 positions of the sequence of a consensus serine protease. The values show peaks at discrete positions along the sequence. *b* and *c*, solvent-accessible surface of chymotrypsin (4CHA) color-coded according to impact values (*b*) or residue conservation (*c*). The color scheme is a continuum spectrum from *blue* to *red* mapping the impact values from 0 (*blue*) to 1 (*red*) or conservation from 0% (*blue*) to 100% (*red*). Residues with high impact are noted (*b*). Of these, residues 36, 174, and 178 have very low conservation. Substrate recognition sub-sites discussed in the text are also indicated (*c*). The correlation between impact values and residue conservation (*d*) is weak ($r = 0.4$), and no correlation ($r = 0.0$) is seen between impact values and water accessibility (*e*). The region defining the S1–S4 sub-sites has the largest impact among all patches of accessible surface area of comparable size (see also Fig. 3). This region also has relatively low conservation and contains the smallest set of residues capable of reproducing the phylogenetic tree in Fig. 1 in its entirety.

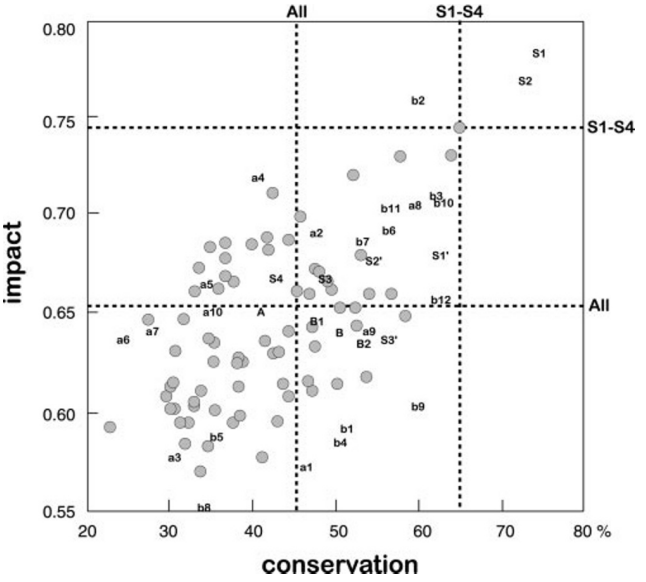

FIG. 3. **Impact values of set of residues *versus* average residue conservation.** Secondary structure sets are defined as individual β-strands b1-(28–36), b2-(37–49), b3-(50–54), b4-(66–68), b5-(81–95), b6-(99–109), b7-(133–144), b8-(150–162), b9-(179–184), b10-(196–203), b11-(206–213), b12-(226–230); β-barrels B1 and B2 comprise β-strands b1–b6 and b7–b12, respectively. B groups all β-strands. Helices and loops are defined as a1-(16–27), a2-(55–65), a3-(69–80), a4-(96–98), a5-(110–132), a6-(145–149), a7-(163–178), a8-(185–195), a9-(214–225), a10-(231–245). A groups all loops. *Gray circles* represent surface patches of clustered residues spanning 1100–1300 Å^2 (104 total). Individual sub-sites are defined as S4-(96–99,174–175,214–216), S3-(173–174,216–218,224), S2-(57,98–99,174,192–196,214–216), S1-(42,57,189–196,213–221,225–227), S1′-(40–42,57–58,73,143,191,193), S2′-(39–42,73,141–143,147,150–152,191–194), S3′-(33–34,38–42,58,73,141,143,151,193). The impact values and average residue conservation referring to the S1–S4 sub-sites (*S1–S4*) and the entire sequence (*All*) are shown by *discontinuous lines*.

and is bound from 0 to 1. The impact of a set of p positions in the sequence is calculated from the sum of p terms as in Equation 3, divided by p. The impact value I_k is maximum when $\sigma_k^2 \approx \sigma^2$, *i.e.* when position k contains a residue or gap that is absolutely conserved among all sequences or when the distribution of amino acids or gap substitutions at position k reproduces closely the functional branching of the parent tree. Therefore, the comparison of the residue or gap conservation at position k with the impact I_k enables the identification of residues with high impact and relatively low conservation that have been replaced during evolution to generate the functional diversity in the parent tree.

Fig. 2*a* shows the profile of impact values of individual positions along the sequence of the catalytic domain of a consensus serine protease. High values of the impact parameter tend to concentrate at few positions. The impact value has practically no correlation ($r = 0.0$, Fig. 2*e*) with the water accessibility of the residue. Notably, all residues with impact values >0.95 are $\geq 95\%$ buried. Fig. 2, *b* and *c*, show the solvent-accessible surface of chymotrypsin color-coded according to the impact value and the residue conservation. The difference in color patterns indicates the poor correlation between impact values and residue conservation (Fig. 2, *b* and *c*, see also *d*). Several residues on the surface are poorly conserved (Fig. 2*c*, *blue*), but are associated with high impact values (Fig. 2*b*, *white* or *red*). Invariant or quasi-invariant positions such as the catalytic His^{57} or Ser^{195} (Fig. 1) have impact values close to 1, as expected from Equations 2 and 3. Notably, residues with high impact tend to concentrate at the active site.

The distribution of impact values in Fig. 2 suggests that most of the information necessary to construct the phylogenetic tree in Fig. 1 is concentrated in a few sites of the protein. This prompted an investigation of the smallest set of residues capable of reproducing the phylogenetic tree in its entirety. Impact values were calculated for secondary structural elements of the protease domain. The impact value of individual β-strands ranged from 0.40 to 0.64 (Fig. 3, *b1* to *b12*). Helices are not conserved among serine proteases and were grouped for the analysis with loops connecting β-strands. These elements show impact values ranging from 0.41 to 0.67 (Fig. 3, *a1* to *a10*). In no case, however, could the tree in Fig. 1 be reconstructed using single secondary structural elements, either β-strands, helices, or loops.

Next we calculated the impact values of patches of accessible surface area ranging 1100–1300 Å^2 in size and plotted them *versus* the average conservation of residues (Fig. 3). More than 100 patches were scanned around each of the water-accessible residues on the surface of the protein. The area with maximum impact clearly stood out from the rest and involved the specificity sub-sites S1–S4. This area encompasses the smallest set of residues capable of reproducing the phylogenetic tree in Fig. 1 in its entirety. Remarkably, the resulting m_{ab} values obtained from the S1–S4 sub-sites give a better approximation of

Seikagaku Corporation Exhibit 2080 - Page 3 of 4
Charles River Laboratories, Inc. v. Seikagaku Corporation
IPR2025-00440

the phylogenetic tree in Fig. 1 than the entire sequence of the enzyme or any combination of secondary structural elements. We have recently shown that the 50 residues of the C-terminal domain of serine proteases are necessary and sufficient to reproduce the architecture of the parent tree in Fig. 1 (8). Twenty out of 30 residues defining the S1–S4 sub-sites are comprised within the 50 C-terminal residues. We conclude that substrate recognition at the S1–S4 sub-sites dominates the evolution and functional diversity of serine proteases.

The definition of the impact value in Equation 3 is a powerful tool to identify sets of residues specifying functional branching in a phylogenetic tree. The information derived from this approach should assist in the rational engineering of enzymes with altered specificity by pointing to regions that likely hold structural determinants of function. In addition, the profile of impact values in a newly discovered gene in a protein family reveals targets for mutagenesis from which the function of the protein may be identified.

REFERENCES

1. Gu, X. (1999) *Mol. Biol. Evol.* **16,** 1664–1674
2. Lichtarge, O., Bourne, H., and Cohen, F. E. (1996) *J. Mol. Biol.* **257,** 342–358
3. Armon, A., Graur, D., and Ben-Tal, N. (2001) *J. Mol. Biol.* **307,** 447–463
4. Landgraf, R., Xenarios, I., and Eisenberg, D. (2001) *J. Mol. Biol.* **307,** 1487–1502
5. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
6. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882
7. Felsenstein, J. (1994) Phylip package, Version 3.57c, University of Washington, Seattle, WA
8. Krem, M. M., Rose, T., and Di Cera, E. (1999) *J. Biol. Chem.* **274,** 28063–28066
9. Felsenstein, J. (1993) *Methods Enzymol.* **266,** 418–427
10. Fitch, W. M., and Margoliash, E. (1967) *Science* **155,** 279–284
11. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242
12. Sali, A., and Blundell, T. L. (1993) *J. Mol. Biol.* **234,** 779–815
13. Bode, W., Turk, D., and Karshikov, A. (1992) *Protein Sci.* **1,** 426–471
14. Matthews, I. I., Padmanabhan, K. P., Ganesh, V., Tulinsky, A., Ishii, N., Chen, J., Turck, C. W., Coughlin, S. R., and Fenton, J. W. (1994) *Biochemistry* **33,** 3266–3279
15. Richards, F. M. (1985) *Methods Enzymol.* **115,** 440–464