Three-dimensional Models of Proteases Involved in Patterning of the Drosophila Embryo

CRUCIAL ROLE OF PREDICTED CATION BINDING SITES

Thierry Rose‡, Ellen K. LeMosy§, Angelene M. Cantwell‡, Dolly Banerjee-Roy‡, James B. Skeath¶, and Enrico Di Cera‡§

From the ‡Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110, the §Department of Cellular Biology and Anatomy, Medical College of Georgia, Augusta, Georgia 30912, and the ¶Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110

Three-dimensional models of the catalytic domains of Nudel (Ndl), Gastrulation Defective (Gd), Snake (Snk), and Easter (Ea), and their complexes with substrate suggest a possible organization of the enzyme cascade controlling the dorsoventral fate of the fruit fly embryo. The models predict that Gd activates Snk, which in turn activates Ea. Gastrulation Defective can be activated either autoprolytically or by Ndl. The three-dimensional models of each enzyme-substrate complex in the cascade rationalize existing mutagenesis data and the associated phenotypes. The models also predict unanticipated features like a Ca\(^{2+}\) binding site in Ea and a Na\(^{+}\) binding site in Ndl and Gd. These binding sites are likely to play a crucial role in vivo as suggested by mutant enzymes introduced into embryos as mRNAs. The mutations in Gd that eliminate Na\(^{+}\) binding cause an apparent increase in activity, whereas mutations in Ea that abrogate Ca\(^{2+}\) binding result in complete loss of activity. A mutation in Ea predicted to introduce Na\(^{+}\) binding results in apparently increased activity with ventralization of the embryo, an effect not observed with wild-type Ea mRNA.

Several genes in the dorsal group (1) are involved in extra-cellular events that lead to dorsoventral polarization of the Drosophila melanogaster embryo. nudel, pipe, and windbeutel are expressed by somatic follicle cells during mid-oogenesis, whereas easter, gastrulation defective, snake, and spätzle are expressed by the nurse cells and oocyte. These genes were identified in several large scale genetic screens for maternal effect mutations that cause homozygous mutant females to produce embryos with abnormal cell fates (2). Among the dorsal group genes, nudel, gastrulation defective, snake, and easter encode proteins containing serine protease domains of the trypsin family. Easter (Ea), Snk, Gd, and Ndl are expressed and secreted during oogenesis as inactive zymogens into a thin, fluid-filled perivitelline space that lies between the eggshell and the oocyte. Genetic and molecular studies suggest that these proteins act in a proteolytic cascade many hours later in the early embryo (1, 3–5). The cascade resembles in its general organization those controlling the innate immune response and blood coagulation (6). Ovulation of the egg in some way triggers the self-activation of Ndl into Ndl\(^{\ast}\). Gd can be activated either by Ndl\(^{\ast}\) or by self-activation in the presence of Snk. Subsequently, Gd\(^{\ast}\) activates diffusible Snk and Snk\(^{\ast}\) activates diffusible Ea. The result of this cascade is cleavage by Ea\(^{\ast}\) of the diffusible dimeric nerve growth factor-like Spz (7). The processed Spz appears to function as a dimer to activate the transmembrane receptor Toll only on the embryo surface that will become ventralized through the Toll signaling pathway. In contrast to the significant knowledge garnered from previous in vivo studies, quantitative information on activity and specificity of various members of the cascade has so far eluded characterization involving purified proteins. Several questions remain regarding the activation of Ndl and Gd (1, 3–5) and the specificity of Gd\(^{\ast}\) and Snk\(^{\ast}\). Elucidation of these timely and important questions would benefit from the knowledge of the structural organization of the enzymes involved in the cascade. However, none of the members of the cascade has been crystallized so far or even expressed successfully for detailed in vitro characterization. Hence, we felt that the construction of three-dimensional models of Ndl\(^{\ast}\), Gd\(^{\ast}\), Snk\(^{\ast}\), and Ea\(^{\ast}\) in complex with their targets could fill a critical structure-function gap in the field as recently shown for thrombin interactions with the platelet receptors (8) and fibrinogen (9). The value of these models stems from their timeliness and the new insight offered for future mutagenesis studies, as illustrated in the present work by the effect on embryo polarization when putative cation binding sites of examined proteases were mutated.

MATERIALS AND METHODS

Sequence Alignment and Comparative Modeling—The fly sequences came from the strain Berkeley in the Flybase (FB) and Swiss Protein (SP) databases: Ea (FBgn0005053, SP-P13582), Gd (FBgn0006806, SP-O62589), Ndl (FBgn0002926, SP-P98159), and Snk (FBgn0003450, SP-P05049). These sequences were aligned with 1800 serine proteases from the non-redundant data base at the National Center for Biotechnology Information (NCBI) (National Library of Medicine, National Institutes of Health, Bethesda, MD) and

\[\text{sin; Ea}^{\ast}\text{, activated Easter; Gd, Gastrulation Defective zymogen; Gd}^{\ast}\text{, activated Gastrulation Defective; Ndl, Nudel zymogen; Ndl}^{\ast}\text{, activated Nudel; Snk, Snake zymogen; Snk}^{\ast}\text{, activated Snake; Spz, Spätzle; Thr, thrombin; Try, trypsin; PDB, Protein Data Bank; FB, Flybase; SP, Swiss Protein.}\]

Received for publication, November 20, 2002, and in revised form, December 18, 2002.

Published, JBC Papers in Press, December 18, 2002, DOI 10.1074/jbc.M211820200

This paper is available on line at http://www.jbc.org
Three-dimensional Models of Fruit Fly Proteases

11321

the Flybase image at the NCBI using trypsin homologues as seeds with the BLAST program and aligned together with ClustalX as described recently (10). Sequences were clustered into 100 groups from a neighbor junction tree accounting 500 bootstraps with ClustalX. One hundred sequences were selected, one per cluster. Three-dimensional models of 70 structures were built by comparative modeling based on 12 of 20 crystal structures of serine proteases used in the sequence core. These models were used to refine the alignment of the 100-sequence core (10). The theoretical three-dimensional models of activated protease domains Nd1* (central or Nd1*+1146–1385) and C-terminal or Nd2* (2017–2616); Gd* (256–528); Snk* (191–430); and Ea* (272–392) were constructed by comparative modeling using the program Modeller 4 (11). The following crystal structures of serine proteases downloaded from the Protein Data Bank (PDB) (12) were used as template structures: the closest hydrophobic side chain from P4 to P10 (seven possibilities). Five three-dimensional models of the complex were provided by Modeller runs, and then the best one was minimized as in the thorough mode.

The quick mode was used to screen possible cleavage sites all along sequence targets in and out of the catalytic domain. We used only one of seven peptide three-dimensional models to template the position for each of the eight residues of the loop between P1 and P2. The ligand, docked it on the best free protease three-dimensional model as a starting point for a new modeling process, and optimized the best complex as described above for the free proteases. The thorough mode screened 1 of 50 three-dimensional models of enzyme-target peptide complexes and allowed to screen every putative activation cleavage site of zymogens with every selected protease to assess activator-activated pairs.

The predicted free energy of association between receptor (R) and peptide (P), \( G_{\text{R-P}} \), was calculated considering that free R and P have the same conformation as in the complex RP from \( G_{\text{R-P}} = G_{\text{R}} - G_{\text{P}} \). The stereochemistry of the three-dimensional models was calculated from its enthalpic and entropic contributions expressed as \( G_{\text{gas}} = G_{\text{conf}} + G_{\text{rt}} + G_{\text{vib}} \). The enthalpy \( G_{\text{gas}} \) is a function of the van der Waals (\( E_{\text{vdw}} \)) and coulombic (\( E_{\text{coul}} \)) components, whereas \( G_{\text{conf}} \) is defined in terms of the rotational, configurational, and vibrational components. \( E_{\text{coul}} \) and \( E_{\text{vdw}} \) were computed from the CFF91 force-field without cut-off with \( e = 2 \). The value of the conformational entropy \( S_{\text{conf}} \) was computed from the loss of side chain angles.

Three-dimensional Models of Enzyme-Substrate Complex—Three-dimensional models of protease-substrate complexes were built by comparative modeling in a thorough or quick mode. In the thorough mode, the protease-fragment complexes were threaded over trypsin-peptide crystal structures (8, 9). We used the following templates: peptide-Ac-DFLAEYGGVR from PDB (19) (19); PPACK from PDB (19b); hiserin peptide-NG-DPDPHPVEE from PDB (19c); and DPDPHPVEE from PDB (19d). Five threading models were built and ranked in terms of stereochemistry quality and lowest potential binding energies. The accepted computer-generated models of protease-peptide substrate complexes had root mean square deviations of <1.5 Å for the protease backbone and peptide residues <10 Å from protease residues. Models containing a ligand with root mean square deviations of <2 Å from a higher ranked model were discarded. We selected the best ten models, extracted the ligand, docked it on the best free protease three-dimensional model as a starting point for a new modeling process, and optimized the best complex as described above for the free proteases. The thorough mode screened 1 of 50 three-dimensional models of enzyme-target peptide complexes and allowed to screen every putative activation cleavage site of zymogens with every selected protease to assess activator-activated pairs.

The quick mode was used to screen possible cleavage sites all along sequence targets in and out of the catalytic domain. We used only one of seven peptide three-dimensional models to template the position for each of the eight residues of the loop between P1 and P2. The ligand, docked it on the best free protease three-dimensional model as a starting point for a new modeling process, and optimized the best complex as described above for the free proteases. The thorough mode screened 1 of 50 three-dimensional models of enzyme-target peptide complexes and allowed to screen every putative activation cleavage site of zymogens with every selected protease to assess activator-activated pairs.

The quick mode was used to screen possible cleavage sites all along sequence targets in and out of the catalytic domain. We used only one of seven peptide three-dimensional models to template the position for each of the eight residues of the loop between P1 and P2. The ligand, docked it on the best free protease three-dimensional model as a starting point for a new modeling process, and optimized the best complex as described above for the free proteases. The thorough mode screened 1 of 50 three-dimensional models of enzyme-target peptide complexes and allowed to screen every putative activation cleavage site of zymogens with every selected protease to assess activator-activated pairs.
to 1.0, G(1.0,2.0), as $\Delta G = G(80,0.2,0) - G(1.0,2.0)$. The radius was fixed to 1.4 Å for solvent molecules and 2 Å for ions. Ionic strength was set at 145 mM, and the protonation state and partial charge distribution were assigned by the program Biopolymer according to the pH fixed at 7.0. The non-polar contribution $G_p$ was considered as linearly dependent on the molecule solvent-accessible surface area using a surface tension coefficient of 25 cal/mol Å² (20), i.e. $\Delta G_p = 25 \times \Delta SASA$.

Based on the above definitions, the free energy for the receptor-peptide complex becomes $\Delta G = \Delta H + T \Delta S + T \Delta S_{gas} - T \Delta S_{conf} + \Delta G_{prot}$

Some of the terms cancel if we compare the association of same length peptides bound to the same protease. The approximation of the relative binding free energy is given by $\Delta G = \Delta H_{prot} - T \Delta S_{conf} + \Delta G_{prot}$.

This approach does not allow comparison of the binding of a peptide to two different proteases unless the vibrational entropy variation upon binding is comparable.

The $\Delta G$ values refer to selected conformations and are affected by the choice of the “best model” according to global potential energy of the system and the goodness of its stereochemistry. The mean ± S.D. is 2.4 kcal/mol for the $\Delta G$ of the 10 best models of Snk* when it is bound to the activation site of Ea. Lower deviations were estimated as 1.2 kcal/mol for Ea* with Snk peptide, 1.7 kcal/mol for Ndl* with Gd peptide, and 1.4 kcal/mol for Gd* with Snk peptide.

**mRNA Preparation of Mutated gd and ea**—The plasmid pNB-GD2 containing a full-length gd cDNA was obtained from J. L. Marsh (University of California, Irvine, CA) (21). The plasmid pGEM7Zf(+) containing a 1534-bp fragment was constructed by inserting a BamHI fragment derived from p3EGFP-1 into the BamHI site of pGEM7Zf(+) (22). Mutations were introduced using the Quick Change kit (Stratagene). We mutated Phe-225 to Ile, Ser, and Tyr to create the putative Na$^+$-dependent cleavage site $\text{GDGR} – \text{VA, score 0.85, threshold 0.48}$; Ndl* 48 $\text{IVGG}$; chymotrypsin-like cleavage; and Ea* 22–127 $\downarrow 128–392$ (LSNR $\downarrow$ IVYG; trypsin-like cleavage) (Fig. 1). The predicted underivatized Ea* A chain (106 amino acids, theoretical mass 12,086 Da), Ea* B chain (285 amino acids, theoretical mass 28,851 Da), Snk* A chain (156 amino acids, theoretical mass 17,372 Da), and Snk* B chain (247 amino acids, theoretical mass 27,319 Da) agree with Western blots described by Dissig et al. (5). In the case of Gd*, no basic or hydrophobic residue occupies the canonical position, and the closest putative cleavage site is either 30 or 22 residues upstream, specifically Gd* 20–211 $\downarrow 212–528$ (GEPK $\downarrow$ SSDG; trypsin-like cleavage) or Gd* 20–220 $\downarrow 221–528$ (TPSV $\downarrow$ FVDD; chymotrypsin-like cleavage). The fragment 212–528 expressed in S2 insect cells is an active protease (3). DeLotto (29) proposed the cleavage site Gd* 20–136 $\downarrow 137–528$ (EHIR$\downarrow$ KLSF; trypsin-like cleavage) located 83 residues upstream of the canonical activation site. The proposed cleavage site is 12 residues upstream of a type A von Willebrand repeat motif (LDDLXX–EXXXRXXD) as described for complement factor B and C2. With such a cleavage, the predicted underivatized Gd* A chain (116 amino acids, theoretical mass 13,396 Da) and Gd* B chain (390 amino acids, theoretical mass 43579 Da) agree with the Western blots described by Dissig et al. (5). The alignment of Spz with the protease activation sites proposes 26–220 $\downarrow 221–326$ (VSSR $\downarrow$ VGGS; trypsin-like cleavage) as the best site of cleavage to produce the fragments documented by SDS-polyacrylamide gel electrophoresis (5). Ndl* could also be purchased to release only the Z-domain domain Ndl1* (241 amino acids, theoretical mass 26,862 Da) by a second cleavage 1145–1385 $\downarrow 1386–2616$ (TTTP $\downarrow$ LLPK; trypsin-like cleavage) as shown by LeMoys et al. (30). The cleavage site 1386–2016 2017–2616 (NLMR $\downarrow$ LLNV; trypsin-like cleavage) is also detectable in the C-terminal domain Ndl2 (600 amino acids).

Ea and Snk show 25% identity overall, 33% within the B chain, and feature the same potential disulfide bridges (Fig. 1). Alignment with other proteases suggests that only one disulfide bridge, 1–122 in the chymotrypsin numbering, 2 links the A and B chains. The disulfide bonds 42–58, 168–182, and 191–220 are highly conserved in the catalytic B chain of serine proteases. Ndl* is proposed to retain the disulfide bonds 42–58 and 168–182 as well as 1–122 linking the A and B chains (Fig. 1). Ndl could have five disulfide bonds in Ndl1* (1–122 between the A and B chains and 42–58, 136–201, 168–182, and 191–220 within the B chain). Ndl2* features only the 42–58 disulfide bond within the B chain and 1–122 between the A and B chains (Fig. 1).

**RESULTS**

**Identification of Cleavage Sites by Alignment of Primary Sequences**—Zymogens Gd (amino acid 528), Snk (amino acid 430), and Ea (amino acid 392) are organized in three domains: an N-terminal signal that is cleaved during protein secretion and a zymogen that gives rise to A (N-terminal) and catalytic B (C-terminal) chains (Fig. 1). The A and B chains remain covalently linked through disulfide bridges after proteolytic activation. The topology of Ndl is more complex and unusual because it carries two S1a protease domains. The first catalytic domain (Ndl1*—(1145–1385)) is central, and the second (Ndl2*—(2017–2616)) is C-terminal. Eleven low density lipoprotein (LDL) receptor-binding repeats intercalate the two protease domains (27). Four LDL receptor repeats are inserted in the second protease catalytic domain.

To locate cleavage positions ( ) of signal peptides, we used the program SignalP (28). This yielded the following sites of cleavage: Ndl 1–47 $\downarrow$ 48–2616 (VYH $\downarrow$ GL, score 0.54, threshold 0.48); Gd 1–19 $\downarrow$ 20–528 (TKA $\downarrow$ VA, score 0.85, threshold 0.48); Snk 1–27 $\downarrow$ 28–430 (LEA $\downarrow$ LD, score 0.75, threshold 0.48); Ea 1–21 $\downarrow$ 22–329 (SAG $\downarrow$ QF, score 0.82, threshold 0.48), and Spz 1–25 $\downarrow$ 26–326 (YEA $\downarrow$ KE, score 0.93, threshold 0.48). Alignment of Ea, Snk, Gd, and Ndl with other serine proteases suggests the following cleavage sites for zymogen activation: Ndl1* 48–1144 $\downarrow$ 1145–2616 (GDGR $\downarrow$ IVVG; trypsin-like cleavage); Snk* 28–183 $\downarrow$ 184–430 (SVPL $\downarrow$ IVVG; chymotrypsin-like cleavage); and Ea* 22–127 $\downarrow$ 128–392 (LSNR $\downarrow$ IVYG; trypsin-like cleavage) (Fig. 1).
the cascade as an active protease. The LDL domain is inserted away from the potential active site in the 186-loop, where insertions of various length also exists in thrombin and tissue-plasminogen activator.

Fig. 2 displays the water-accessible surfaces of Ndl1*, Gd*, Snk*, and Ea*. The overall architecture of the active site is similar in all models, but their surfaces show notable differences in amino acid composition. The four proteases feature the catalytic triad His-57, Asp-102, Ser-195, and the important ancillary residues Cys-42 and Cys-58 (SS-linked), Gly-193, Gly-196, Gly-211, and Ser-214. The Cys-168/Cys-182 disulfide bond stabilizes the intervening loop that forms part of the binding site and is conserved in all four proteases. The Cys-191/Cys-220 disulfide bond is present in Ndl1*, Snk*, and Ea* but not in Gd*. This bond bridges the 186-loop and 220-loop that shape the bottom of the primary specificity pocket. Binding site pockets around residue 189 and the hydrophobic core around residues 99, 174, and 215 differ among the four proteases.

The presence of Asp-189 in the S1 (31) pocket shows that specificity is unambiguously trypsin-like for Ndl1* and Ea*.
FIG. 2. Three-dimensional homology models of Ndl1* (A), Gd* (B), Snk* (C), and Ea* (D). Enzyme residues are numbered according to chymotrypsin for ease of comparison. The models are shown as ribbons on the left side or as solvent-accessible surface areas on the right side, color-coded according to amino acid properties (Asp and Glu in red; Lys and Arg in blue; His in purple; Ala, Ile, Leu, Met, and Val in yellow; Phe, Trp, and Tyr in green; Asn, Cys, Gln, Gly, Pro, Ser, and Thr in white). The Na+ (orange ball) and Ca2+ (purple ball) binding sites are detailed in the insets. Also shown are main residue contacts between primary targets (listed vertically in black as 25-residue peptides) and enzymes (with individual residues color-coded). Cleavage sites are indicated by scissors.
Three-dimensional Models of Fruit Fly Proteases

Enzymes are listed on the top row, and target proteases are listed in the left column. Cleavage sites are reported within the central eight residues of each peptide from the putative activation region of corresponding zymogens. For each enzyme (column), one complex was used as reference ($\Delta\Delta G = 0.0$) to facilitate comparison with other enzyme-target complexes.

| Enzymes $\rightarrow$ substrates | Nd1* | Gd* | Snk* | Ea* | Thr | Try | Chy |
|---------------------------------|------|-----|------|-----|-----|-----|-----|
| Gd                             | 2.0  | 3.9 | -1.0 | +4.1| -1.3| -2.4| +1.5|
| Gd                             | +0.9 | +6.9| +0.5 | -0.6| -0.2| -1.2| +0.8|
| Gd                             | 0.0  | 0.0 | 0.0  | 0.0 | 0.0 | 0.0 | 0.0 |
| Snk                            | 2.1  | -0.5| +0.2 | +1.2| 2.0 | 0.0 | 2.6 |
| Snk                            | +3.8 | -7.6| +2.2 | +1.4| +2.0| 0.0 | -0.8|
| Snk                            | 0.0  | +5.9| -5.9 | +4.7| +0.6| -0.4| +3.7|
| Snk                            | +0.8 | +3.9| +0.9 | -4.8| 0.0 | -1.2| +0.3|

Whereas Ser-189 suggests a chymotrypsin-like specificity for Gd*. The presence of Gly-189 in Snk* makes the prediction of specificity ambiguous. The shape and volume of the S1 pocket in Snk* could accommodate a variety of side chains. Leukocyte elastase, which carries Gly189 and cleaves after Val in P1, shows a 23% identity with Snk* in the catalytic B chain. The structure of elastase (PDB code 1ppg) complexed with the tetrapeptide AAPV (32) shows that Val-190 defines the S1 specificity toward hydrophilic P1 residues. In the model of Snk*, the unusual His-190 (His-371) points out of the S1 pocket and interacts with Asp-194 (Asp-375), thereby leaving the S1 pocket free to interact with a variety of side chains besides hydrophobic residues.

Preferred Cleavage Sites from Enzyme-Substrate Three-dimensional Models—We predicted the position of potential protease targets in every protease sequence based on 25-residue peptide binding energy to each protease active site (Table I) or peptide-target specificity (Table I). The best cleavage sites for Gd* are Arg-1414 in Nd1 and Lys-211 in Gd. The best cleavage site for Gd* is Leu-183 in Snk. The best cleavage site for Snk* is Arg-127 in Ea. The best cleavage site for Ea* is Arg-220 in Spz (Fig. 1, Table I). Interestingly, the other fragments are predicted to bind with high scores and can be regarded as secondary target sites. Fig. 1 reports potential cleavage sites within catalytic domains sorted by trypsin-like cleavage sites (blue, Arg or Lys in P1) or chymotrypsin-like cleavage sites (pink, Leu, Val, Ile, and Phe in P1) when the accessibility score of the corresponding site is $>0.5$. For example, Nd1* can cleave Nd1 at Arg-1385 and can separate Nd12 from Nd1. Nd1* can also cleave at Arg-1094, yielding a fragment that may correspond to the non-diffusible 38-kDa fragment reported by LeMosy et al. (30). Ea* and Snk* are predicted to cleave the prodomain of Gd at Arg-187 and can generate the 50-kDa fragment described by LeMosy et al. (3). Furthermore, Ea* can cleave the prodomain of Snk at Arg-100, thereby producing the 50-kDa fragment observed when Ea* and Snk are coexpressed (3).

The four proteases in complex with their primary targets were examined further using the thorough mode. Relevant contacts of the best targets with each enzyme are shown in Fig. 2, and the structure of the peptide and the epitope of recognition are displayed in Fig. 3. The Nd1 fragment $^{1184}$SDKIEV-GDGR $^{1185}$IVGGSHTSALQWP $^{1186}$ and the two Gd fragments $^{230}$ESLHYAIGEKP $^{231}$KLSFITALPFFVVDDE $^{235}$ (cleavage 30 residues upstream of the standard activation site) and $^{526}$FMTGIRKLSFIPDKKSSLLL $^{531}$ (C2/factor B-type cleavage site 83 residues upstream of the standard activation site) were docked on the active site of Nd1*. Several single mutations have been made in nd1, and their associated phenotypes have been reported previously (33). We introduced these mutations in the three-dimensional model of Nd1* and optimized the structure by 500 cycle-conjugated gradient minimization on the residues within 6 Å from the site of mutation. Mutations are displayed in Fig. 3 where residues are visible at the protein surface in its front view. The mutant C1114S (C1S) loses the disulfide bond with Cys-1252 (Cys-122) that connects the A and B chains, but it is processed and secreted normally and retains partial activity. The A chain is usually inconsequential to function in serine proteases (34). The mutants G1280S (G140S) and G1282R (G142R) affect the position of the highly conserved Trp-1281 (Trp-141) that lines part of the S1 specificity site. Either mutation may change the backbone and side chain orientation of Trp-141 with resulting poor substrate binding and loss of protease activity as seen experimentally (33). The same argument holds for the mutant V1275M (V134M) that destabilizes the scissile bond, thereby explaining the loss of activity seen experimentally (33). The mutant H1355L (H215L) perturbs the hydrophobic core next to the active site (Fig. 3A). The His residue at this position replaces the highly conserved Trp seen in almost all of the serine proteases. The hydrophobic residue Ile-207 of Gd or Val-1140 of Nd1 can make contacts with Leu-1355 (Leu-215). However, non-conservative replacements of residue 215 in thrombin cause a drastic drop in activity (38).
Three-dimensional Models of Fruit Fly Proteases

Mutated residues described in the text are displayed as sticks (domains are shown as solvent-accessible surface areas). Peptide substrates are not visible on the particular displayed view or because they are not accessible to solvent.

The Gd* active site is characterized by very hydrophobic properties of both primed and unprimed subsites (Fig. 2B). Residue Ile-468 (Ile-194) replaces the canonical Asp in the S1 pocket and contributes to the enhanced hydrophobicity of this site together with Ile-463 (Ile-190) and the unusual Ile-511 (Ile-226) that replaces a highly conserved Gly. This largely hydrophobic architecture of the S1 pocket is unusual in serine proteases and probably compensates for the unusual Ala-488 (Ala-215) that replaces the highly conserved Trp at this position. Residue Val-220 of Gd is potentially a good cleavage site for Gd*. To test the possibility of Gd activation by Gd*, the Gd fragment \( ^{210}PKSSDGTSPV \) \( ^{219}FDVDDDEDVLKH-QF^{234} \) was docked onto the active site of Gd*. A potential activation site with chymotrypsin-like specificity is \( ^{128}TQIQ-LEHIRKL \) \( ^{132}SFIPDKKSSLILDP \) located near the C2-factor B-type cleavage site (Fig. 1). Gd* carries two insertions in the 60- and 149-loops relative to chymotrypsin, a feature also observed in thrombin (34) where the insertions contribute to the narrow substrate specificity. The 60-loop in Gd* covers the substrate residues at P1 and P1'. The 149-loop is quite flexible, judging from the various conformations obtained in the fifty best models, and interacts loosely with substrate residues at P3'–P8' (Fig. 3B). Mutant alleles of gd have been identified and grouped in three complementation groups (37).

Mutations in the catalytic domain G466D (G193D), G469E (G196D), and G484D (G211D) are all in the same group complementing with the group with mutations in the propeptide domain. G466D places the acidic side chain at the bottom of the S4' site but does not disturb the aromatic cluster formed by Phe-275, Trp-388, and Phe-390. The hypomorphic effect of this mutation (gd\(^6\)) is moderate (37). On the other hand, G469E introduces a charged side chain into a hydrophobic cluster formed by Ile-265 (Ile-33), Phe-294 (Phe-59), Val-306 (Val-64), Val-328 (Val-90), and Ile-331 (Ile-88). The unfavorable steric hindrance is partially compensated by backbone torsions in the structure core in the vicinity of the catalytic Ser-468 (Ser-194) and His-292 (His-57). A similar effect is observed for the mutant G484D where the charged side chain perturbs the hydrophobic cluster formed by Tyr-377 (Tyr-130), Leu-472 (Leu-201), Phe-445 (Phe-181), Tyr-513 (Tyr-228), and Ala-514 (Ala-229) with resulting rearrangement of the backbone structure of the S3–S4 sites and of the 186-loop and 220-loop shaping the Na\(^+\) binding site. G469E (gd\(^{10}\) allele) and G484D (gd\(^{7}\) allele) dramatically compromise the activity of Gd and yield completely dorsIALIZED embryos.

The Snk fragment \( ^{177}SGKQCVPVSPL \) \( ^{182}IVGGPTRHG-LFPF197 \) was docked onto the active site of Gd* with the P1 residue Leu-183 into the chymotrypsin-like S1 pocket in contact with Ser-468 (Ser-189), Ile-463 (Ile-190), and Ile-467 (Ile-194) (contacts detailed in Fig. 2B and structure detailed in Fig. 3B). The small side chain of Ala-488 (Ala-215) in Gd* opens a cavity bordered by Leu-342 (Leu-97), and Tyr-344 (Tyr-99) that accommodates the P6 residue Val-178 of Snk. Of the other Snk residues, Val-181 at P3 contacts Ala-489 (A216) and Leu-490 (Leu-217), and Phe-195 at P11 stacks favorably against Trp-388 (Trp-141), Phe-390 (Phe-143), and Phe-275 (Phe-34), whereas Leu-194 makes close contacts with Leu-401 (L149d) and Phe-390 (Phe-143).

The Ea fragment \( ^{116}LPGQCNLSRN \) \( ^{121}IVGGMKTIDE-FPW144 \) was docked on Snk* (contacts detailed in Fig. 2C and structure detailed in Fig. 3C). Snk* carries Gly-370 (Gly-189) in the S1 pocket, consistent with either trypsin or chymotrypsin activity. Residue His-371 (His-190) makes the S1 pocket more prone to interact with hydrophilic rather than hydrophobic side chains. Ea residue Arg-127 fills the S1 pocket of Snk*.
making two H-bonds with the backbone carbonyls of His-371 (His-190) and Phe-402 (Phe-218).

The Spz fragment 5'NDLQPTDVSSR ↓ VGGSERFCLR-GRSIR 6' (Fbgm0003495, SP-48607) was docked onto the Ea* active site (contacts detailed in Fig. 2D and structure detailed in Fig. 3D) with Arg-220 at P1 bound to Asp-332 (Asp-189). Several naturally occurring mutations of Ea* have been identified that lead to dominant or recessive phenotypes of dorsoventral differentiation (24). Dominant alleles are A325V (A183V), P373S (P225S), R335C (R192C), G336S (G193S), G371R (G223R), G283S (G142S), V360 M (V213M), and G131E (A183V), P373S (P225S), R335C (R192C), G336S (G193S), G371R (G223R), G283S (G142S), V360 M (V213M), and G131E (A183V), P373S (P225S), R335C (R192C), G336S (G193S), G371R (G223R), G283S (G142S), V360 M (V213M), and G131E (A183V), P373S (P225S), R335C (R192C), G336S (G193S), G371R (G223R), G283S (G142S), V360 M (V213M), and G131E (A183V) or Lys at P1. Other Ea* mutants that are expected to complicate substrate binding are G131E (G19E) that places an acidic side chain in a hydrophobic environment, G339R (G196R) and G363E (G216E) that occlude the S1 pocket, C324Y (C182Y) that removes the disulfide bond and stability of the hydrophobic core next to the active site.

Putative Cation Binding Sites and Their Alteration in Vivo—Many vertebrate serine proteases contain functional cation binding sites that allosterically regulate activity and stability of the enzymes (39, 40), but such sites have not previously been described in invertebrate serine proteases. The inspection of the primary sequence and screening of the dorsoventral protease three-dimensional models with the program VALE (16) identified binding sites for Na\(^{+}\) in Ndl* and Gd* and for Ca\(^{2+}\) in Ea*, each corresponding to the positions of similar sites in the vertebrate proteases. The Na\(^{+}\) binding sites of Ndl* (Fig. 2A) and Gd* (Fig. 2B) have an architecture similar to that described for thrombin (36, 38). Two carboxyl O atoms from residues 221 and 224 contribute together with four buried water molecules to the octahedral coordination of the cation: Arg-1361 (Arg-221) and Glu-1364 (Glu-224) for Ndl*; Cys-506 (Cys-221); and Glu-509 (Gln-224) for Gd*. The Ca\(^{2+}\) binding site of Ea* (Fig. 2D) is similar to that of trypsin (40) with two carboxylic side chains from Glu-193 (Glu-70) and Glu-203 (Glu-80), contributing to the octahedral coordination. In Ea*, three additional carbonyl oxygens from the backbones of Thr-196 (Thr-73) and Thr-198 (Thr-75) and the side chain of Asn-199 (Asn-76) contribute to Ca\(^{2+}\) binding. A water molecule could provide the sixth oxygen in the coordination shell.

To determine whether these putative cation binding sites influence protease function in vivo, we mutagenized key residues in Gd and in Ea and then compared the ability of wild-type and mutant proteases to rescue embryos lacking maternal function for the respective proteins (Table II) (Figs. 4 and 5). In previous studies using the same wild-type mRNAs and recipient embryos, Gd has been shown to act in a dose-dependent manner to cause an abnormal expansion of ventral pattern elements ("ventralization"), whereas Ea rescues to wild type but cannot ventralize the embryo (4, 24). Similar studies could not be undertaken for the putative Na\(^{+}\) binding site in Ndl, as wild-type Ndl is unable to rescue in embryo RNA microinjection assays, presumably because of the complex activation mechanism and early action of this protease.\(^3\)

For Gd, we injected synthetic mRNAs encoding wild-type Gd protein or mutations Y510A (Y225A) and Y510P (Y225P), expecting to disrupt Na\(^{+}\) binding (38, 39). At high doses (0.6 mg/ml), all three RNAs gave a strong ventralization phenotype in which excess Gd activity causes too much signaling through the Toll pathway (4), indicating that the mutant proteins are active. At a 10-fold lower dose (0.06 mg/ml), the wild-type RNA provides a broad range of phenotypes from

---

\(^3\) E. LeMosy, unpublished data.
strong ventralization to moderate dorsalization (partial rescue in Table II) while the mutant RNAs still show predominantly strongly and moderately ventralized embryos. This finding suggests that mutants predicted to lack Na\(^+\) binding have increased catalytic activity compared with wild type, implying that Na\(^+\) binding to Gd\(^+\) may actually result in decreased catalytic activity.

For Ea, we compared the activity of wild type and two mutants of Glu-193 (Glu-70): E193A (E70A) predicted to abrogate Ca\(^{2+}\) binding, and E193K (E70K) predicted to eliminate Ca\(^{2+}\) binding but with the charged Lys partially substituting for Ca\(^{2+}\) (41, 42). These mutants resulted in a complete loss of Ea activity equivalent to the injection of the S337A (S195A) mutant lacking the catalytic serine, indicating that the Glu-193 (Glu-70) residue and possibly Ca\(^{2+}\) binding are critical for Ea function.

**Engineering a Na\(^+\) Binding Site in Ea**—Most invertebrate proteases contain Pro-225, which is incompatible with Na\(^+\) binding rather than Tyr-225 or Phe-225, which are compatible with such binding (39). This usage dichotomy at residue 225 has profound structural (38, 39) and evolutionary (6) implications. One exciting possibility raised by the role of residue 225 in serine proteases (39) is the rational engineering of Na\(^+\) binding with the P225S substitution. We surmised that Ea would be an excellent candidate to engineer a Na\(^+\) site, because it had already been shown that the P373S (P225S) substitution by an EMS mutation resulted in in-
creased activity and ventralized phenotype (24). Ser is an intermediate in the genetic code between Pro and Tyr, and a saturation mutagenesis study has shown that Ser is also intermediate in the genetic code between Pro and Tyr, and a saturation mutagenesis study has shown that Ser is also an important feature that can be exploited in future studies. Ndl1* plays a key role in the function of this enzyme in vivo and offers a structural explanation of the observed mutant phenotypes.

Notably, the three-dimensional models reveal new structural features that can be exploited in future in vivo studies. Of particular importance is the unanticipated identification of a Ca2+ binding site in Ea* and a Na+ binding site in Gd* and Ndl1*. The binding of Ca2+ in trypsin stabilizes the fold of the protease domain (40), and Na+ binding to thrombin and many other serine proteases increases the catalytic activity toward synthetic and natural substrates (36, 38, 39). Based on the results presented here, it is highly likely that Ca2+ binding to Ea* plays a key role in the function of this enzyme in vivo. Likewise, Na+ binding to Gd* and possibly Ndl1* has functional significance. Interestingly, Na+ binding to Gd* may actually result in the inhibition of the catalytic activity of the enzyme in contrast to the effect observed in all other Na+-dependent allosteric serine proteases studied to date (39). The current knowledge on the role of residue 225 in serine proteases (39) predicts that Na+ binding can be introduced in proteases carrying Pro→Tyr replacement. However, the P225Y substitution in tissue plasminogen activator is not sufficient to introduce Na+ binding and actually results in reduced catalytic activity (44). The introduction of Na+ binding in this protease requires substitution of a large number of residues in addition to Pro→Tyr. Therefore, it is remarkable that the P225Y substitution in Ea* has such a profound effect on its catalytic activity, consistent with a gain of function that likely results from Na+ binding. This observation motivates the analysis of this protease in terms of kinetic and direct structural studies and offers new and important insights into ongoing efforts to engineer Na+ binding and enhanced catalytic activity in serine proteases of medical and biomedical relevance.

Acknowledgments—We thank Dr. Carl Hashimoto for offering decisive guidance at early stages of this project. E. K. LeMosy is grateful to Lora LeMosy for computing assistance and to Fu-Shin Yu and Ke-Ping Xu for the use of their phase-contrast microcamera in Fig. 5.

DISCUSSION

The primary cleavage sites of Ndl, Snk, Ea, and Spz have been proposed previously (3, 5, 7, 30) and are confirmed in the present study. Ndl is secreted in the perivitelline space and is required for the ventralization process upstream of Snk activation (30). The activation of Gd remains controversial, but our models propose that Ndl1* has trypsin-like activity and may bind Gd at its activation site Lys-211 better than Snk, Ea, or Spz. We believe that the second protease domain of Ndl*, Ndl2*, is inactive and plays no role in the cascade. Therefore, although Gd can be activated at Val-220 by Gd*, Ndl1* should be retained as a better potential actor in Gd activation from predicted binding energies (3). Gd autoactivation might be detected when Gd is overexpressed either in embryos (4) or in cell culture (3, 5) as suggested by the predicted low affinity of Gd for its own activation site, but this leaky autoactivation might not be as effective at physiologic expression levels of Gd. Hence, the proposed three-dimensional models of Ndl1*, Gd*, Snk*, and Ea* are consistent with the overall organization of the enzymatic cascade defining dorsoventral polarity in the fruit fly as recently described from cell culture and embryo studies (3, 5). The cascade is initiated by Gd activation, more probably by Ndl1* as suggested in vivo (4), and alternatively more weakly by Gd or Gd* as also proposed in vivo previously (5). Gd* then activates Snk and Snk* activates Ea. Ea* then processes Spz for signaling via the receptor Toll. The three-dimensional models are also consistent with previous mutagenesis studies of Ndl1* (33) and Ea* (24) and offer a structural explanation of the observed mutant phenotypes.

REFERENCES

1. LeMosy, E. K., Hong, C. C., and Hashimoto, C. (1999) Trends Cell Biol. 9, 102–107
2. Anderson, K. V., and Nusslein-Volhard, C. (1984) Nature 311, 223–227
3. LeMosy, E. K., Tan, Y. Q., and Hashimoto, C. (2001) Proc. Natl. Acad. Sci. U. S. A. 98, 5055–5059
4. Han, J. H., Lee, S. H., Tan, Y. Q., LeMosy, E. K., and Hashimoto, C. (2000) Proc. Natl. Acad. Sci. U. S. A. 97, 8098–8097
5. Dussing, M., Giordano, H., and DeLotto, R. (2001) EMBO J. 20, 2387–2393
6. Krem, M. M., and Di Cera, E. (2002) Trends Biochem. Sci. 27, 67–74
7. DeLotto, Y., and DeLotto, R. (1998) Biochem. J. 324, 141–148
8. Ayala, Y. M., Cantwell, A. M., Rose, T., Bush, L. A., Arusin, D., and Di Cera, E. (2001) Proteins 45, 107–116
9. Rose, T., and Di Cera, E. (2002) J. Biol. Chem. 277, 18875–18880
10. Rose, T., and Di Cera, E. (2002) J. Biol. Chem. 277, 19243–19246
11. Sali, A., and Blundell, T. L. (1995) J. Mol. Biol. 243, 779–815
12. Berman, H. M., Westbrook, J., Feng, Z., Gililand, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) Nucleic Acids Res. 28, 235–242
13. Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993) J. Mol. Biol. 231, 1049–1067
14. Hoof, R. W. W., Friend, G., Sander, C., and Abola, E. E. (1996) Nature 381, 272–274
15. Luty, R., Bowie, J. U., and Eisenberg, D. (1992) Nature 356, 83–85
16. Nayal, M., and Di Cera, E. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 817–821
17. Pickett, S. D., and Sternberg, P. W. (1995) J. Mol. Biol. 251, 825–839
18. Finkelstein, A. V., and Janin, J. (1989) Protein Eng. 3, 1–3
19. Nicholls, A., Sharp, K. A., and Honig B. (1991) Proteins 11, 281–296
20. Sharp, K. A. (1998) Proteins 33, 39–45
21. Konrad, K. D., Goralski, T. J., Mahowald, A. P., and Marsh, J. L. (1998) Proc. Natl. Acad. Sci. U. S. A. 95, 6819–6824
22. Chasan, R., and Anderson, K. V. (1999) Cell 96, 291–400
23. Konrad, K. D., Goralski, T. J., and Mahowald, A. P. (1988) Dev. Biol. 127, 133–142
24. Jin, Y. S., and Anderson, K. V. (1998) Cell 90, 837–881
25. Konrad, K. V., Jurgen, G., and Nusslein-Volhard, C. (1985) Cell 42, 779–789
26. Jurgen, G., and Nusslein-Volhard, C. (1986) in Drosophilae: A Practical Approach (Roberts, D. B., ed.) pp. 199–227, IRL Press at Oxford University Press, Oxford, United Kingdom
27. Willnow, T. E.,Orth, K., and Herz, J. (1994) J. Biol. Chem. 269, 15827–15832
28. Nielsen, H., Rognliebre, J., Brunak, S., and von Heijne, G. (1997) Int. J. Neural. Syst. 8, 581–599
29. DeLotto, R. (2001) EMBO Rep. 2, 721–726
30. LeMosy, E. K., Kemler, D., and Hashimoto, C. (1998) Development 125, 4843–4853
31. Schechter, I., and Berger, A. (1967) Biochim. Biophys. Res. Commun. 27, 157–162
32. Wei, A. Z., May, L., and Bode, W. (1988) FEBS Lett. 234, 367–373
33. LeMosy, E. K., Leclerc, C. L., and Hashimoto, C. (2000) Genetics 154, 247–257
34. Bode, W., Turk, D., and Karshikov, A. (1992) Protein Sci. 1, 426–471
35. Arosio, D., Ayala, Y. M., and Di Cera, E. (1999) Biochemistry 38, 8065–8101
36. Di Cera, E., Guinto, E. R., Vindigni, A., Dash, Q. D., Ayala, Y. M., Wuyi, M., and Tulinsky, A. (1995) J. Biol. Chem. 270, 22098–22092
37. Ponsonareff, G., Giordano, H., DeLotto, Y., and DeLotto, R. (2001) Genetics 157, 1485–1496
Three-dimensional Models of Fruit Fly Proteases

38. Guinto, E. R., Caccia, S., Rose, T., Futterer, K., Waksman, G., and Di Cera, E. (1999) Proc. Natl. Acad. Sci. U. S. A. 96, 1852–1857
39. Dang, Q. D., and Di Cera, E. (1996) Proc. Natl. Acad. Sci. U. S. A. 93, 10653–10656
40. Bode, W., and Schwager, P. (1975) J. Mol. Biol. 98, 693–717
41. Mathur, A., Zhong, D., Sabharwal, A. K., Smith, K. J., and Bajaj, S. P. (1997) J. Biol. Chem. 272, 23418–23426
42. Lai, T.-S., Slaughter, T. F., Peoples, K. A., and Greenberg, C. S. (1999) J. Biol. Chem. 274, 24953–24958
43. Chang, A. J., and Morisato, D. (2002) Development 129, 5635–5645
44. Vindigni, A., and Di Cera, E. (1998) Protein Sci. 7, 1728–1737