Networks of Coevolving Sites in Structural and Functional Domains of Serpin Proteins

Michael J. Buck[1] *and William R. Atchley*

Department of Genetics and The Center for Computational Biology, North Carolina State University

Amino acids do not occur randomly in proteins; rather, their occurrence at any given site is strongly influenced by the amino acid composition at other sites, the structural and functional aspects of the region of the protein in which they occur, and the evolutionary history of the protein. The goal of our research study is to identify networks of coevolving sites within the serpin proteins (serine protease inhibitors) and classify them as being caused by structural-functional constraints or by evolutionary history. To address this, a matrix of pairwise normalized mutual information (NMI) values was computed among amino acid sites for the serpin proteins. The NMI matrix was partitioned into orthogonal patterns of amino acid variability by factor analysis. Each common factor pattern was interpreted as having phylogenetic and/or structural-functional explanations. In addition, we used a bootstrap factor analysis technique to limit the effects of phylogenetic history on our factor patterns. Our results show an extensive network of correlations among amino acid sites in key functional regions (reactive center loop, shutter, and breach). Additionally, we have discovered long-range coevolution for packed amino acids within the serpin protein core. Lastly, we have discovered a group of serpin sites which coevolve in the hydrophobic core region (s5B and s4B) and appear to represent sites important for formation of the ''native'' instead of the ''latent'' serpin structure. This research provides a better understanding on how protein structure evolves; in particular, it elucidates the selective forces creating coevolution among protein sites.

Introduction

The evolution of a structure results from unobservable selective forces which drive amino acid substitutions at single or multiple sites. These selective forces leave traces discernable as covariation between sites. There has been a great deal of interest in understanding the significance of covariation among amino acid sites and its role in protein evolution and function (Chelvanayagam et al. 1997; Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Atchley et al. 2000; Tuff and Darlu 2000; Afonnikov, Oshchepkov, and Kolchanov 2001; Pritchard et al. 2001; Wang and Pollock 2005). Atchley et al. (2000) proposed that the covariance between amino acid sites i and j, C_{ij}, can be decomposed into its underlying components as

$$C_{ij} = C_{\text{phylogeny}} + C_{\text{structure}} + C_{\text{function}} + C_{\text{interactions}} + C_{\text{stochastic}}. \quad (1)$$

Specifically, $C_{\text{structure}}$ and C_{function} are the covariations due to selective forces maintaining a particular structure or functional domain. For example, if a large amino acid (leucine) within the protein core is substituted with a smaller amino acid (alanine), this change may destabilize the structure of the protein. This substitution can be compensated for by stabilizing substitutions in a few adjacent residues. In addition to structural and functional correlations, the evolutionary history (phylogeny) of the proteins creates correlations between sites ($C_{\text{phylogeny}}$). The evolutionary history will create the strongest correlations between sites and can be difficult to disentangle from $C_{\text{structure}}$ and C_{function}. These main effects (phylogeny, structure, and function) are confounded and are not statistically independent. The $C_{\text{interactions}}$ term accounts for this higher order statistical nonindependence. The unexplained covariation, $C_{\text{stochastic}}$, refers to the lack of fit of the data to the model.

Serpin proteins are an ideal family for studying the evolution of structure and function. Serpins are a large superfamily of serine protease inhibitors characterized by the possession of a single common core domain consisting of three β-sheets and eight to nine α-helices (Gettins 2002*b*). Serpins are widely distributed among eukaryotes and occur in some of the viruses that infect them and have been recently characterized in Archaea and Bacteria (Irving et al. 2000, 2002, 2003; Roberts et al. 2004). Serpins have been the subject of detailed phylogenetic analysis and are currently classified into 16 phylogenetic clades based on sequence and functional analyses (Irving et al. 2000; Atchley et al. 2001; Ragg et al. 2001). Despite the presence of a common fold of approximately 350 residues in all serpins, pairwise identity of primary structures can be as low as 25% (Gettins 2002*b*). Most serpins act as proteinase inhibitors for chymotrypsin-like serine proteinases, although some inhibit other types of proteinases. Serpins also regulate numerous cellular and extracellular processes including blood coagulation, fibrinolysis, complement activation, and tumor suppression (Gettins 2002*b*). Some serpins have lost their inhibitory role and function in blood pressure regulation, hormone binding, and as chaperones or storage proteins.

Serpins have a unique mechanism of action and function like a mousetrap to inhibit proteinases. When a serpin is in the native state, the reactive center loop (RCL) is exposed and accessible for interaction with a proteinase (fig. 1). The target proteinase binds to the RCL and cleaves the P1–P1′ bond (358–359 in α_1-antitrypsin; Supplementary fig. 1). After the RCL is cleaved, it inserts into the A β-sheet (sA). The first residue to insert is P14 (345) and is followed by the flexible hinge region (P15–P9) of the RCL. The hinge portion of the RCL provides the mobility for the conformational change in the native to cleaved transition (Stressed to Relaxed transition) (Hopkins, Carrell, and Stone 1993). The initial point of insertion of

[1] Present address: Department of Biology, The University of North Carolina at Chapel Hill.

Key words: serpin, factor analysis, coevolution, covariation, protein structure, mutual information.

E-mail: mjbuck@bio.unc.edu.

Mol. Biol. Evol. 22(7):1627–1634. 2005
doi:10.1093/molbev/msi157
Advance Access publication April 27, 2005

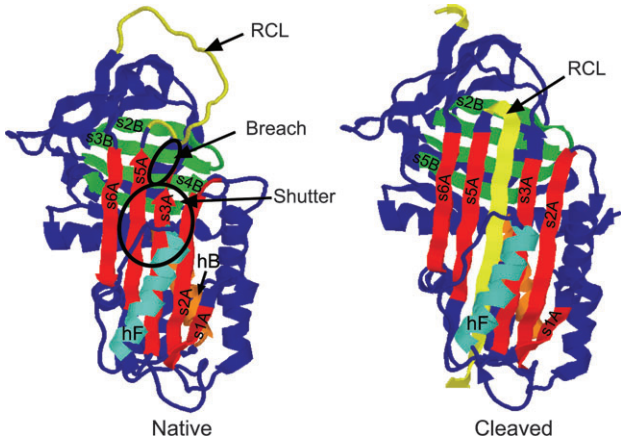

FIG. 1.—The structure of native α_1-antitrypsin and cleaved α_1-antitrypsin indicating important regions. The RCL is colored yellow, sA red, sB green, hF cyan, and hB orange. The positions of the breach and shutter are labeled as defined by Whisstock et al. (2000).

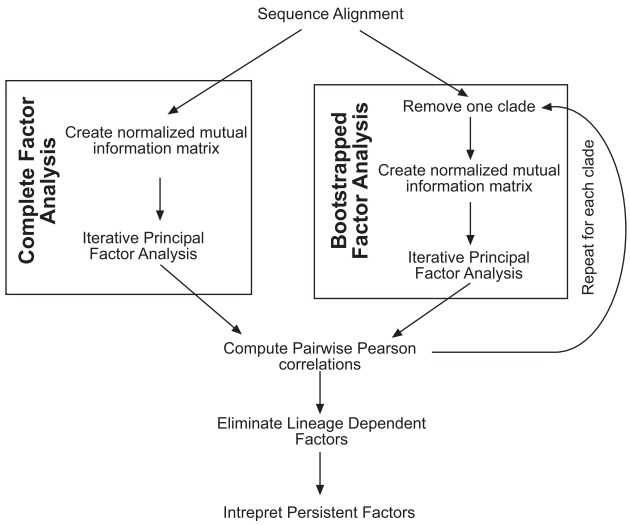

FIG. 2.—Methods flowchart.

the RCL is the breach located at the top of sA (Whisstock et al. 2000). The shutter is located near the center of sA and is composed of s3A, s5A, and neighboring sites. The breach and the shutter are two important regions that facilitate sheet opening and accept the hinge of the RCL as it inserts (Whisstock et al. 2000; Blouse et al. 2003). To insert completely, hF is displaced until the RCL with the bound proteinase passes (Gettins 2002*a*) (Supplementary fig. 2). Then hF returns, covering s3A, and locks the proteinase 70 Å away from the starting position.

In the present paper, we apply the multivariate statistical technique of factor analysis to determine the common underlying correlation structure among amino acid sites within a diverse sample of 211 serpin proteins. The correlation structure or factor patterns were separated into lineage-dependent or lineage-independent patterns using a novel bootstrap factor analysis technique. Three of the lineage-independent patterns were interpreted as a structural or functional pattern of coevolution and identified the coevolving site within the RCL, shutter, breach, and hydrophobic core.

Materials and Methods

The analyses reported here involve a collection of 211 serpin proteins selected from the data set previously analyzed by Irving et al. (2000). The alignment and phylogenetic tree used in this study came from the previous study by Irving et al. (2000). Amino acid sites have been numbered according to the reference sequence α_1-antitrypsin. For the current analyses only sites 23–394 were considered, and columns in the alignment with greater than 20% gaps were removed. A flowchart of the complete analyses is presented in figure 2.

Creating a Covariance Matrix

Covariances among aligned amino acid sites in the studied serpins were estimated using mutual information (MI) (Korber et al. 1993; Clarke 1995; Atchley, Terhalle, and Dress 1999; Atchley et al. 2000). Because sequence elements are symbol variables with no underlying metric, conventional statistical procedures for estimating correlation among variables cannot be used. Thus, normalized mutual information (NMI) values were used to construct a 372 × 372 covariance (**C**) matrix. NMI between sites X and Y, $NMI(X,Y)$, was calculated as follows:

$$NMI(X,Y) = \frac{\left(\sum_i \sum_j P(X_i,Y_j)\log_{20}\frac{P(X_i,Y_j)}{P(X_i)P(Y_j)}\right)}{\sqrt{(H(X)H(Y))}}, \quad (2)$$

where $P(X_i)$ is the probability of i at site X, $P(Y_j)$ is the probability of j at site Y, $P(X_i, Y_j)$ is the joint probability of i at site X and j at site $Y (X \neq Y)$, and $H(X)$ and $H(Y)$ are the entropy values for sites X and Y. The double summation runs over all possible symbols (amino acids) at these sites. Sites that vary independently or are invariant have the minimum NMI value equal to zero. Sites that are perfectly correlated have the maximum NMI value equal to one. NMI is analogous to Pearson's product moment correlation and is bounded by zero and unity.

Factor Analysis

The dimensionality of the **C** matrix was estimated using iterative principal factor analysis (Khattree and Naik 2000). Factor analysis explains the observed covariability among the original variables by a smaller number of unobservable latent variables known as "common factors" (Johnson and Wichern 1988). This technique differs from principal components analysis, in that the latter is primarily a data-compression technique that reduces the dimensionality of all the covariation, whereas factor analysis examines only the common or shared information among elements in the matrix.

To enhance the interpretability of the "common factors" they were rotated to "simple structure" using a Varimax orthogonal rotation (Kaiser 1958). Rotation generally increases the magnitude of large factor coefficients, decreases the size of small coefficients, and generally achieves more interpretable multivariate patterns of

variability. Ideally, each site in the alignment will have a large coefficient on one "common factor" and small to moderate values on all remaining factors. The resulting factor coefficient for each site in the alignment can be considered as the correlation between that site and the "common factor" and varies between −1 and 1. Ten factors were extracted for these analyses. Throughout this paper the common factors from the complete data set will be referred to as a complete data set factor (CDF) followed by the factor numbers 1–10.

Bootstrap Factor Analysis

As described in equation (1), the covariance between any two sites can be caused by phylogeny, structure, function, or the interactions of these major components. To better understand the origins of these patterns of covariation among amino acid sites in serpins, we employed a novel bootstrap factor analysis. The bootstrap factor analysis allowed us to test if any of the 10 CDFs, identified above, depend on a single clade. Any CDF dependent on the covariance from one clade will not appear in the corresponding bootstrap factor analysis. This will limit the spurious factors which may be caused by chance or lineage. Specifically, the procedure was as follows: (1) 21 new data sets were constructed, each missing one clade (A to P); (2) iterative principal factor analysis, extracting 15 common factors, with a Varimax rotation was repeated on each of the 21 datasets, producing 21 new factor patterns (bootstrap data set factors; BDF); (3) pairwise Pearson correlations were computed between all the CDFs and BDFs, producing 21 correlation tables (size 10×15); and (4) lineage-dependent factors were identified by a low correlation, less than 0.5, to every extracted BDF and were eliminated from future analyses. While all the CDFs are important to understand the evolution of serpins, the remaining factors (nonlineage dependent) represent important common factors for all serpin proteins.

Factor Pattern Interpretation

To interpret the biological importance of each factor pattern the factor coefficients were projected on the three-dimensional serpin structure of the native and cleaved form of α_1-antitrypsin (1QLP and 7API; Elliott, Abrahams, and Lomas 1998) (Loebermann et al. 1984), obtained from the Protein Data Bank (www.rcsb.org; Berman et al. 2000). The results were color coded by the factor coefficient using RasMol Molecular Graphics v 2.6 (Sayle and Milner-White 1995), i.e., sites >0.5, red; > 0.4, yellow; >0.3, green; <0.3, blue.

MI statistics were used to estimate the phylogenetic signal and diagnostic sites in the serpin alignment. Phylogenetic signal for each column in the alignment is estimated by the extent of association between the amino acid composition and a dummy variable for clade membership from a phylogenetic tree (Atchley, Terhalle, and Dress 1999). Diagnostic sites were determined for each clade in the alignment by determining the MI from each site to a dummy variable coded as "1" for that particular clade and "0" for all other clades. This was repeated for each clade, and the 10 sites including ties with the highest MI values were retained as the diagnostic sites for that clade (table 1).

Table 1
The Top 10 Diagnostic Sites Including Ties for Each Clade

Clade	Top 10 Sites
A	187, 259, 160, 84, 258, 250, 366, 115, 85, 79
B	219, 23, 110, 149, 390, 276, 209, 232, 143, 24
C	24, 216, 65, 392, 84, 113, 104, 95, 319, 77, 87, 64, 292, 80
D	27, 103, 121, 28, 150, 153, 118, 157, 53, 112
E	222, 260, 259, 129, 245, 287, 236, 302, 104, 55, 30
F	315, 153, 147, 259, 240, 146, 223, 63, 177
G	147, 240, 297, 120, 309, 389, 307, 311, 98, 106, 310, 375, 367, 270, 272
H	297, 142, 345, 112, 164, 180, 212, 121, 347, 56, 342, 192
I	92, 206, 108, 199, 326, 394, 256, 25, 26, 65
J	391, 85, 99, 214, 119, 314, 212, 210, 271, 113
K	59, 275, 60, 106, 322, 385, 31, 35, 133, 276
L	322, 251, 232, 30, 258, 272, 49, 390, 216, 229
M	275, 231, 66, 247, 261, 116, 131, 192, 177, 84, 345, 339, 188, 336, 52, 242, 267
N	66, 79, 229, 234, 146, 97, 164, 142, 322, 96, 171
O	229, 389, 263, 330, 214, 140, 134, 297, 320, 223, 62, 160, 274, 200, 280, 175, 143, 64, 368, 319, 57, 307, 67, 255
P	58, 211, 38, 275, 157, 27, 30, 227, 271, 295

Residue accessibility for both native and cleaved α_1-antitrypsin was determined using the NACCESS program (Hubbard and Thornton 1992). This program calculates the atomic accessible surface defined by rolling a probe of given size around a van der Waals surface (Lee and Richards 1971). For this analysis the total residue accessibility surface was used as the site accessibility.

The rate of evolution at each site was estimated using the Rate4Site algorithm (Pupko et al. 2002). Rate4Site uses a maximum likelihood criterion to estimate the normalized rate of evolution at each site, taking into consideration the topology and branch lengths of the phylogenetic tree. Sites with a positive value are evolving faster than average, and sites with a negative value are evolving slowing than average for that protein. To summarize each factor pattern, we calculated the average rate of evolution and accessibility for all sites with factor coefficients greater than 0.3.

Results and Discussion

Our analyses extracted 10 orthogonal factors which reflect the major independent patterns of amino acid variation among the 211 serpin proteins (table 2). Additional significant factors were identified but account for a very small proportion of the variance. Following this, we present the results for the factors found to withstand the bootstrap resampling procedure (1, 3, 6, and 9). Results of the analysis for all the 10 factors are explained in detail in Supplementary Materials online.

Factor 1 (Phylogenetic)

Factor 1 accounts for the largest amount of covariability among the 372 sites. A total of 272 amino acid sites had

Table 2
Description for the 10 Extracted Factor Patterns

Factor Number	Sites >0.3	Average Rate[a]	Overall Description	Bootstrap
1	272	0.33	Phylogenetic	All
2	33	−0.32	Phylogenetic for clade A	Clade A
3	23	−1.02	RCL and interacting sites	All
4	19	−1.03	Quick turns, hB, and other sites which interact with the inserted RCL	Clade K
5	18	−0.73	Four interaction networks (hA, hB, and hD; s1A-hF; s3A; and breach)	Clade N
6	14	−0.71	Most sites hydrophobic and internal	All
7	7	−1.34	Four interaction networks (Pack against 370; hF-s3A; hB to hE; and hI to hI-s5A)	Clade O
8	10	−1.09	Breach	Clade A
9	5	−1.19	Gate and breach region	All
10	3	−1.28	Two interaction networks (hC to hI-s5A and breach)	Clade A

[a] Average rate of evolution calculated from Rate4Site for sites with coefficients above 0.3.

Factor 1 coefficients greater than 0.3 and are located throughout the serpin structure. These sites had an average evolutionary rate, calculated using Rate4Site, of 0.32, and the Pearson product moment correlation between evolutionary rate and Factor 1 coefficients was 0.56 ($P < 0.001$). Because Rate4Site determines a normalized evolutionary rate, these results suggest that sites with high Factor 1 coefficients are evolving quicker when compared to other sites in the serpin protein.

Phylogenetic signal was also positively correlated to Factor 1 coefficients (0.85, $P < 0.001$) (fig. 3). Thus, sites with large coefficients for Factor 1 are also sites whose residue combinations best predict clade membership. Lastly, the Factor 1 coefficients were positively correlated to accessibility at 0.24 ($P < 0.001$).

Factor 3 (RCL, Shutter, and Breach)

Factor 3 had 23 sites with factor coefficients greater than 0.3 (table 3) and significantly selected for sites found in the RCL, shutter, and breach as supported by Fisher's exact test ($P < 0.001$). These 23 sites had an average evolutionary rate of −1.02, by Rate4Site, implying that they evolve slower than most other sites in the serpin proteins. In addition, these 23 sites had a decrease in accessibility of 30.11 Å^2 when the serpin is cleaved.

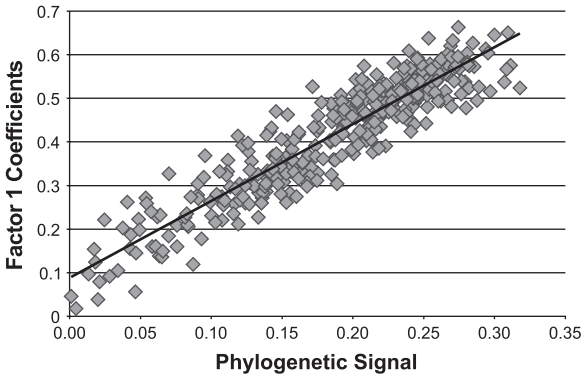

FIG. 3.—Plot of Factor 1 coefficients versus phylogenetic signal. Factor 1 coefficients are significantly correlated to phylogenetic signal; $P < 0.001$.

Table 3
Sites with High Coefficients for Factor 3

Site	Location	Factor 3 Coefficient	Change Accessibility	Connection
342	RCL P17	0.70	3.89	Breach
347	RCL P12	0.60	−67.2	RCL
312		0.54	15.02	Indirect
169		0.53	−0.01	Shutter
56	hB	0.49	−10.8	Shutter
203		0.46	0.19	H-bond to 342
348	RCL P10	0.46	−91.97	RCL
180		0.45	0.14	Indirect
244	s2B	0.44	1.84	Breach
345	RCL P14	0.44	−47.62	RCL
335	s5A	0.42	−9.07	Shutter
346	RCL P13	0.40	−131.66	RCL
351	RCL P8	0.40	−188.47	RCL
161	hF	0.40	−0.1	Shutter
349	RCL P10	0.39	−51.9	RCL
254	s3B	0.39	0.16	?
192		0.36	0.77	Breach
350	RCL P9	0.34	−99.83	RCL
140		0.34	0.47	?
246		0.32	−7.44	H-bond to 244
194		0.31	−4.38	Breach
209	s4C	0.31	13.19	?
112	s2A	0.30	−17.66	?

Specifically, site 342 (P17) had the highest Factor 3 coefficient (0.70) and is located in the breach at the top of s5A (fig. 4*A*). This site is a Glu in 91% of the proteins studied and acts as a pivot for the RCL. Site 347 had the next highest factor coefficient of 0.60 and is part of the hinge of the RCL. All the remaining sites in the hinge also had high Factor 3 coefficients: 345 = 0.44, 346 = 0.40, 348 = 0.46, 349 = 0.39, 350 = 0.34, 351 = 0.40. In addition to sites in the RCL, sites 203, 244, 192, and 188 hydrogen bond to sites in the RCL. Three sites 244, 192, and 194 are part of the breach. Five sites are part of the shutter 169, 56, 335, 161, and 340. The remaining sites with high Factor 3 coefficients have an indirect relationship to the RCL, shutter, and breach. Site 312 is a conserved Phe (90%) that does not interact directly with the RCL but packs underneath sA. Site 180 is a conserved Thr (75%), which stabilizes the turn into s3A, and may be important for shutter opening. Site 246 hydrogen bonds to 244 and may be important for breach function. The remaining four sites 254, 140, 209, and 112 have an unclear relationship to the other sites and the RCL.

In addition to structural-functional relationships for this factor, six of the diagnostic sites for clade H have high coefficients for this factor (56, 112, 180, 347, 342, and 345). This is very interesting and provides functional interpretation. Clade H (Hsp 47) proteins are noninhibitory and function as chaperone proteins and do not require an insertion of the RCL to function. Site 56 is a conserved serine in the shutter of most serpins and hydrogen bonds to another site

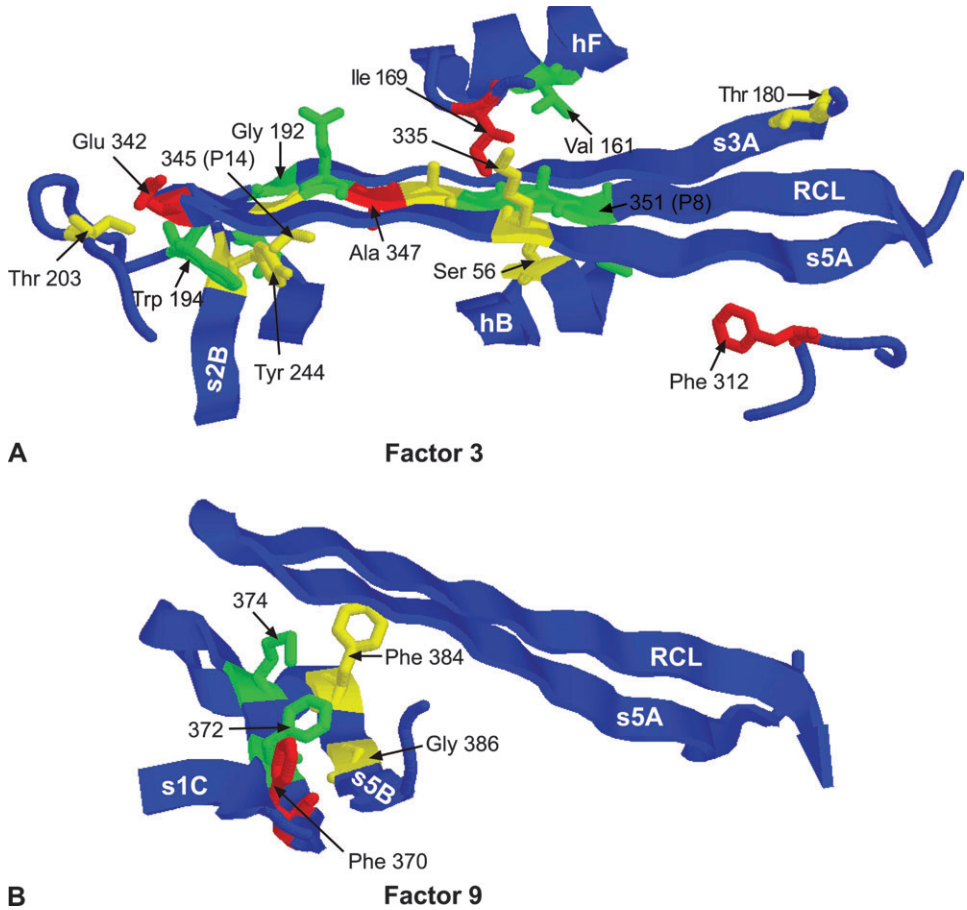

FIG. 4.—Amino acid sites in the cleaved serpin structure are color coded by their factor coefficients. Sites with coefficients >0.5 are color coded red, >0.4 yellow, >0.3 green, and <0.3 blue. Conserved sites (>70%) are labeled with the conserved residue. (*A*) Factor 3 sites on the cleaved structure with RCL inserted. This factor contains the entire hinge region 345–350 and sites which interact with the RCL. Sites are not shown, and their coefficients are 254 = 0.39, 140 = 0.34, 209 = 0.31, and 112 = 0.30. (*B*) Factor 9 hydrophobic core. All side chains face toward the inserted RCL and are part of the hydrophobic core. All sites are shown.

in the shutter, a conserved Asn at 186 (Harrop et al. 1999; Krem and Di Cera 2003). In clade H, site 56 is replaced with a valine which cannot hydrogen bond to site 186; therefore, this bond is lost in these proteins. Mutation of this site from a serine to an arginine in neuroserpin causes severe neurodegeneration in humans (Davis et al. 1999). Site 342, mentioned above, is a conserved glutamate in most serpins, hydrogen bonds to site 203, and has a salt bridge to site 290. Mutation of site 342 to a Lys in α_1-antitrypsin causes loss of the salt bridge to site 290 and decreased inhibitory activity, causing liver disease (Stein and Carrell 1995). All clade H proteins contain a threonine at this position, removing the salt bridge to site 290. Site 180 is a conserved Thr which stabilizes the turn into s3A by hydrogen bonding. In clade H, all proteins contain an aspartate which has an increased potential for hydrogen bonding. This substitution probably increases the number of hydrogen bonds in this region because the side chain of aspartate can accept four hydrogen bonds and threonine can accept only two hydrogen bonds. This substitution may remove the needed flexibility in this region required for shutter opening. Site 347 is a conserved alanine in the hinge region of the RCL. Alanine has a small side chain which makes it easier to be inserted into the hydrophobic core. In clade H proteins, site 347 is an aromatic amino acid, either phenylalanine or tyrosine. Both these amino acids have very large side chains making them difficult to be inserted into the hydrophobic core. In summary, Factor 3 contains the sites which coevolve within three important structural domains that function together to allow the RCL to insert into sA after being cleaved.

Factor 6 (Packed Core)

Factor 6 had 14 sites with coefficients above 0.3 and appeared to select for sites with low accessibility (Fisher exact $P = 0.054$). Most of these sites are hydrophobic residues which are involved in internal packing as well as two charged sites (267 and 387) which are involved in a salt bridge (table 4). The three remaining sites with a high accessibility appear to be related by proximity to other sites within this factor. Overall, this factor represents sites which have coevolved to maintain internal packing interactions. These sites are not in close proximity to each other but represent long-range covariation required to maintain the hydrophobic protein core.

Table 4
Sites with High Factor 6 Coefficients

Site	Location	Factor 6 Coefficient	Connection	Accessibility
220	s3C	0.68	Hydrophobic	**1.16**
157	hF	0.51	Hydrophobic	**1.26**
38	hA	0.40	Hydrophobic	**5.34**
275	hG	0.40	Hydrophobic	**1.23**
246		0.39	?	30.81
185	s3A	0.39	Hydrophobic	**0.11**
66	hB	0.39	Hydrophobic	31.07
264	hG	0.39	Salt bridge 387	11.28
299		0.36	Hydrophobic	**1.25**
30	hA	0.34	?	**4.46**
387	s5B	0.34	Salt bridge 264	11.80
251		0.34	Hydrophobic	**0.07**
322		0.31	?	**3.32**
270	hG	0.31	?	90.35

Site with low accessibility (<10) are shown in bold.

Factor 9 (s4B and s5B)

Factor 9 had 5 sites above 0.3 which where all located in the hydrophobic core region on sheets s4B and s5B (fig. 4*B*). In addition, the side chains for these selected sites extend upward toward the inserted RCL. The side chains from the alternating residues of s4B and s5B are the first side chains which will interact with P14 when the RCL is inserted. In addition, this region of the protein is important for accurate folding into the native structure. The native structure of serpins is metastable, in essence, trapped in a local minimal. There is another conformation of serpins called the ''latent'' form, which is more stable than the native form but is inactive (Lomas et al. 1995). The latent form of plasminogen activator inhibitor-1 revealed that the RCL is inserted into the sA with an accompanying displacement of strand s1C (Mottonen et al. 1992). To ensure that folding of serpins does not enter the latent and inactive state, the hydrophobic core has to fold quickly to trap the molecule in the metastable state (Lee, Seo, and Yu 2001). Factor 9 appears to represent sites which have coevolved on s4B and s5B and may be important for the formation of the native over the ''latent'' structure.

Conclusion

The 10 factors examined in this analysis can be interpreted as patterns of covariation reflecting evolutionary change among serpin proteins or as patterns associated with structural-functional relationships. From our analysis we have found three types of factor patterns, phylogenetic, structural-functional, or clade dependent. The phylogenetic factor pattern, Factor 1, has distinctive characteristics including a high correlation between the factor coefficients and phylogenetic signal. Our results for serpins are very similar to analysis for the basic helix-loop-helix proteins (Atchley et al. 2000). In both studies, the first factor extracted reflects variation in amino acid sites important in resolving the evolutionary history of the proteins or $C_{phylogeny}$. Similarity among amino acid sequences due to descent is the reason multiple sequence alignment algorithms are successful and the major reason for the success of phylogenetic analysis of molecular data.

For the structural-functional factors (3, 6, and 9), the average evolutionary rate, calculated by Rate4Site, for the sites with high factor coefficients was −1.02, −0.71, and −1.19, respectively. Because Rate4Site calculates the relative rate of evolution, sites with a negative rate are evolving slower than average when compared to other sites in the protein. It is expected that sites which are important for proteins' structure or function would evolve slower because of selection acting to maintain their structure and function. It is important to note that in this study the evolutionary rate was estimated across the whole tree for each site, although rate variation can exist among different lineages per site (Gaucher, Miyamoto, and Benner 2001; Desper and Gascuel 2004; Inagaki et al. 2004; Ane et al. 2005). The covarion hypothesis for molecular evolution proposes that selective pressures on sites can change through evolutionary time when the function of a protein changes (Fitch and Markowitz 1970; Fitch 1971). We suspect that the evolutionary rate for sites with high coefficients for Factor 3 would be different in inhibitor versus noninhibitor serpins. To correctly model the evolution of a large protein family, like the serpins, a covarion model should be adopted.

Two subclasses of structural-function factors have been defined: domains and long-range interactions. Factors 3 and 9 represent key functional domains in the serpin proteins. These two factors include most of the important functional domains in serpins including RCL, shutter, breach, and hydrophobic core. The last structural-functional subclass contains sites which share a common amino acid function or type and represent long-range interactions. In Factor 6 almost all the sites are conserved hydrophobic amino acids involved in internal stabilization. Factor 6 represents a type of covariation which frustrates other techniques. The sites in Factor 6 contribute to the overall stability of the protein but do not have direct interactions. This type of covariation may be very important for the evolution of globular proteins. Substitutions which destabilize the hydrophobic core can be compensated by substitutions in many other sites throughout the protein core.

The bootstrap factor analysis presented here greatly enhanced the power of our analysis and can be extended in future experiments. This approach can be used with different classifiers (inhibitory vs. noninhibitory, extracellular vs. intracellular) to further classify the serpin protein superfamily. By using different biological activities as classifiers the underlying factor patterns may be described to more specific activities.

Supplementary Material

Detailed description and figures for serpin movement and descriptions for lineage-dependent factors and supplementary figs. 1 and 2 are available at *Molecular Biology and Evolution* online (www.mbe.oupjournals.org).

Acknowledgments

This work was supported by NIH grant GM45344. We are indebted to Professor Hermann Ragg and Maria Tsompana for their helpful comments and suggestions.

Literature Cited

Afonnikov, D. A., D. Y. Oshchepkov, and N. A. Kolchanov. 2001. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics **17**:1035–1046.

Ane, C., J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. **22**:914–924.

Atchley, W. R., T. Lokot, K. Wollenberg, A. Dress, and H. Ragg. 2001. Phylogenetic analyses of amino acid variation in the serpin proteins. Mol. Biol. Evol. **18**:1502–1511.

Atchley, W. R., W. Terhalle, and A. Dress. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48**:501–516.

Atchley, W. R., K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol. Biol. Evol. **17**:164–178.

Berman, H. M., T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook. 2000. The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. **7**(Suppl.):957–959.

Blouse, G. E., M. J. Perron, J. O. Kvassman, S. Yunus, J. H. Thompson, R. L. Betts, L. C. Lutter, and J. D. Shore. 2003. Mutation of the highly conserved tryptophan in the serpin breach region alters the inhibitory mechanism of plasminogen activator inhibitor-1. Biochemistry **42**:12260–12272.

Chelvanayagam, G., A. Eggenschwiler, L. Knecht, G. H. Gonnet, and S. A. Benner. 1997. An analysis of simultaneous variation in protein structures. Protein Eng. **10**:307–316.

Clarke, N. D. 1995. Covariation of residues in the homeodomain sequence family. Protein Sci. **4**:2269–2278.

Davis, R. L., A. E. Shrimpton, P. D. Holohan et al. (20 co-authors). 1999. Familial dementia caused by polymerization of mutant neuroserpin. Nature **401**:376–379.

Desper, R., and O. Gascuel. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. **21**:587–598.

Elliott, P. R., J. P. Abrahams, and D. A. Lomas. 1998. Wild-type alpha 1-antitrypsin is in the canonical inhibitory conformation. J. Mol. Biol. **275**:419–425.

Fitch, W. M. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. **1**:84–96.

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

Gaucher, E. A., M. M. Miyamoto, and S. A. Benner. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98**:548–552.

Gettins, P. G. 2002*a*. The F-helix of serpins plays an essential, active role in the proteinase inhibition mechanism. FEBS Lett. **523**:2–6.

———. 2002*b*. Serpin structure, mechanism, and function. Chem. Rev. **102**:4751–4804.

Harrop, S. J., L. Jankova, M. Coles, D. Jardine, J. S. Whittaker, A. R. Gould, A. Meister, G. C. King, B. C. Mabbutt, and P. M. Curmi. 1999. The crystal structure of plasminogen activator inhibitor 2 at 2.0 A resolution: implications for serpin function. Structure Fold. Des. **7**:43–54.

Hopkins, P. C., R. W. Carrell, and S. R. Stone. 1993. Effects of mutations in the hinge region of serpins. Biochemistry **32**:7650–7657.

Hubbard, S., and J. Thornton. 1992. NACCESS. Manchester. http://wolf.bi.umist.ac.uk/naccess/.

Inagaki, Y., E. Susko, N. M. Fast, and A. J. Roger. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol. Biol. Evol. **21**:1340–1349.

Irving, J. A., L. D. Cabrita, J. Rossjohn, R. N. Pike, S. P. Bottomley, and J. C. Whisstock. 2003. The 1.5 A crystal structure of a prokaryote serpin: controlling conformational change in a heated environment. Structure (Camb.) **11**:387–397.

Irving, J. A., R. N. Pike, A. M. Lesk, and J. C. Whisstock. 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. Genome Res. **10**:1845–1864.

Irving, J. A., P. J. Steenbakkers, A. M. Lesk, H. J. Op den Camp, R. N. Pike, and J. C. Whisstock. 2002. Serpins in prokaryotes. Mol. Biol. Evol. **19**:1881–1890.

Johnson, R. A., and D. W. Wichern. 1988. Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River, N.J.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika **23**:187–200.

Khattree, R., and D. N. Naik. 2000. Multivariate data reduction and discrimination with SAS Software. SAS Institute Inc., Cary, N.C.

Korber, B. T., R. M. Farber, D. H. Wolpert, and A. S. Lapedes. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. USA **90**:7176–7180.

Krem, M. M., and E. Di Cera. 2003. Conserved Ser residues, the shutter region, and speciation in serpin evolution. J. Biol. Chem. **278**:37810–37814.

Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. **55**:379–400.

Lee, C., E. J. Seo, and M. H. Yu. 2001. Role of the connectivity of secondary structure segments in the folding of alpha(1)-antitrypsin. Biochem. Biophys. Res. Commun. **287**:636–641.

Loebermann, H., R. Tokuoka, J. Deisenhofer, and R. Huber. 1984. Human alpha 1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function. J. Mol. Biol. **177**:531–557.

Lomas, D. A., P. R. Elliott, W. S. Chang, M. R. Wardell, and R. W. Carrell. 1995. Preparation and characterization of latent alpha 1-antitrypsin. J. Biol. Chem. **270**:5282–5288.

Mottonen, J., A. Strand, J. Symersky, R. M. Sweet, D. E. Danley, K. F. Geoghegan, R. D. Gerard, and E. J. Goldsmith. 1992. Structural basis of latency in plasminogen activator inhibitor-1. Nature **355**:270–273.

Pollock, D. D., and W. R. Taylor. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. **10**:647–657.

Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

Pritchard, L., P. Bladon, J. M. O. Mitchell, and M. J. Dufton. 2001. Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng. **14**:549–555.

Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18** (Suppl. 1):S71–S77.

Ragg, H., T. Lokot, P. B. Kamp, W. R. Atchley, and A. Dress. 2001. Vertebrate serpins: construction of a conflict-free

phylogeny by combining exon-intron and diagnostic site analyses. Mol. Biol. Evol. **18**:577–584.
Roberts, T. H., J. Hejgaard, N. F. Saunders, R. Cavicchioli, and P. M. Curmi. 2004. Serpins in unicellular Eukarya, Archaea, and Bacteria: sequence analysis and evolution. J. Mol. Evol. **59**:437–447.
Sayle, R. A., and E. J. Milner-White. 1995. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. **20**:374.
Stein, P. E., and R. W. Carrell. 1995. What do dysfunctional serpins tell us about molecular mobility and disease? Nat. Struct. Biol. **2**:96–113.
Tuff, P., and P. Darlu. 2000. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol. Biol. Evol. **17**:1753–1759.
Wang, Z. O. and D. D. Pollock. 2005. Context dependence and coevolution among amino acid residues in proteins. Methods Enzmol. **395**:779–790.
Whisstock, J. C., R. Skinner, R. W. Carrell, and A. M. Lesk. 2000. Conformational changes in serpins: I. The native and cleaved conformations of alpha(1)-antitrypsin. J. Mol. Biol. **295**: 651–665.

William Martin, Associate Editor

Accepted April 18, 2005