A detailed consideration of a principal domain of vertebrate fibrinogen and its relatives

RUSSELL F. DOOLITTLE

Departments of Chemistry and Biology, Center for Molecular Genetics,
University of California at San Diego, La Jolla, California 92093-0634

(RECEIVED July 17, 1992; REVISED MANUSCRIPT RECEIVED August 28, 1992)

Abstract

Vertebrate fibrinogen is a complex multidomained protein, the structure of which has been inferred mainly from electron microscopy and amino acid sequence studies. Among its most prominent features are two terminal globules, moieties that are mostly composed of the carboxyl-terminal two-thirds of the β and γ chains. Sequences homologous to the latter segments are found in several other animal proteins, always as the carboxyl-terminal contributions. An alignment of 15 amino acid sequences from various fibrinogens and related proteins has been used to make judgments about secondary structure. The nature of amino acids at each position in the alignment was used to distinguish α helices and β structure on the one hand from loops and turns on the other, and the resulting assignments compared with predictions of secondary structure by other methods. Additionally, constraints imposed by the locations of cystines, carbohydrate attachment residues, and proteinase-sensitive points provided further insights into the general organization of the postulated secondary structures. Other ancillary data, including the effects of bound calcium and the locations of labeled or variant residues, were also considered. An intriguing similarity to a portion of the recently reported structure of a calcium-dependent lectin is noted.

Keywords: domains; structure prediction; vertebrate fibrinogen

Vertebrate fibrinogen molecules are dimeric structures composed of three nonidentical but homologous chains ($\alpha\beta\gamma$)₂. Models of the protein are mostly based on electron microscopy (Hall & Slayter, 1959; Fowler & Erickson, 1979; Williams, 1983; Weisel et al., 1985; inter alia) and a host of biochemical data dealing with its covalent structure (for a review, see Doolittle, 1984). The molecule is triglobular, two terminal globular regions being connected to a smaller central domain by two sets of three-stranded coiled coils. The central region contains all six amino-terminals and is the focus of attack by thrombin, which removes the fibrinopeptides A and B. Each of the two terminal globules is composed in turn of two homologous structures contributed by the carboxyl-terminal two-thirds of the β and γ chains, respectively, the sequences of which are about 40% identical in these regions. The γ -chain portion contains the complementary binding site for the polymerization site uncovered by the removal of the fibrinopeptide A; the homologous

β -chain region may contain a complementary binding site for a secondary polymerization site exposed by the removal of the fibrinopeptide B, but the data are not as firm in this case. High resolution three-dimensional information is wholly lacking. In spite of considerable effort in several laboratories, diffracting crystals of the fragments corresponding to these domains have not been obtained.

Recently, the careful consideration of multiple alignments of homologous sequences has proved a successful strategy for predicting not only the secondary structure of proteins but, in some cases, the general arrangement of the folded polypeptide, as well. Thus, Crawford et al. (1987) used a multiple alignment of α subunits from 10 tryptophan synthetases to propose that the structure must be an eight-stranded parallel α - β barrel, a prediction that subsequently proved to be correct. In a similarly successful effort, Bazan (1990) used numerous sequences of cytokine receptors to predict that parts of these proteins are antiparallel, seven-stranded β sandwiches. In an even more challenging case, Benner & Gerloff (1990) used an alignment of 79 protein kinases to attempt a prediction of what turned out to be a complicated folding pattern for a cAMP-dependent protein kinase. Spurred by these

Reprint requests to: Russell F. Doolittle, Departments of Chemistry and Biology, Center for Molecular Genetics, University of California at San Diego, La Jolla, California 92093-0634.

efforts, I have prepared a multiple alignment of 15 fibrinogen and fibrinogen-related sequences and used it to make a proposal about the secondary structure and folding of the domains composing the terminal globules of vertebrate fibrinogen. I have also tried to heed the advice of Schulz (1988) by combining as many prediction methods as possible in the evaluation of each segment of secondary structure.

Nomenclature

It is important to define terms about the parts of fibrinogen and fibrinogen-related proteins (FRePs) that are considered in this article. The most obvious structures visible in electron micrographs of fibrinogen are the two identical terminal globular regions. Because the term "globule" has other connotations, this region will be referred to as a macrodomain. It is wholly encompassed in the product of limited proteolysis called fragment D, an entity that also contains a short residual portion of the severed three-stranded coiled coil. Considerations of the amino acid sequences (Watt et al., 1978) long ago led to the realization that, although fragment D contains parts of all three nonidentical chains, the macrodomain is composed mainly of two homologous structures contributed by the β and γ chains that must be similarly folded (Fig. 1). This aspect of the model was subsequently confirmed by high resolution electron microscopy that

showed the discrete locations of two domains within each macrodomain (Williams, 1983; Weisel et al., 1985).

Amino acid sequences similar to those of these individual domains have been observed in a number of other extracellular proteins. In an earlier article, we referred to these nonfibrinogen polypeptides as fibrinogen-related proteins, or FRePs (Xu & Doolittle, 1990). Given that we are only focusing on a single part of all these proteins, and given also that the unit seems to have spread to other proteins as an authentic, independently folding domain (always at the carboxyl-terminus), I now suggest that these structures be referred to as fibrinogen-related domains, or FReDs. It should be understood, however, that the ancestral domain clearly predates the invention of vertebrate fibrinogen and must have had a function other than clot formation.

It has been demonstrated in the case of fibrinogen that each domain has definable subdomains. For example, it is well known that the kind of fragment D generated by plasmin differs depending on whether or not calcium is present during the digestion (Haverkate & Timan, 1977). In its absence, several additional cleavages occur, and a large portion of the γ chain is removed. The ready removal of this material implies that under these circumstances this long stretch (109 amino acids) is not intimately associated with the main body of fragment D. Furthermore, if fragment D is digested with various enzymes under even more exhaustive conditions, core fragments

Fig. 1. Two-dimensional representation of human fibrinogen fragment D showing the homologous β - and γ -chain regions. The three-stranded coiled coil that connects this macrodomain to the central domain enters at upper left and is held together by six cysteines involved in a disulfide ring. The outlined area corresponds to the γ -chain domain that is the main focus of this paper (from Doolittle & Laudano, 1980).

are generated that also have been referred to as domains (Privalov & Medved', 1982; Medved' et al., 1986). Scanning microcalorimetry conducted in tandem with the digestion efforts showed that different regions of fragment D melt at different temperatures.

Summing to this point, the stable moiety called fragment D generated by the plasmin digestion of fibrinogen is made up of a portion of the coiled-coil connector and a macrodomain corresponding to the terminal globule visualized by the electron microscope. The macrodomain consists of two domains, D β and D γ , each of which is thought to be composed of two subdomains, d1 and d2. The full domains are made up of 240–280 amino acid residues of continuous sequence, depending on the protein in which they are found. The particular structure to be emphasized in this article is the γ -chain domain (D γ) from human fibrinogen, which contains 260 amino acids.

The 15 sequences

The underlying basis for the multiple alignment approach to structure is the observation that homologous sequences always have similar folding because three-dimensional structures are conserved more strongly during evolution than are amino acid sequences (James et al., 1978). All the same, the more distant the sequence relationship, the more likely that the folding will also exhibit some differences. Accordingly, sequences must be grouped with regard to their degrees of similarity, and proportionate heed to details must be paid.

In the present study, the γ chain of human fibrinogen was chosen as the particular benchmark for study and prediction. As such, the most regard has been paid to

comparisons with other fibrinogen γ chains; rat, frog, and lamprey offered a useful range (category A). The second category (B) of homologous sequences, and the next most closely related, are the fibrinogen β chains from human, rat, chicken, and lamprey. The third category (C) is made up of two sequences whose divergence preceded the duplication leading to fibrinogen β and γ chains. These are an unusual, second open reading frame found in the mRNA of chicken fibrinogen α chains (Weissbach & Grienering, 1990) and a second lamprey fibrinogen α chain with the properties expected of a primitive pre-duplication chain (Pan & Doolittle, 1992). The fourth and final category (D) comprises five homologous non-fibrinogen sequences. These sequences are the carboxyl-terminal domains of a mouse T-cell protein (Koyama et al., 1987), chicken cytotactin (Jones et al., 1989), an unidentified reading frame that overlaps a small portion of another human gene (Morel et al., 1989), a sea cucumber sequence (Xu & Doolittle, 1990), and the *scabrous* gene product from fruit fly (Baker et al., 1990). Overall, the most dissimilar sequences in the set of 15 sequences are still almost 30% identical (Table 1).

Secondary structure measurements

Over the years, there has been a scattering of optical rotatory dispersion and circular dichroism experiments conducted on fibrinogen and its proteolytic fragments (Mihalyi & Godfrey, 1963; Mihalyi, 1965; Huseby et al., 1970; Budzynski, 1971). These studies are in general agreement that the native molecule contains about 35% α helix. When fibrinogen is subjected to proteolytic degradation, the helix content of the core fragments is rela-

Table 1. Pairwise comparisons of 15 fibrinogen-related sequences (percent identity based on aligned regions)^a

	FGAH	FGAR	FGAX	FGAL	FBEH	FBER	FBEC	FBEL	BGCH	FAL2	CYAC	URHU	FGSC	FCYM	FGDM
FGAH	—	87.9	75.2	62.7	42.4	42.0	42.2	38.6	43.2	40.2	38.7	33.8	38.7	44.4	28.8
FGAR		—	71.8	60.0	39.1	39.1	39.7	36.5	39.0	39.8	38.3	32.4	36.3	41.3	27.5
FGAX			—	64.3	40.7	41.2	41.3	38.6	42.4	40.3	36.4	33.2	37.3	43.0	29.2
FGAL				—	40.9	40.9	41.1	42.5	40.9	41.6	35.8	35.8	38.4	42.8	28.5
FBEH					—	91.8	81.1	62.9	44.5	38.1	37.9	37.7	38.7	42.3	30.3
FBER						—	81.9	62.1	44.5	38.1	37.9	37.7	38.7	41.9	29.8
FBEC							—	64.4	44.1	39.0	35.7	36.3	40.1	44.1	29.5
FBEL								—	41.1	38.6	40.6	40.6	40.1	43.2	31.5
BGCH									—	55.2	37.1	40.0	41.2	42.7	34.9
FAL2										—	33.6	34.1	37.7	37.1	33.8
CYAC											—	52.5	47.6	40.5	31.5
URHU												—	42.9	38.7	34.3
FGSC													—	44.6	33.5
FCYM														—	30.4
FGDM															—

^a FGAH, human fibrinogen γ chain; FGAR, rat fibrinogen γ chain; FGAX, frog fibrinogen γ chain; FGAL, lamprey fibrinogen γ chain; FBEH, human fibrinogen β chain; FBER, rat fibrinogen β chain; FBEC, chicken fibrinogen β chain; FBEL, lamprey fibrinogen β chain; BGCH, chicken β - γ chain; FAL2, lamprey fibrinogen α -2 chain; CYAC, chicken cytotactin; URHU, human unidentified reading frame; FGSC, sea cucumber FReP; FCYM, mouse T-cell protein; FGDM, fruit fly *scabrous* gene product.

tively greater, indicating that disorganized material is removed from the parent molecule (Mihalyi & Godfrey, 1963; Huseby et al., 1970). Consistent with this conclusion, circular dichroism studies on fragments D and E reported about 40% and 34% α helix, respectively; the only β structure noted amounted to about 15% in fragment D (Budzynski, 1971).

Because these studies were conducted before the amino acid sequence of fibrinogen was known, the relative contribution of various features like the coiled coils could not be distinguished from other regions of α helix. Today the boundaries of the fibrinogen fragments are well delineated, and the contribution of the coiled-coil connectors to the α helix content can be estimated by simple arithmetic. Thus, each half of the human fibrinogen molecule contains 1,482 amino acid residues, 331 of which are in the coiled-coil connectors, amounting to 22%. The maximum contribution of the severed coiled-coil connectors to fragments D and E can also be estimated. In the case of fragment D, which is the focus here, the three-stranded connector should amount to only 25% of the residues (176 residues of 753) and, if not unwound, should contribute to the observed helicity proportionately. It must be concluded that if fragment D contains 40% α helix as reported, then at least 15%, which amounts to about 125

residues, must occur in regions other than the coiled coils. By the same reasoning, the 15% β structure reported for fragment D would amount to about 20% of the portions not involved in the coiled-coil connector.

Bulk amino acid compositions

Some insight into the nature of the expected structures may be gained by a consideration of bulk amino acid compositions alone. Chou (1989) has tallied the amino acid compositions of the four major classes of proteins (all- α , all- β , $\alpha + \beta$, and α/β) and shown statistically significant differences. Individually and as a group, the 15 FReDs are distinctive in having higher than average amounts of Trp, Tyr, Asn, Gly, and Asp, and lower than average amounts of Leu, Val, Pro, Ala, and Ile. As a class they most resemble the ($\alpha + \beta$) group in the conventional four-class system (Table 2).

Taken all together, the FReDs have significantly more negative charge than positive, but the net charge of individual sequences ranges from -8 to $+7$. Fibrinogen γ -chain domains are always more negative than the β -chain kind. The contribution of sialic acids on attached carbohydrate was not considered here.

Table 2. Average amino acid composition calculated from 15 fibrinogen-related domains^a

Amino acid	FReDs	Global	α	β	$\alpha + \beta$	α/β
Ala	5.5 $\pm$ 2.2	7.8	11.6	7.3	9.3	8.3
Arg	5.3 $\pm$ 1.4	5.1	2.2	2.4	4.1	3.4
Asn	6.0 $\pm$ 1.3	4.3	4.0	5.0	6.4	4.2
Asp	7.7 $\pm$ 1.1	5.3	6.7	4.4	5.9	5.6
Cys	1.9 $\pm$ 0.3	1.9	0.9	2.7	3.9	1.5
Gln	4.2 $\pm$ 1.0	4.2	2.7	4.4	3.9	2.6
Glu	4.8 $\pm$ 1.2	6.3	5.5	3.1	4.6	5.9
Gly	11.4 $\pm$ 1.9	7.2	8.1	10.7	9.1	8.7
His	2.6 $\pm$ 0.9	2.3	4.5	1.8	1.7	2.5
Ile	3.5 $\pm$ 1.2	5.3	3.7	4.3	4.9	5.5
Leu	5.8 $\pm$ 1.5	9.1	9.0	6.4	5.8	7.8
Lys	5.1 $\pm$ 2.5	5.9	10.2	4.1	5.9	7.4
Met	2.7 $\pm$ 1.3	2.3	2.0	0.6	1.3	2.1
Phe	4.2 $\pm$ 1.6	3.9	5.0	3.1	2.8	3.6
Pro	2.9 $\pm$ 0.7	5.2	3.4	4.6	3.8	4.3
Ser	6.8 $\pm$ 1.3	6.8	5.0	12.3	6.7	7.5
Thr	5.6 $\pm$ 1.4	5.9	4.9	9.1	6.2	5.5
Trp	3.8 $\pm$ 0.9	1.4	1.3	1.6	1.6	1.7
Tyr	6.4 $\pm$ 1.3	3.2	2.6	4.0	5.7	3.0
Val	3.8 $\pm$ 1.4	6.6	6.8	8.2	6.5	8.7
Difference index			42.9	38.5	25.0	33.6

^a Fibrinogen-related domain (FReD) compositions given as percents $\pm$ SD. Values for (α), (β), ($\alpha + \beta$), and (α/β) protein classes taken from Chou (1989). The global composition is taken from Doolittle (1987). The "difference index," $DI = \sum |di_A - di_B|$, measures the difference between the FReD group and each of the four classes. The DIs are given in parentheses. The least different group, compared with FReDs, is the ($\alpha + \beta$) class of proteins.

The alignment

The 15 sequences were aligned by the progressive method (Feng & Doolittle, 1990). The γ -chain sequences were entered first, followed by the set of separately prealigned β -chain sequences, then the two " β - γ " sequences, also prealigned, and then, finally, the five other sequences (Fig. 2). All told, there are 24 invariant residues among the 240 numbered positions. Of these, 4 are cysteines, 5 are glycines, and 10 are aromatic. There were numerous other positions in which only one or two of the 15 sequences differed from all the others.

Numbered positions were defined as those occupied by an amino acid in more than half of the sequences in the alignment. Points of insertion or deletion were given letter designations corresponding to the previous residue in the sequence. As an example, positions 8a and 8b oc-

cur only in the two category C sequences. Although all the sequences are shown completely through to their carboxyl-termini (Fig. 2), useful comparisons only extend to about position 240, by which point several of the sequences have terminated.

Deriving information from the alignment

General strategy

As a first step, the alignment was used to find the locations of likely connecting turns and loops, on the assumption that breaks in the alignment should not occur in regions of α helix or β structure. Next, the segments between breaks were evaluated for α -helix or β -structure propensity by various procedures. The fact that circular dichroism measurements—indirect as they may have

	-4	-1	v	v	v	40	v	v	80	v
FGAH	ITGKDCQDIANK	G	AKQSGLYFI	KPLKANQQFLVYCEIDGSGNGWTFVQKRLDGSVDFFKNWQYKEGFGHLS					PTGTEFWLGNKIHLLISTQSAIPYA	
FGAR	ITGKDCQDTANK	G	AKESGLYSI	RPLKATQQSLVYCEIDGPGNGWTEFKRLDGSVDFFKNWQYKEGFGHLS					PTGTEFWLGNKHNHLSMQSTIPYA	
FGAX	FTGKDCQEVANK	G	ARLSGLYII	KPLKAKQQFLVYCEIEPSSGSAWTVIQRRLDGSVNFHKNWQYREGFGYLS					PNDKTEFWLGNKIHLLSTQSTIPYV	
FGAL	ITGKDCQQVVDN	G	GKDSGLYII	KPLKAKQPFVFCIE NGNGWTVIQRRHDGSVNFTRDQVSYREGFGYLA					PTLTTEFWLGNKIHLLTGQQ	AYR
FBEH	VSGKECEEIIRK	G	GETSEMYLI	QPDSSVKPYRVYCDMNTENGWTVIQNRQDGSVDFFGRKWDPPYKQFCGNVATNTDGKNYCGPLGGEYWLGNDKISQLTRMG						PTE
FBER	VSGKECEEIIRK	G	GETSEMYLI	QPDTSKPYRVYCDMKTENGWTVIQNRQDGSVDFFGRKWDPPYKQFCGNVATNTDGKNYCGPLGGEYWLGNDKISQLTRIG						PTE
FBEC	VSGRECEIYRK	G	GETSEMYII	QPDPTTTPYRVYCDMETDNGWTVLIQNRQDGSVNFGRWDEYKRGFGNIA					KSGGKPYCDTPGEYWLGNDKISQLTKIG	PTK
FBEL	VSGMHCEIYRN	G	GRTSEAYYI	QPDLFSEPYKVFCDMESHGWWTVVQNRVDGSSNFARDWNTYKAGFGNIA					FDNGKSIENIPGEYWLGTCTVHQLTKQH	TQQ
BGCH	LFLLDCCDIIRKHTS	GA	SGSIFIKI	KPEGSNKVLSVYCDQETTLGGWLLIQRRMDGSVNFNRWQDYRRGFGSDV					KGKQGEWLGNENIHLTQND	TL
FAL2	SEYIDCLDLVLRPPG	GA	SGSLEYEV	RPRGAKRALTVHCEQDQDGGWTLVQOEDGSLNFRSFSAYREGFGTVD					SGSGEELWLGLEAMYLAHED	ST
CYAC	PYPKDCSQALLN	G	EVTSGLYTIYLNQDRTPLOVFCDMAEDGGGWIVFLRRQNGKEDFYRNWKNYVAGFGDP						KDEFWIGLENLHKISSQG	QYE
URHU	PFPDCCGEEMQN	G	AGASRTSTIFLNGNRERPLNVFCDMETDGGGWLVFQRRMDGQDTPFRWDEYAHGFGNI						SGEFWLGNENLHSLTQAG	DYS
FGSC	EYPRDCYDILQSCSG	Q	SPSPGQYII	QPDG GNLIKVYCDMETDEGGWTVFQRRIDGTINFYRSWSYQTCFGNL					NTEFWLGNENIHYLTSSQG	DYE
FCYM	LIYKDCSDHYVL	G	RRSSGAYRV	TPDHRNSSFEVYCDMETMGGGWTVLQARLDGSTNFTREWKDYKAGFGNL					EREFWLGNENIHLTSSKQ	EMI
FGDM	KLPDHCSEVH		TQDGLHLI	APAGQRHPLMTHCTAD					GGGEFWIGNEQLHLLTLDN	CSR
				GWTTVQRRFDSADNFRSWADYAGFGAP						
	v	v	120	v	v	160	v	v		
FGAH	LRVELEDWNGRTSTADYAMFKVGP	EA	KYRLTYAYFAGGDA	DAFGDFGDDP	SDKFFTS	SHNGMQFTSTWDNDNDKFEG	NCAEQDG	SGWMMNKCHAGHLNGV		
FGAR	LRQLKDWSGRTSTADYAMFRVGP	SD	KYRLTYAYFAGGDA	DAFGDFGDDP	SDKFFTS	SHNGMHFTSTWDNDNDKFEG	NCAEQDG	SGWMMNKCHAGHLNGV		
FGAX	MRIELEDWSNQKSTADYSTFR	LGSEKDNRYFTYAYF	IGGDA	DAFGDFGDDP	SDKFFTS	SHNGMQFTSTFDKNDKFEG	NCAEQDG	SGWMMNRCHAAHLNGK		
FGAL	LRIDLTDWENTHRYADYGHFK	LTPESDEYRLFYSMYL	DGDA	DAFGDFGDDP	QDKFYTHLGM	FTSTPERDNDKYE	SCAEQDG	SGWMMNRCHAGHLNGK		
FBEH	LLIEMEDWKGDKVKAHYGGFT	VQNEANKYQISVNKY	KGTAGNALMEGASQLV	GENRMTIHNMGFTSTYDRDNDGWLTSDPRKQCSKEDG			SCAEQDG	SGWMMNRCHAAHPNGR		
FBER	LLIEMEDWKGDKVKAHYGGFT	VQNEANKYQISVNKY	KGTAGNALMEGASQLV	GENRMTIHNMGFTSTYDRDNDGWLTSDPRKQCSKEDG			SCAEQDG	SGWMMNRCHAAHPNGR		
FBEC	VLIEMEDWKGDKVSAHYGGFT	IHNENKYLQSVSNY	SGNAGNALMEGASQLV	GENRMTIHNMGFTSTYDRDNDGWLTSDPRKQCSKEDG			SCAEQDG	SGWMMNRCHAAHPNGR		
FBEL	VLFDMDSWEGSSVYAYASFR	PENEAQGYRLWVEDY	SGNAGNALMEGASQLV	GENRMTIHNMGFTSTYDRDNDGWLTSDPRKQCSKEDG			SCAEQDG	SGWMMNRCHAAHPNGR		
BGCH	LRVELQDWGDNATYAEY	IVQVTEAEGYALTSSY	EGTAGDALVAGWLEEGSE	YTSHAQMQFTSTDRDQDHW	EE	SCAEVYG	SCAEVYG	GGWYNNRCHAAHLNGI		
FAL2	MRVELQDWGAGAHAEY	TVTLRDSSKGYALQVSDY	RTAGNALVSG VADDPE	LTSHGGMFTSTYDRDQDHW	SDG	SCAEWYG	SCAEWYG	GGWYNNRCHAAHLNGI		
CYAC	LRVDLRDR GETAYAVYDK	FSVGDARTRYRLRV	DVGY SGTAGDS	MTYHNGRSFTSTFDKNDSAIT		NCALSYK	NCALSYK	GAFTWYNNCHRVNLNGR		
URHU	IRVDLRAG DEAVFAQYDS	FSHVDSAAEYRLHLEGY	HYTAGDS	MSYHSGSVFSARDRDPNSLLI		SCAVSYR	SCAVSYR	GAFTWYNNCHYANLNGI		
FGSC	LRVELNNLTNGHYAKNGL	FRIGDSFSEYLLVLGAY	SGTAGDS	LAYHNTMRFTSTYDRDNDVYSI		NCASHSYR	NCASHSYR	GAFTWYNNCHYANLNGI		
FCYM	LRIDLEDNFNLTYALYDQFY	YVANEFLKYLRLHIGNY	NGTAGDALRFSR	HYNHLDRFTSTYDRDNDVPSG		NCGLYYS	NCGLYYS	SGWMMNRCHAAHLNGK		
FGDM	LQVQMDDIYDNVVAEYKRFY	ISSRADGYRLHIAEY	SGNADSA	LNYYQGMQFSAIDDDRDISQT		HCAANYE	HCAANYE	GGWMMNRCHAAHLNGR		
	v	200	v	240						
FGAH	YYQGGTYSKA	STPNGYDNGIIWATWKTRW	YSMKKTTMKIIPFNRLTIGEGQ	QHHLGGAKQAGDV						
FGAR	YYQGGTYSKS	STPNGYDNGIIWATWKTRW	YSMKETTMKIIPFNRLSIGDQ	QHGMGSKQVGD						
FGAX	YYQGGTYSADSGPSGYDNGIIWATWRRR		YSMKSVTMKIMPLNRYG	AEGQQTILGSSKKSDFENRGDF						
FGAL	YYFGGNYRKT	DVEFFPYDDGIIWATWHDWR	YSLKMTTMKLLMGRDL	SGHGGQKSKGNSRGDN						
FBEH	YYWGGQYTD	MAKHGTDGIVVMMNWKGSW	YSMRKMSMKIRPFVFPQ							
FBER	YYWGGLYSWD	MSKHGTDGIVVMMNWKGSW	YSMRMSMKIRPFVFPQ							
FBEC	YYWGGTYSWD	MAKHGTDGIVVMMNWKGSW	YSMRKMSMKIRPFVFP							
FBEL	YYWGGIYTKE	QADYGTDDGIVVMMNWKGSW	YSMRKMSMKIRPFVFP							
BGCH	YYPGGHYDPRYNVPYE	IEENGVVWIPFRASD	YSLKVVMMKIRPLETL							
FAL2	YYQGGPYDPRKPPYE	EVENGVVWATYRGSD	YSLKRTAVRFRVQIPIVE							
CYAC	Y	GDNNHSGQVNWFWHKHGE	YSIQFAEMKLRPSSFRNLEGRKKRA							
URHU	Y	GSTVDHQGVSWYHWKGE	YSVPFTMKLRPRNFRSPAGGG							
FGSC	Y	YDYSGAPSIYWSYLPDGN	DQIPFAEMKLRNRSI							
FCYM	Y	YHQYKGVNRNGIFWGTWPGINQAQ	PGGYKSSFFQAKMMIRPKNFKP							
FGDM	Y	NLGL TWFDAARNEW	IAVKSRRMLVKRLPAVECCANASASGAFVSVSGSAADAAPSSGATTTTTTATAAPATVTPKTNNSVQFVAAGQA							

Fig. 2. A 15-way alignment of homologous sequences corresponding to the outlined area in Figure 1 (an additional four residues of lead-in sequence are included). Small v's in number-line indicate gaps in the alignment. Wholly invariant residues are emboldened. FBEH, human fibrinogen γ chain 148–411; FGAR, rat fibrinogen γ chain 148–411; FGAX, frog fibrinogen γ chain 150–417; FGAL, lamprey fibrinogen β chain 149–408; FBEH, human fibrinogen β chain 206–461; FBER, rat fibrinogen β chain 206–461; FBEC, chicken fibrinogen β chain 211–464; FGAL, lamprey fibrinogen β chain 227–479; BGCH, chicken open reading frame found in fibrinogen α -chain message; FAL2, lamprey fibrinogen α -II 399–691; FCYM, mouse T-cell protein 200–431; CYAC, chicken cytactin 1,690–1,912; FGSC, sea cucumber protein 65–282; URHU, human open reading frame 601–820; FGDM, fruit fly *scabrous* gene product 512–774.

been—suggest that there must be both α helix and β structure present was taken into account, as was the nature of the bulk amino acid compositions. Finally, various topologies were considered, especially in the light of known disulfide connections.

Secondary structure interruption points

Five different approaches were used to find those portions of the sequences that were most likely involved in connecting loops and turns. These were: (1) gaps in the alignment, (2) a simple inspection of the interior-surface assignments made on the basis of hydropathy, (3) indexing of the most polar segments on hydropathy plots, (4) Chou–Fasman (CF) prediction of turns, and (5) Garnier–Osguthorpe–Robson (GOR) prediction of turns. Additionally, heed was paid to ancillary data, including either known or potential points of carbohydrate attachment and proteinase-sensitive regions.

Exterior and interior residues

Each position in the alignment was considered individually in the style of Benner and Gerloff (1990), both with regard to the degree of conservation and to the nature of the amino acids involved. Interior locations were assigned only to those positions occupied by amino acids with hydrophobic side chains (L, I, V, M, F, W, Y, A, C) in at least 12 of the 15 sequences. All told, there were 64 such positions (denoted by “I” or “C” in Fig. 3). Additionally, another 10 positions that may be interior only in γ chains (category A) were denoted by the letter “i”. By default, all other positions, amounting to more than two-thirds,

were considered to be on the surface and are denoted “s” in Figure 3.

Carbohydrate-attachment points

As noted, one of the factors considered as evidential support for an exterior location was the existence of potential glycosylation sites, the presumption being that glycosylated asparagine residues must be exposed. Indeed, one survey showed that more than 70% of structurally defined asparagines that are glycosylated occur in β turns (Mononen & Karjalainen, 1984). There are 20 potential glycosylation sites (N-X-T or N-X-S) scattered among the 15 sequences being studied here. Four of these occur in fibrinogen β chains at position 148, although only two of them, human and lamprey, have been shown by experiment to have carbohydrate attached (Watt et al., 1978; Bohonus et al., 1986). Five more potential sites occur at position 54 (or 56), but only one of these, the lamprey fibrinogen γ chain, has been verified experimentally (Strong et al., 1985). It can be concluded that positions 148 and 54 are exterior locations. The consensus occurs twice at two other locations (130–132 and 232–235) and singly in four others. It might be mentioned that, based on the amino acid composition, the expectation for a random distribution would be for 28 N-X-T and N-X-S sequences in all.

Interruption points

There were 15 points of interruption identified by insertions or deletions in the alignment (Fig. 2); most of these occurred in regions of four or more consecutive surface

Fig. 3. “Inside-outside” map showing expected secondary structure elements, disulfide connections, and carbohydrate attachment points. I = interior; C = invariant cysteine; s = surface; i = inside γ chains; KD = Kyte–Doolittle; k = polar extreme; d = hydrophobic extreme. Miscellany: c = sometimes cysteine; cho = potential carbohydrate site; p = plasmin attack site; m = genetic defect known in humans; a = affinity-labeled residue. Solid vertical lines indicate points of deletion or insertion; short vertical lines denote secondary structure interruptions without gaps in sequence alignment. CF = Chou–Fasman; GOR = Garnier–Osguthorpe–Robson; x = α helix, b = β strand, t = turn.

residues (Fig. 3). These locations were automatically presumed to be connecting loops or turns. Three additional regions met various conditions for loops or turns, such as four or more consecutive surface residues or turn prediction by the computer programs, even though gaps in the alignment did not occur (Fig. 3). The 18 separation points (vertical lines in Fig. 3) partitioned the "240-residue" generic sequence into 19 sectors (Roman numerals in Fig. 3).

Secondary structure assignments

β Structure

A priori, it was assumed there must be a hydrophobic core composed of β structure. As such, any consecutive string of three or more interior residues was initially assigned as β structure. There were five such strings among the 240 numbered residues (Fig. 3). In addition, four other segments with alternating interior-exterior assignments were initially considered as β structure, for a total of nine β strands in the whole domain. A 10th sector (XIV) was assigned as β structure to serve as a partner for sector XV, as will be discussed further below.

α Helices

The helical wheel approach (Schiffer & Edmundson, 1967) was employed as an aid in making assignments of amphipathic α -helices. Projections were made with the aligned residues from all 15 sequences (Fig. 4).

Computer prediction

Each of the 15 sequences was subjected to the computer prediction schemes of Chou and Fasman (1974) and Garnier et al. (1978). A control set of 10 fibronectin type III domains, which are known to be seven-stranded antiparallel β barrels, was also examined by the same programs, as were a set of calcium-dependent lectin sequences. The residue assignments for the 15 FReDs were then put in register with the 15-way alignment and the predicted secondary structures (α helix, β structure, or turn) compared among all 15 sequences. Both programs indicated that the FReDs contain a very high proportion of turns, more than 50% by CF (Chou & Fasman, 1974) and as much as 40% by the GOR (Garnier et al., 1978) method. The two methods were in agreement in predicting five regions of β structure that, as expected, matched the regions chosen by the simple interior-exterior assignment method (sectors II, III, IV, X, and XVIII) (Fig. 3). Two other regions of potential β structure were identified by the GOR method that matched the interior-exterior assignments (sectors VIII and IX). The remaining three β -structure designations (VI, XIV, and XV) were not predicted by

either of the computer programs and are discussed further below.

The accuracy of the CF and GOR programs in making assignments in the control groups, in which case comparison could be made with known structures, was better than 60% in both cases, but with less than half of all designations correct by both measures.

Hydropathy plots

Each of the 15 sequences was subjected to hydropathy analysis (Kyte & Doolittle, 1982) and the peaks and troughs examined in the light of the multiple alignment. In general, the most hydrophobic sectors corresponded to regions initially assigned as β structure ("d"s in Fig. 3). Local minima corresponding to polar regions were in very good agreement with known locations of insertions and deletions ("k"s in Fig. 3). The most polar prominence in each of the 15 sequences corresponded to the big surface loop (Fig. 3).

The generic structure

Backbone topology

Throughout this analysis, consideration was given to the possibility that the general fold of fibrinogen-related domains might resemble an already known topology. All the available data were regarded skeptically, and various possibilities ruled out only when more than one approach was clearly negative. For example, neither an all- α nor an all- β structure seemed possible in the light of (1) available circular dichroism data, (2) amino acid compositions, (3) helical wheel projections, or (4) computer predictions of secondary structure. In aggregate, only an ($\alpha + \beta$) structure seemed reasonable, and the secondary structure assignments of some of the sectors were influenced by this consideration. The final analysis assigned five sectors as α helices and nine sectors as β strands (Fig. 5). The α helices encompassed about 36 amino acid residues (15% of the total), and the β strands about 55 residues (23%). It is likely that not every residue of every β strand will be hydrogen-bonded, of course, so the observed ellipticity by circular dichroism would be expected to be somewhat less.

The choice of overall topology was governed by a number of factors. To start, the notion of two subdomains was invoked, based on the observation that the proteolytic removal of the terminal sections of fibrinogen γ chains (in the absence of calcium) implies that strong secondary structure interactions do not occur between these segments and the main body of the domain.

A second consideration had to do with constraints imposed by disulfide bonds. For example, sectors II and III, both predicted to be β structure on several counts (Fig. 3), are logically paired in an antiparallel mode as a result of

Fig. 5. Diagram of secondary structure elements in FReDs. Sector numbering (Roman numerals) and residue numbering (Arabic numerals) are from Figure 3; cho represents potential carbohydrate attachment sites. Plasmin cleavage sites based on human fibrinogen γ chain only.

constraint exists for the disulfide connecting the carboxyl-terminal end of sector V with the segment abutting the amino-terminal end of sector I in fibrinogen β chains. On balance, which includes the necessity to account for a certain amount of helix based on circular dichroism experiments, the α - β - β - α structure seems most likely (Fig. 5).

Sectors VI-X

Positions 71–140 appear to encompass a second center of β structure. Sectors VIII, IX, and X meet most, but not all, criteria for a buried sheet. In contrast, sectors VI and VII were predicted as helical by both the CF and GOR programs (Fig. 3). Helical-wheel depictions indicated that sector VI did not meet the criteria for an amphipathic helix, but sector VII could be accommodated (Fig. 4C). In the end, the two contrary predictions were compromised in the form of a kinked α helix.

The loop region (positions 141–170)

Sectors XI–XIII appear to form a large surface loop, as evidenced by there being only five “interior” residues in the 30-residue skein. Formally, loops are structures of defined length that lack secondary structure but whose termini are within some specified distance of each other (Leszczynski & Rose, 1986). In this case, the region may actually comprise several smaller loops, some of which are anchored by the occasional nonpolar side chain. Both the CF and GOR programs predicted mostly a string of turns, although the GOR method suggested some β structure in the region 150–160 (Fig. 3). This region also contains a known proteinase cleavage point that is only attacked in the absence of calcium.

Fig. 6. Proposed generic structure for fibrinogen-related domains based on secondary structure predictions (Fig. 3) and other considerations. This schematic representation includes all possible disulfides and carbohydrate clusters (cho). In reality, no FReD is known to have more than one of the carbohydrate clusters. Moreover, the majority of known FReDs have only two disulfide bonds; fibrinogen D β chains are known to have three. Heavy shading corresponds to residues in Figure 3 denoted "I"; light shading corresponds to residues denoted "i". Solid triangles correspond to the 15 break points indicated by gaps in the alignment. The positive numbering begins at the site of a known intron in fibrinogen β and γ chains; the connecting segment to amino-terminal domains has negative numbers. Residues denoted with small letters are restricted to one of the FReD classes. For example, residues 89a and b are found only in fibrinogen γ chains and residues 68a–h only in fibrinogen β chains.

The disulfide finger (positions 171–186)

The invariant cysteine residues at positions 172 and 186 are known to be linked by a disulfide bond in the cases of fibrinogen β and γ chains (Bouma et al., 1978), and it is assumed that the same connection exists in all FReDs. The amino acid sequence between these two positions is interrupted by a midpoint insertion in the sea cucumber FReD, and the sequences on either side of the insertion point have been given separate sector assignments (XIV and XV in Fig. 3). They could exist as a β hairpin or a more loosely configured loop. As noted above, neither the CF nor GOR programs predicted β structure for these segments, but modeling on an Iris computer suggested that a β hairpin was altogether feasible. Ancillary circum-

Fig. 7. Schematic representation of γ chain from human fibrinogen patterned on the generic model in Figure 6. The numbering is for the native chain (411 residues); Asp 152 corresponds to position 1 in the generic structure. Two key bonds cleaved by plasmin in the absence of calcium are specified. Solid circles denote nine residues identified in defective polymerization or calcium binding in human variant fibrinogens (Matsuda, 1991). Asterisk denotes Tyr 363, which reacts with a peptide affinity labeling reagent (Yamazumi & Doolittle, 1992).

stances, described more fully below, led to the depiction shown in Figures 5–7.

The carboxyl terminus (positions 187–240)

The disulfide finger is followed by a nine-residue sequence (sector XVI) that has interior residues at the second, fifth, and ninth positions. The prediction programs mostly predicted "turn" throughout, but the helical wheel approach did not rule out a short amphipathic helix. The segment is followed by an expendable 10-residue loop (sector XVII) that is absent in cytactin and other nonfibrinogen proteins. This loop is followed by another nine-amino acid residue stretch of surface residues, likely another exposed loop structure. In contrast, the amino acids at positions 213–216 are invariably hydrophobic, suggesting that this region (sector XVIII) contains a buried β strand. Both prediction programs were in accord (Fig. 3). The strand was arbitrarily paired with sector X of subdo-

main d1. It could equally well have been associated with the β strands that make up the disulfide finger.

Finally, the domain ends with a segment (sector XIX) for which both the CF and GOR programs predict α helix. The helical wheel depiction for this segment is less than canonical, however, an invariant methionine residue being firmly positioned on what would be the polar face of the helix (Fig. 4d). A more reasonable depiction, given the consistent occurrence of four surface residues in a row at locations 226–229, might be an anchored loop, and the assigned helix in Figures 5–7 should be regarded as equivocal.

The fibrinogen γ -chain domain

The human fibrinogen γ -chain sequence was divided into the 19 sectors, and decisions about its secondary structure in particular were made on the basis of the proposed generic model (Table 3). The sectors were arranged along the same lines as the generic model, except that other information about γ chains was included, such as potential calcium-binding sites (Dang et al., 1985; Varadi & Scheraga, 1986), affinity-labeled residues (Yamazumi & Doolittle, 1992), variants that are defective (Matsuda, 1991), and newly exposed cleavage points (Fig. 6). The

Table 3. Sequence sector assignments for human fibrinogen γ -chain^a

HFG number	FReD position	Sector	Sequence	Assignment
d1				
143–151	–9––2		VQIHDITG	(Link to DSR)
152–158	–1–8	I	KDCQDIANK	Helix A
159–164	9–13		GAQQS	Turn
165–170	14–19	II	GLYFIK	β 1
171–177	20–25		PLKANQ	Loop
178–185	26–33	III	QFLVYCGI	β 2
186–189	34–38		DGSGN	Turn
190–196	39–45	IV	GWTVFQK	β 3
197–203	46–52		RLDGSVD	Loop
204–218	53–64	V	FKKNWIQYKEGF	Helix B
219–224	65–73		GHLSPTGTT	Loop
225–228	74–77	VI	EFWL	Helix C
229–232	78–79		GN	Kink
233–237	80–86	VII	EKIHLIS	Helix C
238–243	87–90		TQSAIP	Turn
244–251	91–98	VIII	YALRVELE	β 4
252–259	99–106		DWNGRTST	Loop
260–267	107–114	IX	ADYAMFKV	β 5
268–272	115–119		GPEAD	Turn
273–282	120–129	X	KYRLTYAYFA	β 6
283–302	130–148	XI, XII	GGDAAGDAFDGDFGDDPSDK [‡]	Loops
d2				
303–324	149–170	XIII	FFTSHNGMQFSTWDNDNDKFEG	Loops
325–330	171–175	XIV	NCAEQ	β 9
331–334	176–179		DGSG	Turn
335–340	180–185	XV	WWMNKC	β 10
341–344	186–190		HAGHL	Turn
345–350	191–196	XVI	NGVYYQ	Loop
351–365	197–211	XVII	GGTYSKASTPNGYDN	Loops
366–371	212–217	XVIII	GIIWAT	β 11
372–377	218–223		WKTRWY	Turn
378–387	224–233	XIX	SMKKTMTKII	Helix D
388–395	234–241		PFNRLTIG	Loop
396–411	(242–257)		(EGQQHHLGGAQAGDV)	(Helix E)

^a The sequences shown are from the human fibrinogen γ chain (HFG), the fibrinogen-related domain (FReD) of which begins at residue Asp 152. The first segment shown is the connecting link back to the disulfide ring (DSR) that terminates the coiled coil. The last segment listed is one found only in γ chains and does not have a Roman numeral designation. FReD positions end at number 240. The γ -chain sequence corresponding to the connector between sectors VII and VIII has two more residues (89a and b) than the generic FReD and sector XI has one extra residue (130a). Interior residues (I or C in Fig. 3) are emboldened. Residues identical in all 15 sequences are underscored. Sector XII is composed of several loops; it has been divided into two regions (d1 and d2) separated at a known plasmin cleavage point ([‡]).

fibrinogen γ chains differ from most of the other FReDs in having an additional 18-residue segment at the carboxyl-terminus. The extra segment is depicted as an α helix, although this should not be regarded as a firm prediction. A model proposing two antiparallel α helices formed from these sectors in cross-linked fibrin was presented some time ago (Doolittle, 1973).

Functional considerations bear heavily on how one views this structure. In this regard, this section of the fibrinogen γ chain has been associated previously with both a site for calcium binding and a major polymerization site for clotting. Indeed, the many variant amino acids in human fibrinogens known to exhibit defective polymerization or impaired calcium binding are mostly clustered in sectors X–XIV (m in Fig. 3). Interestingly, some of these mutant proteins have lost both the ability to bind calcium and to polymerize, and in some cases the ability to polymerize can be restored by the addition of excess calcium.

The "disulfide finger" (sectors XIV and XV) has been implicated in the binding of calcium. For example, in the presence of calcium the reduction of this particular disulfide bond is impaired (Lawrie & Kemp, 1979). Also, the vulnerability of fibrinogen γ chains to proteinase cleavage in sectors XII and XIII depends on calcium being absent. For these reasons, it must be presumed that calcium serves as a principal organizing center for subdomain d2. It is reasonable to presume that both the large surface loop and the disulfide finger contribute some of the ligands required to bind calcium. As described below, an intriguingly similar situation exists in calcium-dependent lectins.

Not all FReDs have high-affinity calcium-binding sites. For example, β -chain domains do not bind calcium tightly, whereas it is thought that cytotactin binds calcium like γ chains (Jones et al., 1989). Accordingly, we can ask what it is about fibrinogen β chains that disallows tight calcium binding. The alignment makes it clear that there is a five-residue insertion between the projected loop segment (sector XIII) and the disulfide finger (sectors XIV and XV) in all β chains, and this may be sufficient to distort the binding site. On the other hand, these chains are not excessively vulnerable to proteolysis, and it may be that they possess weak calcium-binding sites that contribute to their folding.

Polymerization sites

So far as is known, only the γ -chain domains from fibrinogen bind the Gly-Pro-Arg derivatives that are involved in the first step in fibrin formation. Accordingly, we can ask what features they have in common that are absent in the other 11 sequences. There is a one-residue insertion at position 128a that initiates a quite different sequence in the following sector. The region is rich in aspartic acid and may well contribute to the binding of Gly-Pro-Arg (Bohonus et al., 1986).

The various FReDs differ significantly with regard to their carboxyl-termination points. Logically, the longest, which include all fibrinogen γ chains, must have an exposed terminal. It is well established that the carboxyl-terminal 15 residues of fibrinogen γ chains are involved both in intermolecular cross-linking in fibrin (Chen & Doolittle, 1970) and in platelet binding (Hawiger et al., 1982).

Similarity to calcium-dependent lectins

During the course of this analysis, the three-dimensional structure of a mannose-binding protein was reported (Weis et al., 1991). The mannose-binding domain is a member of a very diverse family of calcium-dependent lectins (Drickamer, 1988). Different family members bind one or two calcium atoms. Among the characteristics of this family are a signature sequence, denoted "WIGL", and a disulfide-bonded finger of about the same size as those occurring in FReDs. In the crystallographic molecular model of the mannose-binding protein, the calcium-binding sites are formed by a large surface loop in conjunction with the disulfide finger. The loop is directly adjacent to the disulfide finger (Fig. 8). The WIGL sequence is in a buried strand that forms the third segment of a three-stranded sheet with the finger.

As noted above, several FReDs are known to bind calcium, and the region around the disulfide finger is thought to be involved. Moreover, FReDs appear to have the equivalent of a WIGL sequence (exact in at least one instance) (Fig. 2). In FReDs, however, the WIGL region is in a different subdomain from the disulfide finger. Still, the similar arrangement and context is intriguing (Fig. 8). Coincidentally, also, the often shuffled lectin domain has been found in association with coiled coils of the sort found in fibrinogen (Beavil et al., 1992).

On the other hand, no statistically significant similarity could be demonstrated between the FReDs and the lectins, nor did a 3D-Profile search (Bowie et al., 1991) of the mannose-binding domain uncover any fibrinogens or fibrinogen-related proteins in the Protein Identification Resource protein sequence database. At the very least, the lectin structure demonstrates a remarkably similar setting for binding extracellular calcium.

Discussion

The value of a multiple sequence alignment in gauging structures is very much a function of how different the sequences are in the set under consideration. Thus, a very large number of sequences that are highly similar may not be very helpful. At the other extreme, if all the sequences in the set are highly divergent from each other, there may be too much ambiguity. The sequences employed in the present study cover a good range, the sequences in group A (fibrinogen γ chains) ranging from 60 to 90% identi-

Fig. 8. Schematic representation of calcium-binding regions of mannose-binding protein (MBP-A) and human fibrinogen γ chain (FGAH). An X-ray structure of MBP-A has revealed those residues providing ligands for two bound calcium atoms; solid diamonds denote ligands for calcium-site 1 and solid circles for calcium-site 2 (data from Weis et al., 1991). In the case of FGAH, asterisks denote residues known to be changed or deleted in human variant fibrinogens that exhibit defective calcium binding (data from Matsuda, 1991).

cal, groups B and C to group A averaging 40% identity, and group D, the most divergent, being 30–40% identical to group A.

Fibrinogen-related domains (FReDs) occur at the carboxyl ends of a variety of polypeptide chains, where they are linked by short flexible tethers that often include a proline residue. In at least two instances, fibrinogen and cytactin, they occur as extensions of three-stranded molecules and must coexist in proximity with one or two domains like themselves. Potential interactions between the domains have been neglected in the present study; in defense of this position, it can be said that the domains appear quite discrete in the electron microscope (Williams, 1983).

Each FReD is likely composed of two subdomains. The amino-terminal section, d1, contains approximately 130 amino acid residues. There were 10 breakpoints in the multiple sequence alignment in this region, delineating 10 sections of secondary structure. Six of these were designated as β strands and were arranged as two three-stranded sheets. The remainder includes one long α helix and two shorter ones, one of the latter being kinked and occupying two sectors.

The second subdomain, d2, is made up of about 100 residues. It is connected to the d1 subdomain by a series

of loops, which in some FReDs contribute to a calcium-binding site. The d2 subdomain also contains a disulfide-defined finger. Although the general structure of d2 appears much more open than that of subdomain d1, some β structure has been proposed here, as well as a carboxyl-terminal helix.

Assessing the model

As in the case of some previous attempts at prediction, the FReD assignments seem more certain in the amino-terminal region, perhaps a reflection of natural selection favoring stronger local interactions as an aspect in the initiation of folding (Schulz, 1988). Of necessity, the final assessment of any predicted model must await an experimentally observed structure. Nonetheless, past predictions offer some idea of what can be expected. The α/β barrel predicted by Crawford et al. (1987) belongs to a protein class that includes up to 10% of all known X-ray structures. Similarly, the seven-stranded β barrel proposed by Bazan (1990) also conforms to a previously observed pattern that appears to be common. In contrast, the cAMP-dependent kinase predicted by Benner & Gerloff (1990) was not a simple, previously known, kind of structure. Although the report of the X-ray structure (Knighton et al., 1991) generously comments that Benner and Gerloff's predictions were "remarkably accurate," the fact is that about a third of the assignments for α helices and β strands were incorrect (7 of 19), even with liberal allowances for starting and stopping points. Further, the topology of the predicted structures fell far short of being a realistic representation, only a single pairing of two consecutive β strands as a hairpin being correct. Although reasonable models can be constructed when the information from multiple alignments is carefully considered, in the final analysis the accuracy is a function of how complicated the fold really is. As such, success must be viewed in the light of the degree of difficulty.

In the case of the fibrinogen-related domain, known regular structures were considered and rejected. None of the data were consistent with either an all- α structure or a parallel α/β barrel. An all- β structure might have been possible except for the circular dichroism data that suggest the occurrence of α helix in the domain. The circular dichroism data, the bulk amino acid compositions, and two computer prediction schemes were most consistent with an ($\alpha + \beta$) structure. In this light, a survey of known three-dimensional structures did not reveal any obvious candidates for related structures. Among the possibilities considered were the actin-heat shock 70-hexokinase group (Flaherty et al., 1991), the platelet factor 4 family (St. Charles et al., 1989), and the HLA family (Rippmann et al., 1991). Although each of these had some similar features, none was close enough to warrant a pronouncement of ancient homology. Still it is not impossible that subdomain d1 will turn out to resemble one

or another of these and subdomain d2 something else. For the time being, however, the available data best fit the more complex arrangement shown in Figure 5.

Acknowledgments

I thank William Weis and Wayne Hendrickson for sending me the coordinates of the mannose-binding protein and David Eisenberg for performing a 3D-Profile search. I also acknowledge discussions with Hiroyuki Toh about the significance of sequence relationships and Jack Kyte for helpful comments on the manuscript. I thank Tripos and Associates for the extended use of the Sybyl Programming Package. The assistance of Karen Anderson during the preparation of this manuscript is gratefully acknowledged. This investigation was supported by NIH grant HL26873.

References

- Baker, N.E., Mlodzik, M., & Rubin, G.M. (1990). Spacing differentiation in the developing *Drosophila* eye: A fibrinogen-related lateral inhibitor encoded by *scabrous*. *Science* 250, 1370-1377.
- Bazan, J.F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci. USA* 87, 6934-6938.
- Beavil, A.J., Edmeades, R.L., Gould, H.J., & Sutton, B.J. (1992). α -Helical coiled-coil stalks in the low-affinity receptor for IgE (Fc ϵ RII/CD23) and related C-type lectins. *Proc. Natl. Acad. Sci. USA* 89, 753-757.
- Benner, S.A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.* 31, 121-181.
- Bohonus, V., Doolittle, R.F., Pontes, M., & Strong, D.D. (1986). Complementary DNA sequence of lamprey fibrinogen β chain. *Biochemistry* 25, 6512-6516.
- Bouma, H., III, Takagi, T., & Doolittle, R.F. (1978). The arrangement of disulfide bonds in fragment D from human fibrinogen. *Thromb. Res.* 13, 557-562.
- Bowie, J.U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164-170.
- Budzynski, A.Z. (1971). Difference in conformation of fibrinogen degradation products as revealed by hydrogen exchange and spectropolarimetry. *Biochim. Biophys. Acta* 229, 663-671.
- Chen, R. & Doolittle, R.F. (1970). Isolation, characterization, and location of a donor-acceptor unit from cross-linked fibrin. *Proc. Natl. Acad. Sci. USA* 66, 472-479.
- Chou, P.Y. (1989). Prediction of protein structural classes from amino acid compositions. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), ch. 12, pp. 549-586. Plenum Press, New York.
- Chou, P.Y. & Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry* 13, 222-245.
- Crawford, I.P., Niermann, T., & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: Application to the α subunit of tryptophan synthase. *Proteins Struct. Funct. Genet.* 2, 118-129.
- Dang, C.V., Ebert, R.F., & Bell, W.R. (1985). Localization of a fibrinogen calcium binding site between γ -subunit positions 311 and 336 by terbium fluorescence. *J. Biol. Chem.* 260, 9713-9717.
- Doolittle, R.F. (1973). Structural aspects of the fibrinogen to fibrin conversion. *Adv. Protein Chem.* 27, 1-109.
- Doolittle, R.F. (1984). Fibrinogen and fibrin. *Annu. Rev. Biochem.* 53, 195-229.
- Doolittle, R.F. (1987). *Of URFS and ORFS. A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.
- Doolittle, R.F. & Laudano, A.P. (1980). Synthetic peptide probes and the location of fibrin polymerization sites. *Protides Biol. Fluids* 28, 311-316.
- Drickamer, K. (1988). Two distinct classes of carbohydrate-recognition domains in animal lectins. *J. Biol. Chem.* 263, 9557-9560.
- Feng, D.F. & Doolittle, R.F. (1990). Progressive alignment of protein sequences. *Methods Enzymol.* 183, 375-387.
- Flaherty, K.M., McKay, D.B., Kabsch, W., & Holmes, K.C. (1991). Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. *Proc. Natl. Acad. Sci. USA* 88, 5041-5045.
- Fowler, W.E. & Erickson, H.P. (1979). Trinodular structure of fibrinogen. *J. Mol. Evol.* 134, 241-249.
- Garnier, J., Osguthorpe, D.J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120, 97-120.
- Hall, C.E. & Slayter, H.S. (1959). The fibrinogen molecule: Its size, shape, and mode of polymerization. *J. Biophys. Biochem. Cytol.* 5, 11-17.
- Haverkate, F. & Timan, G. (1977). Protective effect of calcium in the plasmin degradation of fibrinogen and fibrin products D. *Thromb. Res.* 10, 803-812.
- Hawiger, J., Timmons, S., Kloczewiak, M., Strong, D.D., & Doolittle, R.F. (1982). γ and α chains of human fibrinogen possess sites reactive with human platelet receptors. *Proc. Natl. Acad. Sci. USA* 79, 2068-2071.
- Huseby, R.M., Mosesson, M.W., & Murray, M. (1970). Studies of the amino acid composition and conformation of human fibrinogen: Comparison of fractions I-4 and I-8. *Physiol. Chem. Phys.* 2, 374-384.
- James, M.N.G., Delbaere, L.T.J., & Brayer, G.D. (1978). Amino acid sequence alignment of bacterial and mammalian pancreatic serine proteases based on topological equivalences. *Can. J. Biochem.* 56, 396-402.
- Jones, F.S., Hoffman, S., Cunningham, B.A., & Edelman, G.M. (1989). A detailed structural model of cytotactin: Protein homologies, alternative RNA splicing, and binding regions. *Proc. Natl. Acad. Sci. USA* 86, 1905-1909.
- Knighton, D.R., Zheng, J., Ten Eyck, L.F., Ashford, V.A., Xuong, N.-H., Taylor, S.S., & Sowadski, J.M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253, 407-414.
- Koyama, T., Hall, L.R., Haser, W.G., Tonegawa, S., & Saito, H. (1987). Structure of a cytotoxic T-lymphocyte-specific gene shows a strong homology to fibrinogen β and γ chains. *Proc. Natl. Acad. Sci. USA* 84, 1609-1613.
- Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157, 105-132.
- Lawrie, J.S. & Kemp, G. (1979). The presence of a Ca²⁺ bridge within the γ chain of human fibrinogen. *Biochim. Biophys. Acta* 577, 415-423.
- Leszczynski, J.F. & Rose, G.D. (1986). Loops in globular proteins: A novel category of secondary structure. *Science* 234, 849-855.
- Matsuda, M. (1991). Abnormal fibrinogens: The present status of structure elucidation. *Biomed. Progr.* 4, 51-54.
- Medved', L.V., Litvinovich, S.V., & Privalov, P.L. (1986). Domain organization of the terminal parts in the fibrinogen molecule. *FEBS Lett.* 202, 298-302.
- Mihalyi, E. (1965). Physicochemical studies of bovine fibrinogen. III. Optical rotation of the native and denatured molecule. *Biochim. Biophys. Acta* 102, 487-499.
- Mihalyi, E. & Godfrey, J.E. (1963). Digestion of fibrinogen by trypsin. II. Characterization of the large fragment obtained. *Biochim. Biophys. Acta* 67, 90-103.
- Mononen, I. & Karjalainen, E. (1984). Structural comparison of protein sequences around potential N-glycosylation sites. *Biochim. Biophys. Acta* 788, 364-367.
- Morel, Y., Bristow, J., Gitelman, S.E., & Miller, W.L. (1989). Transcript encoded on the opposite strand of the human steroid 21-hydroxylase/complement component C4 gene locus. *Proc. Natl. Acad. Sci. USA* 86, 6582-6586.
- Pan, Y. & Doolittle, R.F. (1992). cDNA sequence of a second fibrinogen α chain in lampreys: An archetypal version alignable with full-length β and γ chains. *Proc. Natl. Acad. Sci. USA* 89, 2066-2070.

- Privalov, P.L. & Medved', L.V. (1982). Domains in the fibrinogen molecule. *J. Mol. Biol.* 159, 665-683.
- Rippmann, F., Taylor, W.R., Rothbard, J.B., & Green N.M. (1991). A hypothetical model for the peptide binding domain of hsp70 based on the peptide binding domain of HLA. *EMBO J.* 10, 1053-1059.
- St. Charles, R., Walz, D.A., & Edwards, B.F.P. (1989). The three-dimensional structure of bovine platelet factor 4 at 3.0-Å resolution. *J. Biol. Chem.* 264, 2092-2099.
- Schiffer, M. & Edmundson, A.B. (1967). Use of helical wheels to represent the structure of protein and to identify segments with helical potential. *Biophys. J.* 7, 121-135.
- Schulz, G.E. (1988). A critical evaluation of methods for prediction of protein secondary structures. *Annu. Rev. Biophys. Biophys. Chem.* 17, 1-21.
- Strong, D.D., Moore, M., Cottrell, B.A., Bohonus, V.L., Pontes, M., Evans, B., Riley, M., & Doolittle, R.F. (1985). Lamprey fibrinogen γ chain: Cloning, cDNA sequencing, and general characterization. *Biochemistry* 24, 92-101.
- Varadi, A. & Scheraga, H.A. (1986). Localization of segments essential for polymerization and for calcium binding in the γ -chain of human fibrinogen. *Biochemistry* 25, 519-528.
- Watt, K.W.K., Takagi, T., & Doolittle, R.F. (1978). The amino acid sequence of the β -chain of human fibrinogen: Homology with the γ -chain. *Proc. Natl. Acad. Sci. USA* 75, 1731-1735.
- Weis, W.I., Kahn, R., Fourme, R., Drickamer, K., & Hendrickson, W.A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science* 254, 1608-1615.
- Weisel, J.W., Stauffacher, C.V., Bullitt, E., & Cohen, C. (1985). A model for fibrinogen: Domains and sequence. *Science* 230, 1388-1391.
- Weissbach, L. & Grieninger, G. (1990). Bipartite mRNA for chicken α -fibrinogen potentially encodes an amino acid sequence homologous to β - and γ -fibrinogens. *Proc. Natl. Acad. Sci. USA* 87, 5198-5202.
- Williams, R.C. (1983). Band patterns seen by electron microscopy in ordered arrays of bovine and human fibrinogen and fibrin after negative staining. *Proc. Natl. Acad. Sci. USA* 80, 1570-1573.
- Xu, X. & Doolittle, R.F. (1990). Presence of a vertebrate fibrinogen-like sequence in an echinoderm. *Proc. Natl. Acad. Sci. USA* 87, 2097-2101.
- Yamazumi, K. & Doolittle, R.F. (1992). Photoaffinity labeling of the primary fibrin polymerization site: Localization of the label to γ -chain Tyr-363. *Proc. Natl. Acad. Sci. USA* 89, 2893-2896.