PROTEINS: Structure, Function, and Genetics 27:279–289 (1997)

A Predicted Consensus Structure for the C Terminus of the Beta and Gamma Chains of Fibrinogen

Dietlind L. Gerloff,[1] Fred E. Cohen,[2] and Steven A. Benner[3]*

[1,2]Departments of Cellular and Molecular Pharmacology, [1,2]Pharmaceutical Chemistry, [2]Biochemistry and Biophysics, and [2]Medicine, University of California San Francisco, San Francisco, California; [3]Department of Chemistry, ETH Zurich, Switzerland; [3]Department of Chemistry, University of Florida, Gainesville, Florida

ABSTRACT **A secondary structure has been predicted for the C termini of the fibrinogen β and γ chains from an aligned set of homologous protein sequences using a transparent method that extracts conformational information from patters of variation and conservation, parsing strings, and patterns of amphiphilicity. The structure is modeled to form two domains, the first having a core parallel sheet flanked on one side by at least two helices and on the other by an antiparallel amphiphilic sheet, with an additional helix connecting the two sheets. The second domain is built entirely from β strands. Proteins 27:279–289** © 1997 Wiley-Liss, Inc.

Key words: protein structure prediction; prediction contest; protein sequence alignment; compensatory covariation; CASP2

INTRODUCTION

One of the defining problems in modern protein chemistry challenges the biological chemist to deduce the conformation (secondary and tertiary structure) of a protein from sequence information (primary structure). Both at the ETH in Zurich[1] and elsewhere,[2–6] progress toward solution of this problem has come through an analysis of patterns of conservation and variation in the sequences of homologous proteins.[7] Such an analysis is especially powerful when it is aided by detailed models of divergent evolution.[8,9] Predictions made using this approach are "consensus" models for conformation of a protein family and assume that proteins related by common ancestry have similar conformations.[10]

The value of these methods has been demonstrated by their application to make bona fide predictions, those published before an experimental structure becomes available. To date, nearly two dozen bona fide predictions have been made using these methods (reviewed in Ref. 11). For about half of these, a subsequently determined crystal structure has emerged to allow these predictions to be evaluated. In most cases, the predictions have proven to be remarkably accurate. Further, misassignments generally fall into only a few categories: secondary structure elements near an active site, internal helices, and noncore regions.

Nevertheless, "perfect" predictions are possible, defined as secondary structural models that miss no core secondary structural elements, misassign no α helices as β strands (or vice versa), and do not overpredict any significant secondary structural element.[12] Predictions that meet this criterion are satisfactory as starting points for assembly of a tertiary structural model of a protein family. Predicted secondary structures for the pleckstrin homology domain,[13,14] the Src homology 2 domains,[2,3] the hemorrhagic metalloproteinases,[15] phospho-β-galactosidase,[16] synaptotagmin,[16] cyclin,[17] the von Willebrand factor,[18] the serine/threonine protein phosphatases,[19] the tyrosine protein phosphatases,[20] and the proteasome[21] come close to perfection by this definition.

Continuing bona fide prediction efforts are necessary to define the scope of this or any other prediction method. Gradually, a large set of examples will emerge that, in time, will become statistically representative of proteins as a whole. It is important, now to move past simple secondary structure modeling, especially to learn how secondary structures might be refined hand-in-hand with efforts to assemble secondary structural elements into tertiary structural models. This will require the development of new tools and more bona fide predictions. As with other areas of chemistry, the first steps taken must necessarily be manual, computer-assisted but not fully automated.

As part of the structure prediction contest to be held in Asilomar in December 1996, we now add to this growing collection of bona fide predictions by examining the secondary and tertiary structure of a segment of fibrinogen. This protein is part of a complex system involved in the clotting of blood.[22] Considerable effort has been devoted to analyzing the structure of fibrinogen using both crystallo-

Contract grant sponsor: NIH, contract grant number GM 39900.

*Correspondence to: Dr. Steven A. Benner, Department Chemie, Universitastr. 16, ETH-Zentrum, Ch-8092 Zurich, Switzerland.

Received 21 June 1996; accepted 15 August 1996.

graphic and noncrystallographic techniques.[23] The protein is organized into multiple domains, many of which can be resolved by partial proteolysis. This paper concerns the C-terminal fragment of the β and γ chains of fibrinogen.

METHODS

A multiple alignment for the protein family was built from sequences extracted from SwissProt[24] using the DARWIN system.[25,26] Surface and interior residues were assigned by automated procedures similar to those described elsewhere,[27] the multiple alignment was parsed into units forming independent secondary structures, and elements of secondary structure were predicted within the parsed segments from patterns of conservation and variation, as described elsewhere.[9,13,15,16,28] Many of the automated routines used in this prediction are available to the public on a server accessible via electronic mail at the address cbrg@inf.ethz.ch, or using the World Wide Web with URL http://cbrg.inf.ethz.ch/.

New in this prediction is an increased reliance on "parsing strings," consecutive positions that contain Pro, Gly, Ser, Asn, or Asp, to assign breaks in secondary structure. Recent work in these laboratories (T. F. Jenny and M. Turcotte, unpublished observations) has suggested that these are significantly more reliable than gaps in assigning breaks in secondary structure.

SECONDARY STRUCTURE PREDICTION

The secondary structure prediction is presented residue-by residue in Figure 1, and summarized in Table I, based on an evolutionary tree shown in Figure 2. The following comments can be made about the predicted secondary structural model.

First, the DARWIN tool generated a coherent multiple alignment including all sequences starting only at position 2037. This is because DARWIN uses stringent criteria to ensure that the multiple alignment is of high quality. The Cys at position 2043 forms a disulfide bond to Cys 2010, however, and it is likely that the folded domain begins somewhere near residue 2000 (in the alignment numbering generated by DARWIN, Figure 1). Additional sequences were added by hand for positions 2006–2037 in Figure 1.

Second, large segments of the fibrinogen family have undergone substantial amounts of divergent evolution, making the precise placement of gaps impossible by automated methods. The multiple alignment was therefore adjusted by hand, at points noted on Figure 1. This manual adjustment followed no objective criteria; in some cases, the adjustment was influenced by the predicted secondary structures. In at least one case,[16] such adjustment was later found to be a source of error in predicting secondary structure, and consideration was given to this possibility here as well.

Experience to date has shown that it is desirable in each prediction to identify secondary structural elements that are not reliably assigned, examine them in detail, and consider alternative assignments. When modeling tertiary structure, both alternatives are considered separately for these elements. This procedure can be followed only if the number of ambiguities is small, of course, as the number of possible structures increases rapidly (2^n for n twofold ambiguities).

In the fibrinogen prediction, several segments are problematic. The first concerns segment 2215–2217, canonically is assigned as a strand. However, Cys 2204 forms a disulfide with position 2220. It is difficult to bring the two cysteines together if they are separated in the polypeptide sequence by a single β strand without the return strand. Further, the conserved tryptophan residues at positions 2215 and 2216 might form protein-protein contacts. Therefore, the coil assignment is preferred for positions 2215–2217. However, the structure must form a type of

Fig. 1. Residue-by-residue secondary structure prediction for fibrinogen. The SIAPrediction assigns positions to the surface (S, s), to the interior (I, i), or to lie near the "active site." Automated output is given, with manual output also noted when different to the right of the automated output. Where the multiple alignment is adjusted, the surface/interior assignments may no longer correspond. Asterisks denote parse positions; residues participating in parsing strings are underlined. Sequences, designated by single letters, are from the SwissProt database, as summarized below. Secondary structure is indicated by E (strong strand assignment), e (weak strand assignment), H (strong helix assignment), and h (weak helix assignment).

a. (P02679) FIBG HUMAN Fibrinogen gamma-A chain precursor. Homo sapiens .
b. 12799) FIBG BOVIN Fibrinogen gamma-B chain precursor (gamma′). *Bos taurus*.
c. (P02680) FIBG RAT Fibrinogen gamma-A and B chain precursor S. *Rattus norvegicus*.
d. (P17634) FIBG XENLA Fibrinogen gamma chain precursor. *Xenopus laevis*.
e. (P04115) FIBG PETMA Fibrinogen gamma chain precursor. *Petromyzon marinus* (lamprey).
f. (Q02020) FIBB CHICK Fibrinogen beta chain precursor (fragment). *Gallus gallus* (chicken).
g. (P02675) FIBB HUMAN Fibrinogen beta chain precursor. *Homo sapiens*.
h. (P14480) FIBB RAT Fibrinogen beta chain precursor (fragments). *Rattus norvegicus*. i
i. (P02676) FIBB BOVIN Fibrinogen beta chain. *Bos taurus*.
j. (P02678) FIBB PETMA Fibrinogen beta chain (fragments). *Petromyzon marinus* (lamprey).
k. (P33573) FIB2 PETMA Fibrinogen alpha-2 chain precursor. *Petromyzon marinus* (lamprey).
l. (P12804) FIBX MOUSE cytotoxic T-lymphocyte specific protein). *Mus musculus* (mouse).
m. (P19477) FIBA PARPA Fibrinogenlike protein A precursor (FREP-A). *Parastichopus parvimensis* (sea cucumber).
n. (P10039; P13132) TENA CHICK Tenascin precursor (TN). *Gallus gallus* (chicken).
o. (P21520) SCA DROME Scabrous protein precursor. *Drosophila melanogaster* (fruit fly).
p. (P24821) TENA HUMAN Tenascin precursor (TN). *Homo sapiens*.
q. (P22105) FIBL HUMAN Fibrinogenlike protein (fragment). *Homo sapiens*.

Pos	protein sequences C jhigf	q	pn	lm	o	k	edabc	SS	SIAPred	Parse	Comments
2006	SSSSS	F	FY	IY	L	E	TTTTT				
2007	GGGGG	P	PP	YP	P	Y	GGGGG		.		*
2008	MKKKR	R	KK	KR	H	I	KKKRK		s		helix to 2018 possible
2009	HEEEE	D	DD	DD	D	D	DDDDD		s		
2010	CCCCC	C	CC	CC	C	C	CCCCC		a		disulfide to Cys 2043
2011	EEEEE	G	SS	SY	S	L	QQQQQ		s		strand 2012-2015 possible
2012	DEKED	E	QQ	DD	E	D	QEDDD		s		coil is preferred to
2013	IIIII	E	AA	HI	V	V	VVIVT		i		accomodate disulfide
2014	YIIIY	M	ML	YL	H	L	VAAAA		i		see text
2015	RRRRR	Q	LL	VQ	T	Q	DNNNN		s		DNGG tetrapeptide parse
2016	NKNKK	N	NN	LS	Q	R	NKKKK		s		*
2017	GGEGG	G	GG	GC	R	G	GGGGG		s		*
2018	GGGGG	A	DE	RS	P	G	GAAAA		s		*
2019	_____	G	TV	_G	_	_	_____		s		**
2020	_____	_	__	__	_	_	_____				***
2021	_____	_	__	__	_	_	_____				***
2022	_____	_	__	_Q	_	_	_____		s		***
2023	REEEE	A	TT	RS	_	K	KRKKK		s		**
2024	TTTTT	S	SS	SP	T	A	DLQEE		s		* SPPSG pentapeptide parse
2025	SSSSS	R	GG	SP	D	S	SSSSS		s		*
2026	EEEEE	T	LL	GS	G	G	GGGGG		s		
2027	AMMMM	S	YY	AG	L	L	LLLLL	E	i		interior strand
2028	YYYYY	T	TT	YQ	H	Y	YYYYY	E	i		core
2029	YLLLI	I	II	RY	L	E	YYFFS	E	i		
2030	IIIII	F	YY	VY	I	V	IIIII	E	i		
2031	QQQQQ	L	LL	TI	A	R	KKKRR	E	s		
2032	PPPPP	N	NN	PQ	P	P	PPPPP		s		* PDSS tetrapeptide parse
2033	DDEDD	G	GG	DP	A	R	LLLLL		.		**
2034	LTDSP	N	DD	HD	G	G	KKKKK		s		*
2035	FSSSF	R	KR	RG	Q	A	AAAAA		s		* 4 consecutive surface residues
2036	SSSVT	E	AT	NG	R	K	KKNKT		s		NSS tripeptide parse
2037	EKKKT	R	QQ	SN	H	R	QQQKE	h	S		
2038	PPPPP	P	AP	SL	P	A	PQQQQ	h	i		
2039	YYYYY	L	LL	FI	L	L	FFFFS	he	I		strand
2040	KRRRR	N	EQ	EK	M	T	LLLLL	he	iS		strand note possible short helix
2041	VVVVV	V	VV	VV	T	V	VVVVV	he	I		interior strand
2042	FYYYY	F	FF	YY	H	H	FYYYY	he	I		
2043	CCCCC	C	CC	CC	C	C	CCCCC	he	A		disulfide to Cys 2010
2044	DDDDD	D	DD	DD	T	E	EEEEE	he	s		
2045	MMMMM	M	MM	MM	A	Q	IIIIT	he	I		
2046	EKKNE	E	TA	EE	D	D	EEDDD	h	S		DGPGNG hexapeptide parse
2047	STTTT	T	SE	TT	_	T	_PGGG		.		* confirmed by gap
2048	HEEED	D	DD	MD	_	D	NSSSP		S		*
2049	GNKNN	G	GG	GE	_	G	GGGGG		S		** 5 consecutive surface
2050	GGGGG	G	GG	GG	_	G	NSNNN		s		** confirmed in all members
2051	GGGGG	G	GG	GG	G	G	GAGGG		.		* indisuptable parse
2052	WWWWW	W	WW	WW	W	W	WWWWW	E	I		
2053	TTTTT	L	II	TT	T	T	TTTTT	E	I		4 consecutive interior assignments
2054	VVVVL	V	VV	VV	T	L	VVVVE	E	sI		beta strand assignment
2055	VIIII	F	FF	LF	V	V	IIFFF	E	I		
2056	QQQQQ	Q	LL	QQ	Q	Q	QQQQK	E	S		
2057	NNNNN	R	RR	AR	R	Q	HRKKK	E	s		
2058	RRRRR	R	RR	RR	R	R	RRRRR		A		conserved Arg
2059	VQQQQ	M	KQ	LI	F	E	HLLLL		s		
2060	DDDDD	D	NN	DD	D	D	DDDDD		s		* parse, conserved G then surface
2061	GGGGG	G	GG	GG	G	G	GGGGG		.		* DGS tripeptide parsing string
2062	SSSSS	Q	RK	ST	S	S	SSSSS	h	s		* confirmed by following helix
2063	SVLVV	T	EE	TI	A	L	VVVLV	H	I		alpha helix 1 2063-2080
2064	NDDDN	D	ND	NN	D	N	NNDDD	H	S		* very exposed
2065	FFFFF	F	FF	FF	F	F	FFFFF	H	I		hydrophobic contact at positions
2066	AGGGG	W	YY	TY	N	N	THKKL	H	S		2065, 2069, 2072, 2076, 2079
2067	RRRRR	R	QR	RR	R	R	RKKKK	H	S		some subfamilies bent at 2070-2071
2068	DKKKA	D	NN	ES	S	S	DNNNN	H	S		some subfamilies missing final turn
2069	WWWWW	W	WW	WW	W	F	WWWWW	H	I		
2070	NDDDD	E	KK	KS	A	S	VVIII	H	.		
2071	TPPPE	D	AN	DY	D	A	SQQQQ	H	s		
2072	YYYYY	Y	YY	YY	Y	Y	YYYYY	H	I		
2073	KKKKK	A	AV	KQ	A	R	RRKKK	H	s		
2074	AKQQR	H	AA	AT	Q	E	EEEEE	H	S		
2075	EGGGG	G	GG	GG	G	G	GGGGG	H	s		*

Fig. 1a.

```
2076 FFFFF F FF FF F F FFFFF   H   I       *
2077 GGGGG G GG GG G G GGGGG   h   .       *
2078 NNNNN N DD NN A T YYHHH   h   .       *
2079 IIIVI I RP LL P V LLLLL   h   i       *
2080 AAAAA S RK EN G D ASSSS   h   i       *  parsing strings
2081 FTTTK _ __ __ _ G PPPPP       i       *  confirmed by indels
2082 GNNNS _ __ __ _ S TNTTT       .      **
2083 _EAT_ _ __ __ _ G LDGGG       S      **
2084 NDEDG _ __ __ _ H TKTTT       S     ***
2085 GTGGG _ __ __ _ _ ___G_       S     ***  adjusted multiple alignment
2086 KKKKK _ __ __ _ _ ___N_       s     ***
2087 SKKNK _ __ __ _ _ _____       S     ***
2088 IYYYY _ __ __ _ _ _____       i     ***
2089 CCCCC _ __ __ _ _ _____       a     ***
2090 NGGGD _ __ __ _ _ _____       s    ****
2091 ILVLT _ __ __ _ _ _____       i     ***
2092 PPPPP _ __ __ _ _ _____       .     ***  dipeptide PG parse
2093 GGGGG G ED RT G G TTTTT   e   .      **
2094 EEEEE E EE EE E E EEEEE   e   A     conserved Glu
2095 YYYYY F FF FF F L FFFFF   E   I
2096 WWWWW W WW WW W W WWWWW   E   I     three internal assignments
2097 LLLLL L LI LL I L LLLLL   E   I
2098 GGGGG G GG GG G G GGGGG       .       * GNDN tetrapeptde parse
2099 TNNNN N LL NN N L NNNNN       s
2100 KDDDD E DE DD E E EEEEE   h   S     conserved G adjacent to 3 surface
2101 TKRKK A NN KN Q A KKKKK   h   S     helix assignment
2102 VIIII L LL II L M IIIIN   h   i
2103 HSSSS H NH HH H Y HHHHH   h   Is
2104 QQQQQ S KK LY H L LLLLL   h   Is
2105 LLLLL L II LL L L LLIII   h   I
2106 TTTTT T TS TT T A TSSSS   h   .i
2107 KRNRK Q AS KS L H GTTTM   h   S
2108 QIMMI A QQ SQ D E QQQQQ   h   .s    4 consecutive surface assignments
2109 HGGGG G GG KG N D QSSSS   h   s       *  GP dipeptide parse
2110 _____ _ __ __ C _ ATAST       S      **  confirmed by indels
2111 _____ _ __ __ _ _ _IIII       i      **
2112 TPPPP D QQ ED _ _ _PPPP       s      **
2113 QTTTT Y YY MY S S YYYYY   E   I     readjusted alignment
2114 QEKEK S EE IE R T RVAVA   E   S     amphiphilic strand
2115 VLLLV I LL LL L M LMLLL   E   I     on edge of folded structure
2116 LLLLL R RR RR Q R RRRRR   E   is
2117 FIIII V VV IV V V IIVII   E   I
2118 DEEEE D DD DE Q E DEEQQ   E   S
2119 MMMMM L LL LL M L LLLLL   E   I
2120 SEEEE R RR EN Q Q TEEEK   E   S
2121 DDDDD A DD DN D G DDDDD       s       *
2122 WWWWW _ __ FT I W WWWWW       I     confluence of weak parse signals
2123 EKKKN G HR NL Y D ESNNS       S       * indel, tripeptide parses
2124 GGGGG D GG GG D G NNGGG       s       *
2125 SDDDD E EE LN N A TQRRR       S
2126 SKKKK A TT TH V G HKTTT   e   s
2127 VVVVV V AA LY W A RSSSS   E   Is    amphiphilic strand
2128 YKTKS F FY YY V H YTTTT   E   Si
2129 AAAAA A AA AA A A AAAAA   E   I
2130 QHLHL Q VV LK E E DDDDD   E   S
2131 YYYYY Y YY YY Y Y YYYYY   E   I
2132 AGEGG D DD DN K _ GSAAA       S     GG dipeptide parse; single indel
2133 SGGGG S KK QK R T HTMSM       s     database error?
2134 FFFFF F FF FF F V FFFFF   E   I     short edge strand
2135 RTTTT H SS YR Y T KRKKR   E   S
2136 PVVVI V VV VI I L LLVVV   E   I
2137 EQQQH D GG AG S R TGGTG   e   S     GPGSD pentapeptide parse
2138 NTNNN S DD ND S D PSPGP       S       * one strong dipeptide parse
2139 EEEEE A AA ES R D EEEEG       S     one tripeptide parse
2140 AAAAG A KK FF A S SKANS       S     consecutive surface residues
2141 QNNNN E TT LS D K DDDDD       s
2142 GKKKK Y RR KE G G ENKKK   E   S
2143 YYYYY Y YY YY Y Y YYYYY   E   I     amphiphilic strand
2144 RQQQQ R KR RL R A RRRRR   E   s     on edge of fold
2145 LVLIL L LL LL L L LFLLL   E   I
2146 WSSSS H KR HV H Q FTTTT   E   Is
2147 VVVVV L VV IL I V YYYYY   E   I
2148 ENSNS E ED GG A S SAAAA   E   S
2149 DKKKN G GG NA E D MYYYY   e   .s
```

Fig. 1b.

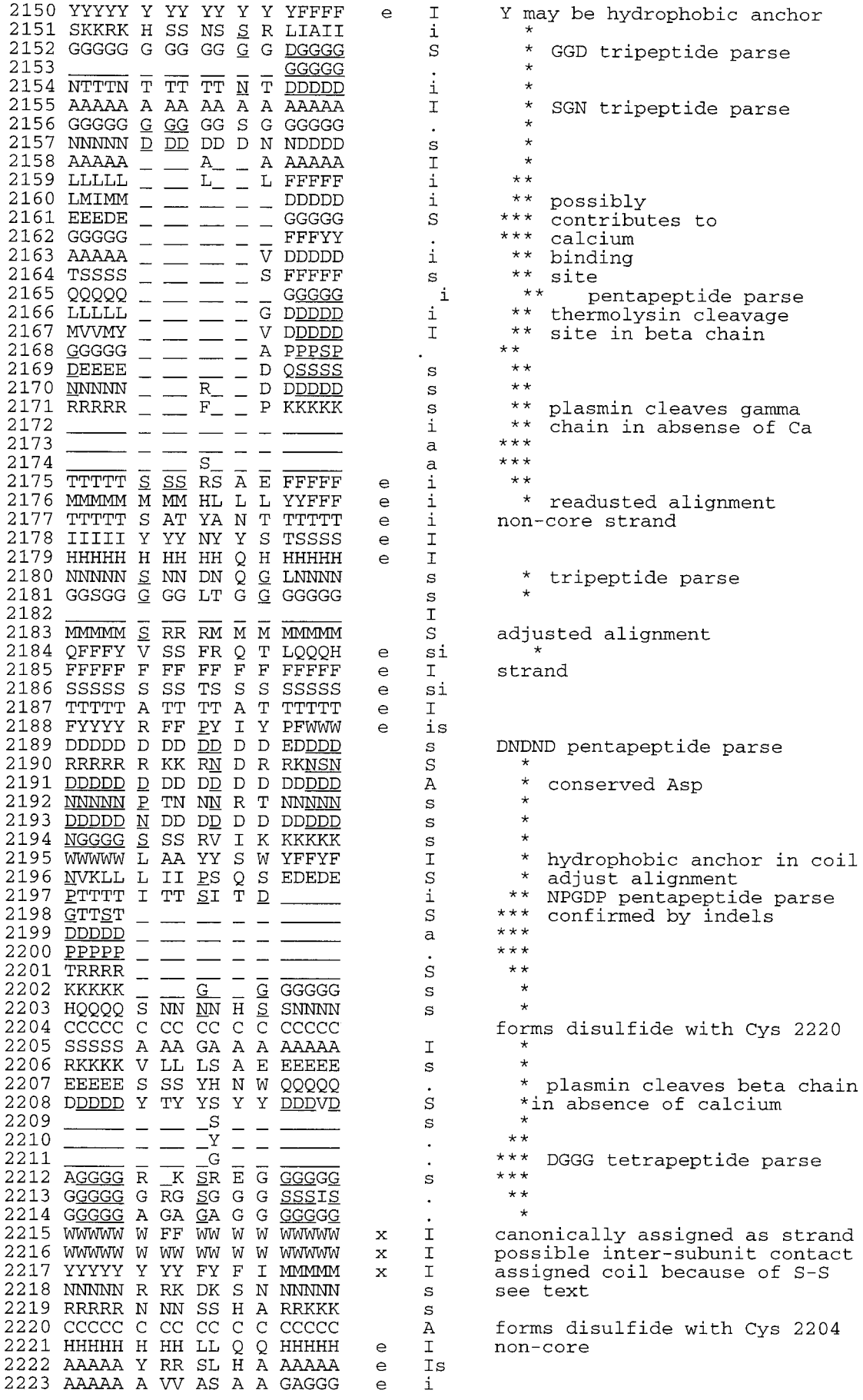
2150 YYYYY Y YY YY Y Y YFFFF e I Y may be hydrophobic anchor
2151 SKKRK H SS NS S R LIAII i *
2152 GGGGG G GG GG G G DGGGG S * GGD tripeptide parse
2153 _____ _ __ __ _ _ GGGGG . *
2154 NTTTN T TT TT N T DDDDD i *
2155 AAAAA A AA AA A A AAAAA I * SGN tripeptide parse
2156 GGGGG G GG GG S G GGGGG . *
2157 NNNNN D DD DD D N NDDDD s *
2158 AAAAA _ __ A_ _ A AAAAA I *
2159 LLLLL _ __ L_ _ L FFFFF i **
2160 LMIMM _ __ __ _ _ DDDDD i ** possibly
2161 EEEDE _ __ __ _ _ GGGGG S *** contributes to
2162 GGGGG _ __ __ _ _ FFFYY . *** calcium
2163 AAAAA _ __ __ _ V DDDDD i ** binding
2164 TSSSS _ __ __ _ S FFFFF s ** site
2165 QQQQQ _ __ __ _ _ GGGGG i ** pentapeptide parse
2166 LLLLL _ __ __ _ G DDDDD i ** thermolysin cleavage
2167 MVVMY _ __ __ _ V DDDDD I ** site in beta chain
2168 GGGGG _ __ __ _ A PPPSP . **
2169 DEEEE _ __ __ _ D QSSSS s **
2170 NNNNN _ __ R_ _ D DDDDD s **
2171 RRRRR _ __ F_ _ P KKKKK s ** plasmin cleaves gamma
2172 _____ _ __ __ _ _ _____ i ** chain in absense of Ca
2173 _____ _ __ __ _ _ _____ a ***
2174 _____ _ __ S_ _ _ _____ a ***
2175 TTTTT S SS RS A E FFFFF e i **
2176 MMMMM M MM HL L L YYFFF e i * readusted alignment
2177 TTTTT S AT YA N T TTTTT e i non-core strand
2178 IIIII Y YY NY Y S TSSSS e I
2179 HHHHH H HH HH Q H HHHHH e I
2180 NNNNN S NN DN Q G LNNNN s * tripeptide parse
2181 GGSGG G GG LT G G GGGGG s *
2182 _____ _ __ __ _ _ _____ I
2183 MMMMM S RR RM M M MMMMM S adjusted alignment
2184 QFFFY V SS FR Q T LQQQH e si *
2185 FFFFF F FF FF F F FFFFF e I strand
2186 SSSSS S SS TS S S SSSSS e si
2187 TTTTT A TT TT A T TTTTT e I
2188 FYYYY R FF PY I Y PFWWW e is
2189 DDDDD D DD DD D D EDDDD s DNDND pentapeptide parse
2190 RRRRR R KK RN D R RKNSN S *
2191 DDDDD D DD DD D D DDDDD A * conserved Asp
2192 NNNNN P TN NN R T NNNNN s *
2193 DDDDD N DD DD D D DDDDD s *
2194 NGGGG S SS RV I K KKKKK s *
2195 WWWWW L AA YY S W YFFYF I * hydrophobic anchor in coil
2196 NVKLL L II PS Q S EDEDE S * adjust alignment
2197 PTTTT I TT SI T D _____ i ** NPGDP pentapeptide parse
2198 GTTST _ __ __ _ _ _____ S *** confirmed by indels
2199 DDDDD _ __ __ _ _ _____ a ***
2200 PPPPP _ __ __ _ _ _____ . ***
2201 TRRRR _ __ __ _ _ _____ S **
2202 KKKKK _ __ G_ _ G GGGGG s *
2203 HQQQQ S NN NN H S SNNNN s *
2204 CCCCC C CC CC C C CCCCC forms disulfide with Cys 2220
2205 SSSSS A AA GA A A AAAAA I *
2206 RKKKK V LL LS A E EEEEE s *
2207 EEEEE S SS YH N W QQQQQ . * plasmin cleaves beta chain
2208 DDDDD Y TY YS Y Y DDDVD S *in absence of calcium
2209 _____ _ __ _S _ _ _____ s *
2210 _____ _ __ _Y _ _ _____ . **
2211 _____ _ __ _G _ _ _____ . *** DGGG tetrapeptide parse
2212 AGGGGG R _K SR E G GGGGG s ***
2213 GGGGGG G RG SG G G SSSIS . **
2214 GGGGG A GA GA G G GGGGG . *
2215 WWWWW W FF WW W W WWWWW x I canonically assigned as strand
2216 WWWWW W WW WW W W WWWWW x I possible inter-subunit contact
2217 YYYYY Y YY FY F I MMMMM x I assigned coil because of S-S
2218 NNNNN R RK DK S N NNNNN s see text
2219 RRRRR N NN SS H A RRKKK s
2220 CCCCC C CC CC C C CCCCC A forms disulfide with Cys 2204
2221 HHHHH H HH LL Q Q HHHHH e I non-core
2222 AAAAA Y RR SL H A AAAAA e Is
2223 AAAAA A VV AS A A GAGGG e i

Fig. 1c.

```
2224 NNNNN N NN NN N N HHHHH       is    NPNG tetrapeptide parse
2225 PPPPP L LL LL L L LLLLL       I       *
2226 NNNNN N MM NN N N NNNNN       s       *
2227 GGGGG G GG GG G G GGGGG       .       *
2228 RRRRR L RR KQ R V KKVVV   e   .s    shifted multiple alignment
2229 YYYYY Y YY YY Y Y YYYYY   e   I     strand
2230 YYYYY G GG YY N Y YYYYY   e   i      GDNN tetrapeptide parse
2231 WWWWW S DD HD L Q FQQQQ   e   Is      *
2232 GGGGG _ __ Q_ G G GGGGG       .      ** GGP tripeptide parse
2233 GGGGG _ __ K_ - G GGGGG       .      **
2234 ILAQT _ __ __ - P NTTTT       S      **
2235 YYYYY _ __ __ - Y YYYYY       i       *
2236 TSTTS _ __ __ - D RSSSS       S       *
2237 KWWWW _ __ __   P KEKKK       s       *  end of a sequence
2238 EDDDD _ __ __   R TAATS       S       *
2239 QMMMM _ __ __   E DD___       i       *  hexapeptide parse
2240 ASAAA _ __ __   K VSSSS       s       *
2241 DKKKK _ __ __   P EGTTT       s      **
2242 _____ _ __ __   _ F____       .     ***
2243 _____ _ __ __   P PPPPP       .     ***  tripeptide parse
2244 YHHHH T NN YY   Y _SNNN       s      **
2245 GGGGG V NN KS   E _GGGG       s       *
2246 TTTTT D HH GG   V YYYYY       i       *
2247 DDDDD H SS VA   E DDDDD       s       *  tripeptide parse
     _____ _ __ RP   _ _____              adjusted alignment
2248 DDDDD Q QQ N_   N DNNNN       s       *
2249 GGGGG G GG GS   G GGGGG   e   .       *
2250 VVVVI V VV II   V IIIII   E   i     consecutive interior
2251 VVVVV S NN FY   V IIIII   E   i     assignments
2252 WWWWW W WW WW   W WWWWW   E   i
2253 MMMMM Y FF GS   A AAAAA   E   i
2254 NNNNN H HH TY   T TTTTT   E   i
2255 WWWWW W WW WL   Y WWWWW   E   i
2256 KKQKK K KK PP   R HRKKK   e   S     PGDND pentapeptide parse
2257 GGGGG G GG GG   G DRTST       S       *
2258 SSSSS F HH ID   S RRRRR       s
2259 WWWWW E EE NN   D WWWWW       i
2260 YYYYY F HY QD   Y YYYYY   e   i
2261 SSSSS S SS AQ   S SSSSS   e   s
2262 MMMMM V II QI   L LMMMM   e   i
2263 RRKRK P QQ PP   K KKKKK   e   S
2264 QRKKK F FF _F   R MSKKE   E   S     bad multiple alignment
2265 MMMMM T AA GA   T TVTTT   E   i     bent at 2263-2264
2266 ASSSS E EE GE   A TTTTT   E   is
2267 MMMMM M MM YM   V MMMMM   E    i
2268 KKKKK K KK KK   R KKKKK   E    s
2269 LIIII L LL SL   - LIIII   E   i
2270 RRRRK R RR SR   - LMIII   E   i
2271 PPPPP P PP FN   - PPPPP       s       *
2272 ----- - -- --   - MLFLF       i       *
2273 ----- - -- --   - GNNNN       s       *
2274 ----- - -- --   - RRRRR       a
2275 ----- - -- --     D____       a       *
2276 ----- - -- --     LYLLL       i       *
2277 ----- - -- --     SGTAS       S       *
2278 ----- - -- --     G_III       i       *
2279 ----- - -- --     HAGGG       s       *
2280 ----- - -- --     GEEED       S       *
2281 ----- - -- --     GGGGG       .       *
2282 ----- - -- --     QQQQQ       a
2283 ----- - -- --     QQQQQ       a
2284 ----- - -- --     Q_HHH       i
2285                   STHQH       S
2286                   KLLLM       i
2287                   GGGGG       .       *
2288                   NGGGG       s       *
2289                   SSAAS       .       *
2290                   RKKKK       s
2291                   --QQQ       a
2292                   - AV        .
2293                     GG        .
```

Fig. 1d.

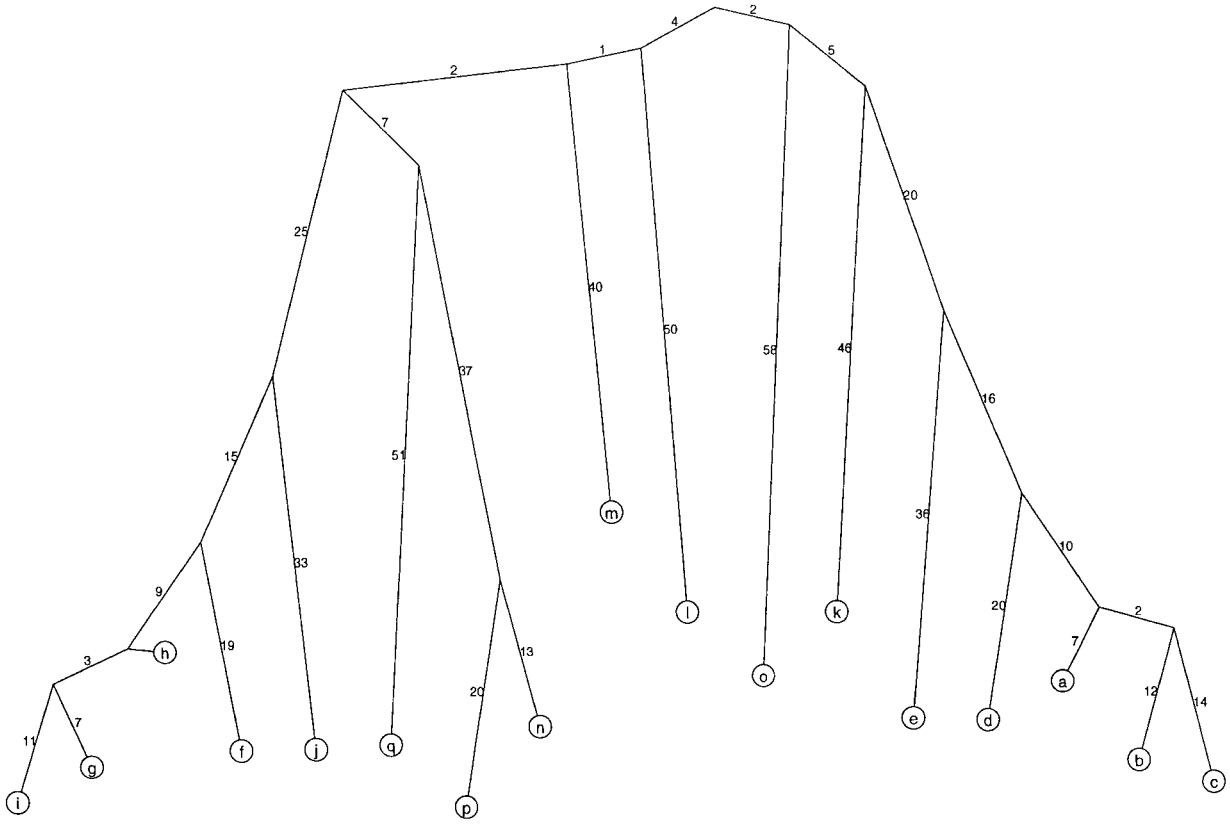

Fig. 2. Evolutionary tree interrelating protein sequences used in this work (numbers indicate evolutionary distance in PAM units).

hairpin, which may be assigned β structure by at least some secondary structure assignment programs.

Segment 2126–2137 is problematic to assign because a single residue gap in a single protein in the family disrupts the multiple alignment. This gap is difficult to align due to substantial sequence divergence in the family. DARWIN aligns the gap with a G that is part of a GG dipeptide at positions 2132–2133. This is a weak dipeptide parse. If the gap is accepted as a parse, a strand is assigned to the first part of this segment (positions 2126–2131), and a second strand is assigned to the second part of the segment (positions 2134–2137). The segment has been assigned as two β strands, but might be regarded in tertiary structure modeling as a single unit.

Finally, the segment comprising positions 2037–2046 is assigned as a helix, but with an alternative strand a possibility. The helix is assigned provided that Cys 2043, which forms a disulfide bond, is at the surface-interior interface. Here, both alternative secondary structures need to be considered when modeling tertiary structure, and both are listed in Figure 1. The need to bury other strands in the structure in particular, the strand before it and the two strands following it, has created a need for an additional helix in this domain. Therefore, the helix conformation is preferred in this modeling.

TERTIARY STRUCTURAL MODELING

It is appropriate in light of the secondary model predicted here to speculate on possible supersecondary and tertiary molecules that are built from the predicted secondary structural elements. Indeed, to date, most of the secondary structure predictions made in Zurich have been accompanied by at least some supersecondary structural modeling.[16] Again, the core fold is modeled most productively.

An interesting but controversial approach to assembling secondary structural elements involves the search for compensatory covariation, substitutions at pairs of positions distant in the sequence that appear to be compensatory. The first time compensatory covariation analysis was used in a bona fide prediction setting was, we believe, in the protein kinase prediction.[28] In this family, LLPLRRR at position 87 was matched with QQQQEEE at position 108 (alignment numbering). This led the prediction to suggest that these side chains were in contact, which imposed a long distance constraint on the fold that required two β strands to lie antiparallel. When

TABLE I. Secondary Structure Assignments in the C-Terminal Domain of the Beta and Gamma Chains of the C-Terminal Fragment of Fibrinogen

Unit	Positions	Comments
Beginning of multiple alignment for some family members		
Position	2010	Cys forming disulfide with Cys 2043
Segment	2011–2014	Edge strand, short helix, ambiguous, not core, ignored in model
Parse	2015–2026	DNGG, PPSG tetrapeptide parses
Strand	2027–2031	May be extended in some members
Parse	2032–2037	PDGGN, NSS, and NGN parsing strings, reliable
Beginning of reliable multiple alignment over all family members		
Helix	2037–2046	Strand is alternative, see text
Position	2043	Cys forming disulfide to Cys 2010
Parse	2047–2051	GSGNG, GPGNG, reliable
Strand	2052–2057	2052–2055, four consecutive internal positions
Position	2058	Conserved Arg
Parse	2059–2062	Weaker parse, DGS tripeptide parse, start of helix possible 2062
Helix	2063–2080	Highly reliable, last turn 2078–2081 weak
Parse	2081–2092	PGG, SP, PG parsing strings, confirmed by gap
Strand	2093–2097	4 consecutive interiors, segment may extend next helix (see text)
Parse	2098–2099	GNDN tetrapeptide parse
Helix	2100–2109	See text for discussion
Parse	2110–2112	GP dipeptide parse confirmed by gap
Strand	2113–2120	Amphiphilic strand
Parse	2121–2125	Tripeptide parses, confirmed by gap, 4 consecutive surface positions
Strand	2126–2131	Issue of following parse, see text
Parse	2132–2133	Weak GG dipeptide parse, may fuse strand before and after
Strand	2134–2137	Issue of preceding parse, see text
Parse	2138–2141	GPGSD pentapeptide parse, 6 consecutive surface residues
Strand	2142–2150	Amphiphilic strand, 2150 may be hydrophobic anchor
Parse	2151–2174	GDS, DDPSD parses, gaps, possible Ca ligands
Strand	2175–2179	5 consecutive interior, noncore, bad alignment
Parse	2180–2183	SGS tripeptide parse, confirmed by gap
Strand	2184–2188	Largely, but not entirely, buried strand
Position	2189	Conserved Asp, Ca binding
Parse	2190–2194	DNDND pentapeptide parse, Ca-binding loop?
Position	2195	Possible hydrophobic anchor of a loop
Parse	2196–2203	NPGDP pentapeptide parse
Position	2204	Cys forming disulfide with Cys 2220
Parse	2205–2214	DGGG tetrapeptide parse, confirmed by gap, assigned hairpin
Segment	2215–2219	Canonical strand 2215–2217; hairpin because of disulfide, see text
Position	2220	Cys forming disulfide with Cys 2204
Strand	2221–2223	Noncore
Parse	2224–2227	NPNG tetrapeptide parse
Strand	2228–2231	Multiple alignment bad, possible noncore strand
End of coherent multiple alignment with distant homologs		
Parse	2232–2248	A variety of parsing strings confirmed by gaps
Strand	2249–2256	Buried strand
Parse	2257–2259	PGDND parsing string
Strand	2260–2270	Multiple alignment bad, see text

the crystal structure of a representative protein kinase was ultimately solved, it was found that positions 87 and 108 were in fact in contact, and that the two strands were indeed antiparallel. The post hoc analysis pointed out that one reason compensatory covariation was so successful in this case was because the side chains were largely buried in the structure.

Since this initial use of covariation analysis, several papers have examined the overall statistics of the approach.[29–33] In general, it is agreed that a compensatory covariation signal is present, but weak, during divergent evolution of protein sequences under functional constraints. Much discussion remains as to whether such a weak signal is useful in a bona fide prediction setting. With the exception of Chelvanayagam and colleagues,[33] none of this discussion has centered on instances where compensatory covariation analysis has been used productively in a bona fide prediction setting.

In the protein kinase prediction, the weak compensatory covariation signal was identified because of its context. The possibility of two secondary structural elements lying antiparallel was recognized. This constrained the search for compensation to a small number of pairs of positions. Further, it was recognized that compensatory variation should be sought within strict guidelines of evolutionary distance, and that charge compensation was likely to persist for longer evolutionary distances than other types of covariation.

It is clear that this sort of analysis is ad hoc, and extremely difficult to test in any but a bona fide prediction setting. Thus, we have experimented with compensatory covariation analysis in the fibrinogen prediction reported here.

For example, segment (2027–2031) and segment (2037–2046) might either lie adjacent or not. An intriguing charge variation is observed within subfamily jhigf at position 2023 (REEEE) and position 2046 (EKKNE). This change is compensatory in the first two proteins of the subfamily, and neutral elsewhere. These residues are on the surface of the folded structure, and are flanked on one (position 2046) or both (position 2023) sides by surface positions. Thus, we interpret this as normal variation within the family at surface positions, variation that need not reflect proximity in the side chains.

The RY variation at position 2029 in subfamily lm is not, however, likely to be on the surface. This variation is embedded within an internal segment, and is more likely to be compensated for this reason. The fact that proteins l and m have diverged 91 PAM units requires that only charge compensation be examined.[33] If the strand is antiparallel and adjacent in the sheet to the following strand, compensatory covariation might be able to be observed in the second segment. Indeed, at position 2040, an EK substitution is observed. Therefore, this compensatory covariation may indicate an antiparallel orientation of segments 2027–2031 and 2037–2046.

The following strand (2052–2057) also has some intriguing charge variation in internal segments. For example, family edabc has residues VVVVE at position 2054, and residues QQQQK at position 2056. The PAM distance between proteins b and c is quite low (only 25 PAM units), making this a strong case for compensation. Here, the compensatory covariation does not allow us to detect long distance contacts; it is almost certainly the case that the compensation is between residues i and $i+2$ in a strand. However, the compensatory covariation is useful because it allows us to confirm the hypothesis that segment 2052–2057 adopts a β strand conformation as a secondary structure or, more precisely, that the side chains of positions 2054 and 2056 are in proximity.

Further, this provides an interesting case where secondary structural assignments allow us to reconsider the surface-interior assignments made from analysis of sequence data alone. The automated computer program implemented in DARWIN assigns both positions 2054 and 2056 to the surface. Upon inspection, however, it is clear that these positions depend heavily on the appearance of a Glu in this subfamily at position 2054 and a Lys at position 2056. If these are in fact internally compensatory, the positions themselves are not as likely to be on the surface. This is illustrative of a general rule that secondary structure models, although assembled from sequential models, should be used to reevaluate the sequential information, just as tertiary structure models, assembled from secondary structural models, should be used to reevaluate the secondary structural models.

Finally, this allows us to make a comment on the role of abundant sequences to structure predictions from multiple alignments. We noted some time ago that the more sequences, the better. Recently, di Francesco suggested that this might not be generally the case.[34] Clearly, additional sequences provide additional information, something that is always useful, provided that the analytical tools are constructed to handle the additional information correctly. Here, it is clear that if the database happened not to contain protein c, then the analysis would not be possible. Positions 2054 and 2056 would be normal interior positions.

Relevant to the tertiary structural modeling is the fact that strands 2027–2031, 2052–2057, and 2093–2097 must be buried in the structure. The assignment of secondary structure to the segment around position 2040 is ambiguous; it can either be a short helix or a somewhat exposed strand. We must now consider how best to use this segment to bury the segments that are almost certainly buried strands. To do this, we must consider first the domain structure in this protein.

The γ chain of fibrinogen is cleaved by plasmin following position 2171 in the absence of calcium, and a domain boundary is believed to occur near here. If this is the case, the first domain in this model must be completed by three β segments, strand 2113–2120, strand 2126–2137 (interrupted at positions 2132–2133), and strand 2142–2150. The first and third are canonically amphiphilic, almost text-

book in extent. Thus, it is appropriate to assemble these into an antiparallel β sheet, and to use this sheet to bury secondary structural elements that precede it in the domain, in particular, strands 2027–2031, 2052–2057, and 2093–2097, in a sandwich structure. Two alternative β meanders are conceivable, depending on whether segment 2126–2137 is treated as one strand or two. In this model, strands 2027–2031, 2052–2057, and 2093–2097 form the core of the first domain of the C-terminal fragment.

What then buries the other side of the sheet formed by strands 2027–2031, 2052–2057, and 2093–2097? Clearly, helices 2063–2075 and 2100–2109 are available, the first connecting strand 2052–2057 to strand 2093–2097, the second connecting strand 2093–2097 to the amphiphilic sheet. If the second helix is indeed a connecting helix, it will do little to bury these strands, in particular, strand 2027–2031. Additional material is needed. If the ambiguous segment is assigned as a helix (positions 2037–2046), it can help bury the hypothetical core sheet. For this reason, the secondary structure in Figure 1 is preferred, and a specific tertiary structural model follows. This ends us with a three-strand parallel sheet. This might require that an additional β unit be obtained from positions preceding position 2027. The alignment is poor, however, making this difficult to assign.

The second segment of the fibrinogen fragment considered here is assigned entirely a β structure. The β strands in this region are both amphiphilic and internal. Many come in segments where the multiple alignment must be adjusted by hand. These presumably form an all β barrel or sandwich structure as well, perhaps a six-stranded Greek key structure as found in serine proteases, but time is inadequate to build a comprehensive model.

Since this prediction was prepared, we realized that Russell Doolittle prepared some time ago a prediction of the structure of fibrinogen.[35] Doolittle applied a variety of methods, including an analysis similar to that used here.[28] Much of Doolittle's prediction corresponds to the prediction reported here, and where the prediction disagrees, it is often in regions where the multiple alignments are difficult to construct.

ACKNOWLEGMENTS

NIH (GM 39900) to F.E.C. and a postdoctoral fellowship of the Leukemia Society (D.L.G.)

REFERENCES

1. Benner, S.A. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv. Enzymol. Reg. 28:219–236, 1989.
2. Pananoyotou, G., Bax, B., Gout, I., Federwisch, M., Wroblowski, B., Dhand, R., Fry, M.J., Blundell, T.L., Wollmer, A., Waterfield, M.D. Interaction of the p85 subunit of PI 3-kinase and its N-terminal SH2 domain with a PDGF receptor phosphorylation site. EMBO J. 11:4261–4272, 1992.
3. Russell, R.B., Breed, J., Barton, G.J. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. FEBS Lett. 304;1520, 1992.
4. Musacchio, A., Gibson, T., Lehto, V.-P., Saraste, M. SH3: An abundant protein domain in search of a function. FEBS Lett. 307:55–61, 1992.
5. Bazan, J.F. Structural design and molecular evolution of a cytokine receptor superfamily. Proc. Natl. Acad. Sci. USA 87:6934–6937, 1990.
6. Moe, G.R., Koshland Jr., D.E. Transmembrane signalling through the asparte receptor. In "Microbial Energy Transduction, Genetics, Structure and Function of Membrane Proteins." Youvan, D.C., Daldal, F. (eds.). Cold Spring Harbor, NY: Cold Spring Harbor Press, 1986:163–168.
7. Benner, S.A., Gerloff, D.L., Jenny, T.F. Predicting protein crystal structures. Science 265:1642–1644, 1994.
8. Benner, S.A., Ellington, A.D. Evolution and structural theory: The frontier between chemistry and biochemistry. Bioorg. Chem. Front. 1:1–70, 1990.
9. Benner, S.A. Predicting the conformation of proteins from sequence data. In "Protein Engineering: A Guide to Design and Production." Craik, C.S., Cleland, J. (eds.). 1996:71–99.
10. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826.
11. Benner, S.A., Chelvanayagam, G., Turcotte, M. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Rev., submitted, 1996.
12. Jenny, T.F., Benner, S.A. Evaluating predictions of secondary structure in proteins. Biochem. Biophys. Res. Commun. 200:149–155, 1994.
13. Jenny, T.F., Benner, S.A. A prediction of the secondary structure of the pleckstrin homology domain. Proteins 20:1–4, 1994.
14. Musacchio, A., Gibson, T., Rice, P., Thompson, J., Sarasate M. The PH domain: A common piece in the structural patchwork of signaling proteins. Trends Biochem. Sci. 18:343–348, 1993.
15. Gerloff, D.L., Jenny, T.F., Knecht, L.J., Benner, S.A. A secondary structure prediction of the hemorrhagic metalloprotease family. Biochem. Biophys. Res. Commun. 194:560–565, 1993.
16. Benner, S.A., Gerloff, D.L., Chelvanayagam, G. The phospho-beta-galactosidase and synaptotagmin predictions. Proteins 23:446–453, 1995.
17. Bazan, J.F. Helix fold prediction for the cyclin box. Proteins 24:18–34, 1996.
18. Edwards, Y.J.K., Perkins, S.J. The protein fold of the von Willebrand factor type A domain is predicted to be similar to the open twisted beta-sheet flanked by alpha helices found in human Ras-p21. FEBS Lett. 358:283–286, 1995.
19. Jenny, T.F., Gerloff, D.L., Cohen, M.A., Benner, S.A. Predicted secondary and supersecondary structure for the serine/threonine specific protein phosphatase family. Proteins 21:1–10, 1995.
20. Livinston, C.D., Barton, G.J. Secondary structure prediction from multiple sequence data: Blood clotting factor XIII and *Yersinia* protein-tyrosine phosphatase. Int. J. Peptide Protein Res. 44:239–244, 1994.
21. Lupas, A., Koster, A.J., Walz, J., Baumeister, W. Predicted seconday structure of the 20 S proteasome and model structure of the putative peptide channel. FEBS Lett. 354:45–49, 1994.
22. Doolittle, R.F. Fibrinogen and fibrin. Annu. Rev. Biochem. 53:195–229, 1984.
23. Weisel, J.W., Stauffacher, C.V., Bullitt, E., Cohen, C. A model for fibrinogen: Domains and sequence. Science 230: 1388–1391, 1985.
24. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 20:2019–2022, 1992.

25. Gonnet, G.H., Benner, S.A. Computational biochemistry research at ETH. Tech. Rep. Dept. Inform. 154, 1991.
26. Gonnet, G.H., Cohen, M.A., Benner, S.A. Exhaustive matching of the entire protein sequence dtabase. Science 256: 1443–1445, 1992.
27. Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona fide prediction of aspects of protein conformation: Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. J. Mol. Biol. 235:926–958, 1994.
28. Benner, S.A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: The catalytic domain of protein kinases. Adv. Enzymol. Reg. 31:121–181, 191.
29. Gobel, U., Sander, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.
30. Neher, E. How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. USA 91:98–102, 1994.
31. Shindyalov, I.N., Kolchanov, N.A., Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 7:349–358, 1994.
32. Taylor, W.R., Hatrick, K.L. Compensating changes in protein multiple sequence alignments. Protein Eng. 7:341–348, 1994.
33. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G.H., Benner, S.A. An analysis of simultaneous vriation in protein structures. Proteins, submitted.
34. di Francesco, V., Garnier, J., Munson, P.J. Improving protein secondary structure prediction with aligned homologous sequences. Protein Sci. 5:106–113, 1996.
35. Doolittle, R.F. A detailed consideration of a principal domain of vertebrate fibrinogen and its relatives. Protein Sci. 1:1563–1577, 1992.