

Structure Prediction in a Post-genomic Environment: A Secondary and Tertiary Structural Model for the Initiation Factor 5A Family

Dietlind L. Gerloff,* Marcin Joachimiak,*† Fred E. Cohen,*‡ Gina M. Cannarozzi,§
Stephen G. Chamberlin,[¶] and Steven A. Benner§^{||}1

*Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143;

†Graduate Group in Biophysics, University of California, San Francisco, California 94143; ‡Department of Medicine,

University of California, San Francisco, California 94143; §Department of Chemistry, University of Florida,

Gainesville, Florida 32611; ¶Sulfonics, Inc., 12085 Research Drive, Alachua, Florida 32615; and

||Department of Anatomy and Cell Biology, University of Florida, Gainesville, Florida 32611

Received July 15, 1998

Two predictions have been prepared for the fold of initiation factor 5A (IF5A) starting from a set of homologous sequences. In the first, a secondary structural model was predicted for the protein in 1994, when only eleven homologs (and no eubacterial homologs) had been sequenced. The second was made recently, after genome projects had generated a total of 33 sequences for the protein family from species of all three kingdoms of life. With the second set of sequences, but not with the first, it was possible to predict that the N-terminal domain of the protein folds in a possibly open beta-barrel/sandwich core structure, with a short helix capping one side of the barrel. We place the pair of predictions in the public domain before an experimental structure is known. This example illustrates the impact of genome sequencing projects on structure prediction from sequence alignments. © 1998

Academic Press

Key Words: initiation factor 5A (IF5A); hypusine-containing protein; CASP3 prediction experiment; protein structure prediction; genome projects; beta-barrel fold.

Predictions of protein structure made and announced before an experimental structure is available have shown that useful predictions of secondary structure can be obtained from a set of aligned sequences of homologous proteins diverging under functional constraints (1). Especially important in this context have been predictions made in the project known as “Critical Assessment of Techniques for Structure Prediction” (CASP), now in its

third round (URL: <http://PredictionCenter.llnl.gov/>). In former rounds of this project, we used secondary structural predictions to identify the 8-fold alpha-beta barrel fold of phospho-beta-galactosidase (2), identify the beta-sandwich of synaptotagmin as one of three alternative folds (3), and build a tertiary structure model for the heat shock protein 90 family that predicted that it was a distant evolutionary relative of the DNA gyrases (4), *inter alia*. The quality of a prediction of a protein fold made from a set of aligned homologous protein sequences depends on the number of sequences in the alignment, the extent to which these have diverged, their relationship on an evolutionary tree, and the extent to which function has been conserved within the family (1). Such parameters are difficult to quantitate by automated tools for assessing the quality of a prediction. They have therefore been frequently overlooked in the process of evaluating prediction projects within the CASP framework (1). Understanding how the quality of the prediction “output” depends on the nature of the sequence “input” will nevertheless be central to efforts to rationally improve prediction methodology.

Initiation factor 5A (IF5A) is a widely conserved protein that is post-translationally modified on one lysine to incorporate the unusual amino acid hypusine (5, 6). Hypusinylation is known in archaea and eukaryotes, but not in eubacteria, where the two necessary enzymes, deoxyhypusine synthase (7) and hydrolase, appear to be missing. This may imply that the modification was either developed after eubacteria diverged from archaea and eukaryotes, or was present in the universal ancestor and lost selectively in eubacterial lineages (8).

Initiation factor 5A is best known as a monomer, although dimeric and higher oligomers might be formed as well (9). Deletion studies suggest that a core

¹ Corresponding author. Fax: (352) 846 2580. E-mail: benner@chem.ufl.edu.

segment extending from F30 to D70 is required for human IF5A to be recognized by deoxyhypusine synthase; this may represent a core folding unit (10). While original assay of the protein was based on its ability to stimulate the formation of the first peptide bond in protein biosynthesis, its mechanism of action is now clearly more complicated (11, 12). For example, Ruhl et al. showed that IF5A binds to Rev, a nuclear phosphoprotein that accumulates in the nucleoli of cells that are expressing RNA molecules from HIV carrying the "Rev response element" (RRE). These messages are retained in the nucleus and appear in the cytoplasm only when Rev is present. IF5A appears to be necessary for Rev to function in mammalian cells (13, 14), and its expression is significantly increased in T-lymphocytes when they are activated (15). Liu et al. showed that IF5A binds to RRE in gel shift assays (16).

As a prediction target, IF5A was first encountered in 1994, when the protein was considered as a potential therapeutic target. At that time, only ten IF5A homologs with clearly alignable sequences were available (Figure 1), and these were not widely distributed on the evolutionary tree (Figure 2). Nevertheless, predictions were prepared using both transparent prediction tools (1) and the neural network program PHD (17) in the version then available by server. In the first prediction, a key element of secondary structure was ambiguously assigned as either a long strand or as an internal helix. Tertiary structure models were built using both alternatives. These predictions were published as supplementary information to the structural characterization of different isoforms of human IF5A (18). However, because no experimental structure has become available in the meantime, the accuracy of the predictions could not be assessed.

Very recently, a homolog to IF5A, translation factor 5A from the archaeobacterium *Pyrobaculum aerophilum*, was announced as a target for the CASP3 project (19) by T. Peat (Los Alamos National Laboratory), implying that a crystal structure would shortly emerge. In the intervening time, genome sequencing projects have made available a total of 33 homologs in the superfamily. In particular, members of a protein family in the eubacterial

kingdom, the eubacterial elongation factors P (EF-P), display significant sequence similarity with the N-terminal 90 residues of the hypusine-containing proteins (Figure 1). Thus, IF5A provides an opportunity to compare the impact of the systematic development of sequence databases on protein structure prediction in a *bona fide* prediction setting. This comparison makes evident how valuable genome projects, especially those selected from organisms widely dispersed on the universal tree of life, are in generating predictions.

MATERIALS AND METHODS

The multiple sequence alignment shown in Figure 1 was prepared using PileUp (Genetics Computer Group, (20)). Sequences marked with × were available in 1994. Positions in the multiple alignment predicted to lie on the surface, in the interior, in the active site and in parses in the protein fold were assigned (1) using the DARWIN tool available via server (21), using phylogenetic trees based on global pairwise alignments. Secondary structure predictions were made manually from these assignments following the procedures recently reviewed (1). In addition, PHD (17) neural network predictions were obtained via a server featuring the program in the version available at the time the predictions were obtained (22).

Maximum likelihood trees were prepared using the DARWIN server. To obtain reliable evolutionary distances (expressed as PAM units, the number of point accepted mutations per 100 amino acids), the IF5A family was first divided into subfamilies where each protein in a subfamily is essentially the same length. These subfamilies corresponded to the three major kingdoms of life. A maximum likelihood Darwin tree was built for each subfamily, comparing sequence fragments pairwise over the length of the corresponding multiple subalignments. This procedure yielded the PAM distances for each subtree. The sequences of all IF5A proteins were then truncated to a common core alignable over all kingdoms, and a second maximum likelihood tree was built. From this tree, PAM distances were extracted for the edges of the tree that join the subfamilies. The resulting tree is shown in Figure 2, with the sequences available in 1994 highlighted.

Tertiary structure modelling for the putative N-terminal domain followed analyses used for the prediction of the tertiary fold of protein kinase (23), synaptotagmin, and phospho-beta-galactosidase (24), aided by submissions to fold recognition servers on the World Wide Web (see below).

RESULTS AND DISCUSSION

A secondary structure model was generated in 1994 for IF5A from the 10 sequences available at the time

FIG. 1. Multiple sequence alignment and secondary structure predictions for the initiation factor 5A protein superfamily. Subalignments for each of the three primary kingdoms (eubacteria, eukaryotes, and archaea) are shown aligned in a master alignment, generated using PileUp (GCG Wisconsin Package) and edited manually. SIAP shows surface, interior, active site assignments made separately for each kingdom: S/s = strong/weak surface, I/i = strong/weak interior, A/a = strong/weak active site. Parse assignments indicate break in secondary structure, indicated by numbers 1-5 in order of decreasing reliability. Secondary structure assignments, H/h = strong/weak helix, E/e = strong/weak strand are made by experts (ETH.94 made by method in (1) with sequences marked with ×; DLG.Eub, DLG.Euk, and DLG.Arc, made by Gerloff on the eubacterial, eukaryotic, and archaeal sequences respectively; SAB.98 made by Benner for the superfamily as a whole) or the PHD neural network made through the PredictProtein-server (22) (PHD.94 in 1994 submitting sequence 11; PHD.1 in 1998 submitting sequence 33; PHD.2 in 1998 through the UCLA-DOE Fold Recognition-server (31) submitting sequence 33). ETH.94 had an ambiguous assignment around Ali# 050 (18). Sequences were extracted from publicly available databases. Sequences are: 1: efp_mycty, 2: efp_syny3, 3: efp_synp7, 4: efp_anasp, 5: efp_bacsu, 6: aquae950 *Aquifex aeolicus*, 7: efp_helyp, 8: bbur213 *Borrelia burgdorferi*, 9: if52_chick, 10: if5a_rabit, 11: if5a_human, 12: 2119941 *Homo sapiens*, 13: if51_chick, 14: if5a_dicdi, 15: if52_caeel, 16: if51_caeel, 17: if5a_neucr, 18: if52_yeast, 19: if51_yeast, 20: 3024014 *Solanum tuberosum*, 21: 3024018 *Zea mays*, 22: 3024019 *Solanum tuberosum*, 23: g3024022 *Solanum tuberosum*, 24: if51_nicpl, 25: if52_nicpl, 26: if5a_medsa, 27: 124230 *Nicotiana glauca*, 28: afu634, *Archaeoglobus fulgidus*, 29: 2696455 *Pyrococcus horikoshii*, 30: mthe858 *Methanobacterium thermoautotrophicum*, 31: if5a_metja, 32: if5a_sulac, 33: CASP3-T0063 *Pyrobaculum aerophilum*.

```

Eukaryotes
¥ 9 xxxQCSALRKNGFVVLKGRPCKXXXSTSKTGKHGHAKVHLVGIDIF
¥10 ADDLDFET_GDAGASATFPMQCSALRKNGFVVLKGRPCKIVEMSTSKTGKHGHAKVHLVGIDIF
¥11 ADDLDFET_GDAGASATFPMQCSALRKNGFVVLKGRPCKIVEMSTSKTGKHGHAKVHLVGIDIF
12 MADDLDFET_GDAGASATFPMQCSALRKNGFVVLKGRPCKIVEMASKTGKHGHAKVHLVGIDIF
13 MADELDFTT_GDAGASSTYPMQCSALRKNGFVVLKGRPCKIVEMSTSKTGKHGHAKVHLVGIDIF
¥14 MKPLIMEYNKMSDNEALDVEDYAQ_ASGASLTFPIQCSALRKNGFVVIKGFPCIKIVDMSTSKTGKHGHAKVNITAIDIF
15 MSDDHHDEHFHT_GDSGAAATFPKQCSALRKNNEHVMIKGRPCKIVEMSTSKTGKHGHAKVHMAIDIF
16 MSDDHHDEEQFDS_AESGAAATFPKQCSALRKNNEHVMIRGRPCKIVEMSTSKTGKHGHAKVHMAIDIF
¥17 MSAAHDDAQHEHTFDS_VDAGRSATYPMQCSALRKGHHVVIKNRPCIKIVDMSTSKTGKHGHAKVHLVAIDIF
¥18 SDEEHT_FET_ADAGSSATYPMQCSALRKNGFVVIKSRPCIKIVDMSTSKTGKHGHAKVHLVAIDIF
¥19 SDEEHT_FEN_ADAGASATYPMQCSALRKNGFVVIKGRPCKIVDMSTSKTGKHGHAKVHLVTLIDIF
20 MAEEHHVDFEG_GEAGASLTFPMQCSALRKNHVVVIKGRPCKIVDMSTSKTGKHGHAKVHIVALDIF
21 MSDSEHHHFESKADAGASKTYPQQAGTVRKNGFIVIKNRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
22 MSD_EEHHHFESKADAGASKTYPQQAGTIRKSGYIVIKGRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
23 MSD_EEHHHFESKADAGASKTYPQQAGTIRKSGHIVIKNRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
¥24 MSD_EEHHHFESKADAGASKTYPQQAGTIRKNGHIVIKNRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
¥25 MSD_EEHQFESKADAGASKTYPQQAGTIRKNGHIVIKGRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
¥26 MSD_EEHQFESKADAGASKTYPQQAGTIRKNGYIVIKNRPCKVVEVSTSKTGKHGHAKCHFVAIDIF
27 MSD_EEHQFESKADAGASKTYPQQAGTIRKNGYIVIKNRPCKVVEVSTSKTGKHGHAKCHFVAIDIF

SIA .....ssiiSSssisSaiSi..isai.sAI.IIAAssiIIIssS.AAiiSiASAAA.AA.AIAiII.IAII
parse 3211 111111 1 1 111

DLG.Euk eeeeeeeee eeeee eeeee eeeeeAAAAAAAAA eeeeeeee

PHD.94: ....._.....HHHHHHHH...EEEE...EEEEEE.....EEEEEEEEEE.
ETH.94: ....._..EEEEEE...EEEE...EEEE...EEEEEE.....hhhhhhhhhhh
                                         eeeeeeeeeee

```

175

Ali#	060	070	080	090	100	110	120	130
<i>Eubacteria</i>								
1	SGKVVDKTFN	AGVKVDATATVDRD	TTYLYRDGSD	FV_FMDSDQDYE_QHPLP	_____	EALVGDAARFLLEGMPVQVAFHNGV		
2	TGNVVEKTFR	AGETVPQANIEKSV	MQHTYKDGQYV	FMDMETFE_EVSIA	_____	PTDLGDKAKFIKEEMSVSVTWDGT		
3	TGNVVEKTFR	AGETVPQAVLEKST	LQYTYKGDGDFV	FMDMETYE_EGRLT	_____	AATIGDRVVKYLKEGMEANVITWNGQ		
4	SGKVLEKTFR	AGETVPQATLEKIT	MQHTYKEGDEFV	FMDMESYE_EGRLS	_____	AAQIGDRVVKYLKEGMEVNVIRWGEQ		
5	TGAIQEKTFR	AGEKVAKAQIETKT	MQYLYANGDQHV	FMDTSSYE_QLELS	_____	ATQIEEELKYLLNMSVHIMMYQDE		
6	TGNVTELTFK	ASDRIPLADFEQVY	ATYSYNDGENYY	FMNTQTYD_MIAVP	_____	KEKIEEAKFLKEGMEVIVFLYKQG		
7	DGKVIEKTFH	AGDKCEPNLVEKTM	QYLYHDGDTYQ	FMDIESYE_QIALN	_____	DSQVGEASKWMLDGMQVQVLLHNDK		
8	NKFVIRETLK	GADTAAEIEIYEV	SAQYLYKDKDVLV	FMDLETYD_QVSL	DLKESANLQDKVPFLQ	ESEIYSLVTFD	NV	
SIA	SsSIisSAIS_I	.sSISSiSISSSii	.sISSsSSii_IISsSSIS	.s.sIS.aaSSSISSsSisIIS	SsiSisIISiSSS			
parse	_11		111	-	-	23332	1	11
DLG.Eub	hhhhhhhhh_h	eeee	eeee	eeee	eeee	eeee	eeee	eeee
<i>Eurkaryotes</i>								
¥ 9	TGKKYEDIC_PSTH	NMDVPNIKRND	FQLIGI_QDGFLS	LLQ_DSGEVREDLRL	PEGE_LGRxxx			
¥10	TGKKYEDIC_PSTH	NMDVPNIKRND	FQLIGI_QDGYLS	LLQ_DSGEVREDLRL	PEGD_LGKEIEQKYDSGE	EILITVLS		
¥11	TGKKYEDIC_PSTH	NMDVPNIKRND	FQLIGI_QDGYLS	LLQ_DSGEVREDLRL	PEGD_LGKEIEQKYDCGE	EILITVLS		
12	TGKKYEDIC_PSTH	NMDVPNIRND	FQLIGI_QDGYLS	LLQ_DSGEVPEDLRL	PEGD_LGKEIEQKYDCGE	EILITVLS		
13	NGKKYEDIC_PSTH	NMDVPNIKRND	FQLIGI_QDGYLS	LLT_ESGEVREDLKL	PEGD_LGKEIEGKFANE	DVQISVIS		
¥14	TGKKYEEIC_PSTH	NIDVPNVSRKEY	TVMDV_QDGYLS	LLD_AGGEVKEDLAL	PEDD_IGKEITQMLKEGK	EPLVSVIS		
15	TSKKLEDIC_PSTH	NMDVPPVKRREY	LLMAI_DDGYCS	LMDPESCEQKDDL	KLPTDE_LGQQIRDAYEK	DEGSVLVQVVS		
16	TTKKLEDIC_PSTH	NMDVPPVKRREY	ILMSI_EDGFC	SLMDPESCEKDDL	KMPEDG_LGNTIREALEK	DEGSVLVQVVA		
¥17	TGKKLEDLC_PSTH	NMDVPPVKRRTDY	QFSYI_DEDFL	VLID_SNGEEKREL	KMPGE_LAKRIEKL	EEGK_DFFVGVQT		
¥18	TGKKLEDLS_PSTH	NMEVPPVKRNEY	QLLDI_DDGYLS	LMN_MGDTKDDVKA	PEGE_LGDSLQTA	FDEGK_DLMVTIIS		
¥19	TGKKLEDLS_PSTH	NLEVPFVKRSEY	QLLDI_DDGYLS	LMT_MDGETKDDVKA	PEGE_LGDSMQAAF	FDEGK_DLMVTIIS		
20	NGRKYEDMS_PSTH	NMDVPPVKRDEY	QLVNI_DDGYL	NLMT_TDGTTKDD	VRLPEGE_LGNEIEEG	FDEGR_DLIITVVS		
21	NGKKLEDI_VPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_SDGNTKDDL	RLPTDET	TLVAQIKGEGFESGK	DLVVTVQS	
22	NGKKLEDI_VPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	RLPTDD	DALLNQVKGFEQK	DLVLSVMS	
23	TGKKLEDI_VPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	RLPTDD	TLLAQVKDGAEGK	DLVLSVMS	
¥24	TGKKLEDI_VPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	RLPTDD	NLLALIKDGAEGK	DLVLSVMS	
¥25	TGKKLEDI_VPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	RLPTDD	NLLTQIKDGAEGK	DLVLSVMS	
¥26	TSKKLEEVYPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	KLPTDD	SLLTQIKDGAEGK	DLVLSVMS	
27	TSKKLEEVYPSSH	NCDVPHVNRTEY	QLIDISEDGFV	SLLT_ENGNTKDDL	KLPTDD	SLLTQIKDGAEGK	DLVLSVMS	
SIA	sssAIAsIii	.AiAAISI	.sISASSiIi	SIaSSsIIi	SIiSsSsSI	SI.ssSsIi	SISSs.sSsS	.Siiisiis
parse	1	1221	1	1 1111	11111	111111		1121
DLG.Euk	hhhhhhhhh	(eeee)	eeee	eeee	(hhhhhhhhh)	eeee	hhhhhhhhhhhhh	eeee
PHD.94:	_.....	EEEE		EEEE			HHHHHHHH	
ETH.94:	hh....._.....	EEE		EEEEEE		EEEE		EEEE
ee								
<i>Archaea</i>								
28	DSQK_RSIVQPVT	AKIYVPIVERKRA	QIISVTGN_VAQL	MDLETYET_FELEV	_____	PEELKDKMEQGRE_VIYLE		
29	DGKV_RSIVKPTS	AEVDVPIIDKKT	AQVIAITPD_TVQ	IMMETYET_FEVP	IDTGVAD	IRDQLKEGIN_VEYWE		
30	DNQK_RSIVKPV	DSKVDPIDDKRT	AQVIAIMGG_DVQ	LMLETYET_FETPI	_____	PDELSEQLVEGVE_VEYIE		
31	EKVK_KEFVAPT	SSKVEVPIIDRR	KGQVLAIMGD_MVQ	IMDLQTYET_LELPI	PEGI_____	EGLEPGGE_VEYIE		
¥32	TGQK_RSLMAP	VDQVQVEVPIE	KHVGQILADKVG	NLTIMDLESYET_FDLE	_____	KPTENEIVSKIRPGAE_IYWS		
33	DGGK_RTLSLP	VDQAQVEVPIE	KFTAQILSVSGD_VI	QLMDMRDYKT_IEVPM	_____	KYVEEAKGRLAPGAE_VEVWQ		
Target#	57 60	70	80	90	100	110	120	127
SIA	sSss_ssIis.i	SsSISI.IISS	SSiAii.s.S	SIiIIAIS	Isa_____	IsiSisS.ssSai	SSSISS.Ss_____	ISIIS
parse	1	_	1 1	1	22_1	1_1	11222111	1 1112_1
DLG.Arc		eeeeeeee(ee)	eeeeee	eeee	e_eee	hhhhhhhhh	e_eeee	
SAB.98	HHHH_HHHHH	EEEEEEEEeee		EEEEEE		EEEEEE		
PHD.1:	_EEEE		EEEEEE		EEEEEE		EE	
PHD.2:	_.....	EEEE		EEEE			HHHHHHHH	

FIG. 1—Continued

Ali#	140	150	160	170	180	190	200
	.	.	.	.	.	.	.
Eubacteria							
1	PLYIELPVTVELEVTHTEPGLQGRSSAGTKPATLQTGAQINVPLFINTGDKLKVDSRDGSYLGRVNA						
2	ILDVELPTSVVLEIVETDPGVKGDATGGTKPAILETGAQVMVPLFIAQGERIKVDTRDGSYLGRDNA						
3	VIEVELPNSVVLEVIETDPGVKGDATGGTKPAKVETGAQVMVPLFISVGERIKIDTRNDSYLGRE						
4	VLEVELANSVVLEVIQTDPGVKGDATGGTKPAIVETGATVMVPLFISQGERIKIDTRDDKYLGRE						
5	TLGIELPNTVELKVVETEPGIGKDTASGGTKPAKTETGLVVNVPPFFVNEGDTLVVNTSDGSYVSRA						
6	PIGIELPKHVELQVVETEPAFKGDTQAGGTKPAKLETGAVIQVPPFFVKEGDIVKVDTRTGSYVERVKEAK						
7	AISVDVPQVVALKIVETAPNFKGDTSSASKKPATLETGAVVQVPPFHVLEGEI IKVNTETEEYLEKVK						
8	VIDIKLAPKIAFEVVEVEAAVKGDTVNTAMKNITLNTGLVVKAPLFINVGDVKLINSETKEYAERIKN						
SIA	sIsISI.SsISISiSiS.Sis.Assss.SAsISisA.I.IsI.IiISS.SsIsISsSSSSIIIsSsss.a						
	11	11	1	11	1	1	11 1221
DLG.Eub	eeeeee	eeeeeeeeee		eeeeee	eeee	eeeeee(eeee)	eeeeee
Eukaryotes							
¥ 9							
¥10	AMTEEA AVAIKAMAK						
¥11	AMTEEA AVAIKAMAK						
12	AMTEEA AVAIKAMAK						
13	AMNEECAVAIKPCK						
¥14	ALGKEGVVSVKVSNN						
15	AIGEEAILGWKVSTKE						
16	ACGEEAILGYKISTKE						
¥17	AMGEEAAIDVKEASNKD						
¥18	AMGEEAAISFKEAARTD						
¥19	AMGEEAAISFKEAPRSD						
20	AMGEETALACRDAPSS						
21	AMGEEQICALKDVGP						
22	AMGEEQICAVKDIGTKS						
23	AMGEEQICGIKDIGPK						
¥24	AMGEEQICGIKDVGP						
¥25	AMGEEQICALKDIGPK						
¥26	AMGEEQICALKDIGGKN						
27	AMGEEQICALKDIGGKN						
SIA	iissasiisiSSisSss.						
	1		12234				
DLG.Euk	eeeeee						
	hh)						
PHD.94:	...HHHHHHH....						
ETH.94:	EE....EEEE....						
Archaea							
28	SLGKRKIERMA						
29	TLGRIKIMRIKGEGE						
30	ALGQRKLMRTKG						
31	AVGQYKITRVIGGK						
¥32	VMGRRKIVRVK						
33	ILDYKIIRVK						
Target#	128	138					
SIA	.IsSSAIsAiS.ssa						
	1	12344					
DLG.Arc	(eeeeee)						
PHD.1:	H..EEEEEEE.						
PHD.2:	EEEEEEE....						

FIG. 1—Continued

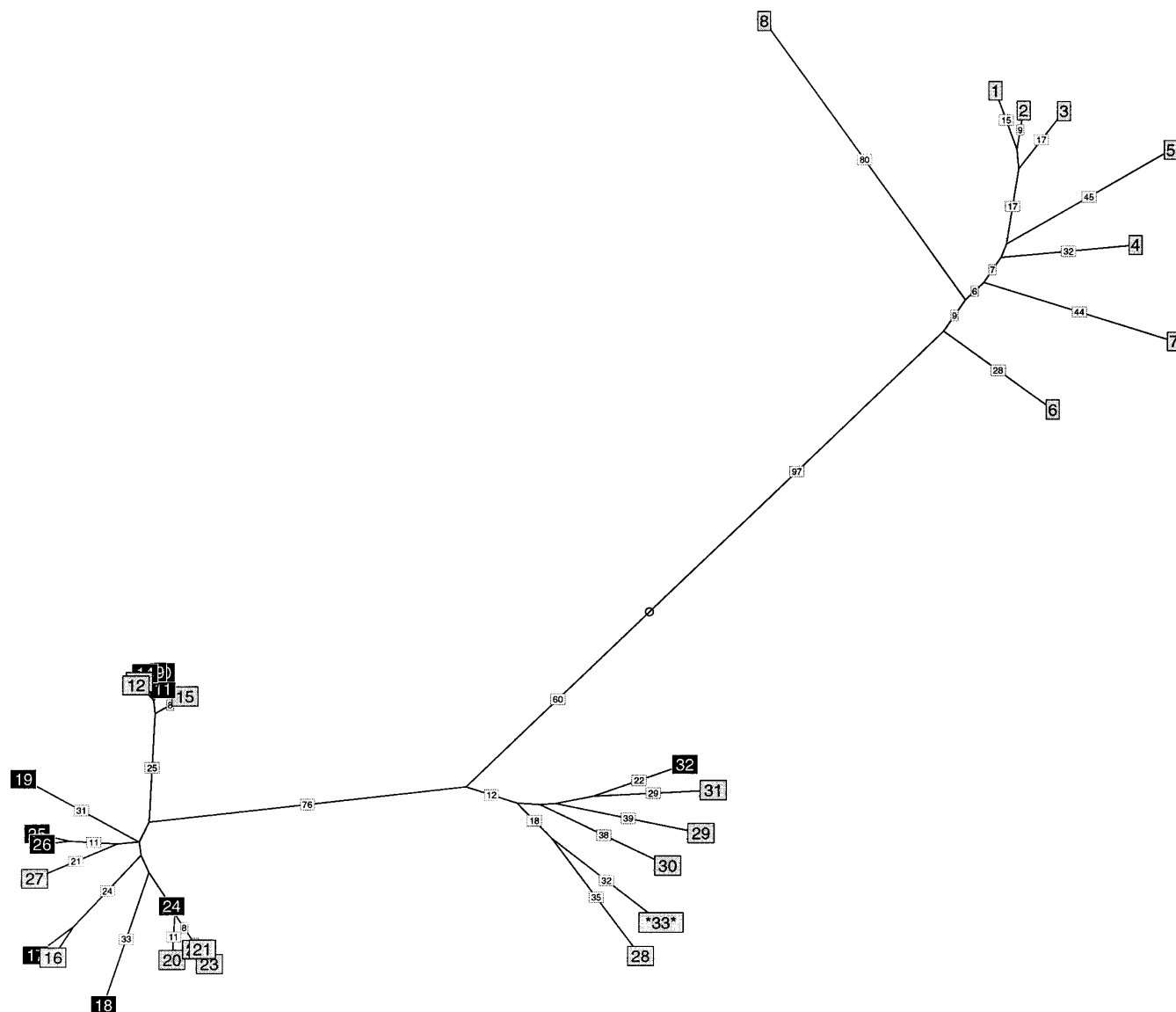


FIG 2. Maximum likelihood tree for the initiation factor 5A superfamily. Derived from the sequences aligned in Figure 1. Numbers in gray boxes denote sequences as listed in Figure legend 1, small boxed numbers are evolutionary distances in PAM units, see Methods for details. Sequences available for the 1994 prediction are marked black, the target sequence for the CASP3 prediction experiment, IF-5A from *Pyrobaculum aerophilus*, is marked with asterisks.

(Figure 1), using an archaeobacterial sequence as an outgroup (18). Several ambiguities characteristic of predictions made from multiple sequence alignments were present in this prediction, including an uncertain internal helix (frequently confused with an internal strand) and several possible edge strands (frequently

confused with coils). These ambiguities were manifest when transparent prediction tools (1) were used, and tertiary structure modelling based on the predicted secondary structure therefore was recognized as being unreliable. Figure 3 allows the comparison of secondary structure predictions made in 1994 (ETH.94,

FIG 3. Summary of secondary structure predictions and assignment of core tertiary structure elements. Secondary structure predictions are labeled and aligned as in Figure 1 and shown with CASP3-target sequence 33. DLG.98 is a consensus prediction for all three subfamilies over the N-terminal approx. 100 positions, and over the eukaryotic and archaeal subfamilies in the C-terminal part, where structures may have diverged in the eubacteria. Elements predicted to form the N-terminal tertiary structural core are numbered corresponding to the OB-fold description by Murzin (25), although the tertiary structure may adopt a more open, twisted Greek-key beta-barrel/sandwich topology than found in the "classic" OB-fold topology, see Discussion.

```

=====
Ali#           001       010       020       030       040       050
                |       |       |       |       |       |
DLG.Eub                (eeee)  eeeee  eeeee  eeeee  eeeee
DLG.Arc                eeeeeee eeeee  eeeee  AAAAAAeeeeee
DLG.Euk                eeeeeee eeeee  eeeee  eeeeeAAAAAaaaa eeeeeee

PHD.94 (Euk)      ....._.....HHHHHHHH...EEEE...EEEEEE.....EEEEEEEEEE.
ETH.94 (Euk)      ....._..EEEEEE...EEEE...EEEE...EEEEEE.....hhhhhhhhhh
                                     eeeeeeeeeee

PHD.1 (Arc)      ...EEEE.EEEE...EE...EEEEEE...EEEEEEEE.....EEEEEEEEEE
PHD.2 (Arc)      .....HHHH...EEEE...EEEEEEEE.....EEEEEEEEEE
SAB.98 (Arc)      .....EEEEEEEE...EEEEEE...EEEEEEEE.....EEEEEE..HH
DLG.consensus.98 .....EEEEEE.EEEEE...EEEEEE...EEEEEE...
3D-CORE ELEMENTS          Sla      S1b          S2 ActSite-loop S3
33 (Target)      MVLKWVMSTKYVEAGELKEGSYVVIDGEPCRVVEIEKSKTGKHGSAKARIVAVGVF
Target#          1          10          20          30          40          50          56

```

```

=====
Ali#           060       070       080       090       100       110       120       130
                |       |       |       |       |       |       |       |
DLG.Eub hhhhhhhh_h      eeeee  eeeee  eeeee  eeeee  eeeee  eeeee
DLG.Arc                eeeeeee (ee) eeeee  eeee  e_eee  hhhhhhhh  e_eee
DLG.Euk hhhhhhhh      (eeee) eeeee  eeeee  eeee  eeee  hhhhhhhhhhh  eeee
                                     (hhhhhhhh)

PHD.94 ....._.....EEEE_...EEEE_....._..HHHHHHHH..._..EEEEEE.
ETH.94 hh....._.....EEE.....EEEEEE_...EEEE_...EEEE...H_HHHHHHHHHHH..._..EEEEEE
ee

PHD.1 ....._EEEE.....EEEEEE.....EEEEEE..._EEEEEE....._EE..._..HHHHH.....E_EEEHH
PHD.2 ....._.....EEEEEE..._EEEEEE....._....._.....HHHHHHHH..._...EE
SAB.98 HHHH_HHHHH.EEEEEEEeee...EEEEEEEE..._EEEEEEEE...
DLG.98 .hhhhhhhhh...eeeeeeee.EEEEEEE...EEEE...EEEE...HHHHHHHHH...EEEE..
3D-CORE   H1      S4      S5      (S6?)
33        DGGK_RTLSPVDAQVEVPPIIEKFTAQILSVSGD_VIQLMDMRDYKT_IKVPM_KYVEEEAKGRLAPGAE_VEVWQ
Target# 57 60          70          80          90          100          110          120          127

```

```

=====
Ali#           140       150       160       170       180       190       200
                |       |       |       |       |       |       |
DLG.Eub eeeee  eeeeeeeee  eeeee  eeee  eeeee (eeee)  eeeee
DLG.Arc                (eeee)
DLG.Euk                eeeee
hh)

```

PHD.94: ...HHHHHHH.....
ETH.94: EE....EEEE.....

PHD.1: H..EEEEEEEE.
PHD.2: EEEEEEE....
SAB.98
DLG.98eeeeee

33 ILDRYKIIRVK*
Target# 128 138

PHD.94) with 1998 predictions based on better populated sequence alignments and more balanced trees. While the exact sequence submitted to the PHD-servers differed (see Figure 1), we find that the 1998 predictions by the different methods seem to converge better than those generated in 1994, with the exception of PHD.2, returned by the UCLA-DOE server, which seems to be based on a preformatted multiple sequence alignment. As the differences in the PHD outputs could reflect differences in the alignment used, and the UCLA-DOE prediction was discarded.

Two points become apparent when the secondary structure predictions for the three kingdoms are compared. First, the correspondence between the prediction for the eubacterial proteins and the prediction for the other two kingdoms becomes doubtful in the C-terminal part of the alignment (Figure 3, after alignment position # 110, approximately). This suggested an end of the core domain and/or significant structural divergence between the eubacterial EF-P structure and the IF5A structure in eukaryotes and archaea. Second, we can identify seven, or possibly eight, elements of secondary structure which are likely to form the core of the N-terminal part of the protein. These elements are marked in Figure 3 and include (a) three consecutive beta strands, with a possible bend in the first of the three, (b) an alpha helix, which may be replaced by a surface loop in the archaeal structures, and (c) two (or possibly three) beta strands. Together, the predicted core elements can be described as S1a-S1b-S2-S3-H1-S4-S5-(S6?) (Figure 3). While secondary structure content based on circular dichroism spectroscopy is rarely reliable quantitatively, it is interesting to note that the reported, high, beta-sheet content (18) agrees well with the prediction, and supports it qualitatively.

Next, an analysis of highly conserved, functional and/or aromatic amino acid residues indicative of functionally important sites (where sequence divergence has been constrained during divergent evolution) was used to orient the predicted secondary structural elements (1, 23). Position 042 in the alignment holds the hypusylated Lys; residues in the surrounding positions are well conserved in the eukaryotic and archaeal proteins. This indicates the presence of either a single, bipartite, or two separate functional sites in the three-dimensional structure. Finally, while some of the strands appear to be bent, or bulged, none of them could be identified with certainty to be an edge strand, as it had been the case for the combinatorial tertiary structure prediction of synaptotagmin (3).

These observations indicate that the folded structure of the first domain is composed of highly twisted, anti-parallel beta-sheet and a single alpha-helix, and that the residues involved in functional interactions are located at both "ends" of the resulting barrel/sandwich-structure. With respect to the strand order in the

sheet, our preference is for a "Greek-Key" topology, or similar, with the predicted helix in the long connection. For core strand segments and a closed barrel structure, this topological arrangement is exemplified in the superfold described by Murzin as the "OB-fold" (25). While we cannot exclude with certainty the possibility that the core is made of more than five strands, or that the chain arrangement follows a different topology (due to possible mispredictions in both the exact location and the core assignment of the strands), one of our preferred tertiary structural models would bear resemblance to the OB-fold topology, but with some noticeable deviations from the structural properties typically conserved in the "classic" members of this fold family (26), which has been found in many variations (see the SCOP database, (27)).

One of the deviations would be the non-classic location of active site residues in the connecting loop between strands 2 and 3 in our model (25). While unusual, this location is not without precedent in the structures listed with OB-fold topology in SCOP. Namely, the NMR structure of the ribosomal protein S17 from *Bacillus stearothermophilus* (PDB:1rip) has an insertion of functional residues at what would be the equivalent location (28). The observation that the RPS17 structure would be a somewhat unusual member of the OB-fold family in that the barrel is opened, and maybe better described as a strongly twisted beta-sheet, is further compatible with a second deviation from the OB-fold characteristics, the exclusion of aromatic side-chains from the packed core (26). Finally, the reported minimum human substrate for deoxyhypusine synthase (10) corresponds to positions 022-073 in our alignment (Figure 1). This would be too short to include the entire open barrel domain proposed in Figure 3. Reasons for this inconsistency could be a misassignment and/or misalignment of any core secondary structure, supersecondary structure formation in the fragment, or misprediction of the tertiary structure altogether.

The proposed tertiary structural features were compared them with the output of the automated fold recognition methods by Fischer (29) and Rice (30), accessible through the UCLA-DOE server (31), and the ProCyon method by Sippl (King's Beech, H. Floeckner & M. Sippl, URL:<http://www.horus.com/sippl/download.html>) (data not shown). The top 15 ranked folds by all methods were a mix of beta sheet-rich folds, dominated by several examples of the immunoglobulin-like beta sandwich and OB-superfolds. However, all returned matches had sub-significant scores. Nevertheless, the structure of ribosomal protein S17 (which bears a functional site between strands 2 and 3), was ranked 11 by one of the UCLA-methods (gonnet+predss, (29)). The structure ranked among the top five by most of the UCLA-methods, Rous Sarcoma Viral Protease (PDB: 2rsp),

forms a 6-stranded barrel of unusual topology, but would lead to a remotely similar structure to the open barrel discussed above, if the assignment of core strands and active site locations, with an alternative position for strand S3, were used.

This communication shows the impact of genomic sequences on the quality of secondary and tertiary structural models derived from analyses of multiple sequence alignments. As is evident from a comparison of the tree in Figure 2 with the universal tree of life, the set of sequences available for this target is now as widely dispersed as possible. It will, of course, be necessary to wait for the experimental structure to emerge to learn whether prediction methods were successful in this case even with a set of sequences as effectively distributed as possible on the universal tree. When the experimental secondary structure of the protein becomes known, and our prediction can be evaluated in its individual elements, this comparison will be useful for estimating the reliability that can be expected from contemporary prediction methods using multiple sequence analysis, and for guiding the design of additional genome projects to support a balanced distribution of organisms for best results in protein structure prediction.

ACKNOWLEDGMENTS

We are indebted to Dr. Heidi C. João and Dr. Manfred Auer (Novartis Research Institute, Vienna) for sharing their expert knowledge on IF5A. DLG is supported by a Postdoctoral Fellowship from the Leukemia Society of America.

REFERENCES

- Benner, S. A., Cannarozzi, G., Gerloff, D., Chelvanayagam, G., & Turcotte, M. (1997) *Chem. Rev.* **97**, 2725–2843.
- Gerloff, D. L., & Benner, S. A. (1995) *Proteins: Struct. Funct. Genet.* **21**, 273–281.
- Gerloff, D. L., Chelvanayagam, G., & Benner, S. A. (1995) *Proteins: Struct. Funct. Genet.* **21**, 299–310.
- Gerloff, D. L., Cohen, F. E., Korostensky, C., Turcotte, M., Gonnet, G. H., & Benner, S. A. (1997) *Proteins: Struct. Funct. Genet.* **27**, 450–458.
- Chen, K. Y., & Liu, A. Y.-C. (1997) *Biol. Signals* **6**, 105–9.
- Park, M. H., Lee, Y. B., & Joe, Y. A. (1997) *Biol. Signals* **6**, 115–123.
- Liao, D.-I., Wolff, E. C., Park, M. H., & Davies, D. R. (1998) *Structure* **6**, 23–32.
- Bartig, D., Lemkemeier, K., Frank, J., Lottspeich, F., & Klink, F. (1992) *Eur. J. Biochem.* **204**, 751–758.
- Chung, S. I., Park, M. H., Folk, J. E., & Lewis, M. S. (1991) *Biochim. Biophys. Acta* **1076**, 448–451.
- Joe, Y. A., & Park, M. H. (1994) *J. Biol. Chem.* **269**, 25916–25921.
- Park, M. H., Wolff, E. C., & Folk, J. E. (1993) *BioFactors* **4**, 95–104.
- Kang, H. A., & Hershey, J. W. B. (1994) *J. Biol. Chem.* **269**, 3934–3940.
- Ruhl, M., Himmelsbach, M., Bahr, G. M., Hammerschmid, F., Jaksche, H., Wolff, B., Aschauer, H., Farrington, G. K., Probst, H., Bevec, D., & Hauber, J. (1993) *J. Cell Biol.* **123**, 1309–1320.
- Bevec, D., & Hauber, J. (1997) *Biol. Signals* **6**, 124–133.
- Bevec, D., Klier, H., Holter, W., Tschachler, E., Valent, P., Lottspeich, F., Baumrucker, T., & Hauber, J. (1994) *Proc. Nat. Acad. Sci.* **91**, 10829–10833.
- Liu, Y. P., Nemeroff, M., Yan, Y. P., & Chen, K. Y. (1997) *Biol. Signals* **6**, 166–174.
- Rost, B., & Sander, C. (1993) *J. Mol. Biol.* **232**, 584–599.
- Klier, H., Csonga, R., Joao, H. C., Eckerskorn, C., Auer, M., Lottspeich, F., & Eder, J. (1995) *Biochemistry* **34**, 14693–14702.
- CASP3-Organizers (1998) WWW-site: 3rd community wide experiment on the critical assessment of techniques for protein structure prediction (CASP3), (URL: <http://PredictionCenter.llnl.gov>).
- Devereux, J., Haeberli, P., & Smithies, O. (1984) *Nucl. Acids Res.* **12**, 387–395.
- Hallett, M., Korostensky, C., Knecht, L., & Gonnet, G. H. (1992) WWW: Computational Biochemistry Research Group (CBRG) Server (ETH Zürich, Switzerland, URL: <http://cbrg.inf.ethz.ch>).
- Rost, B., & Sander, C. (1993) WWW: The PredictProtein Server (EMBL Heidelberg, Germany, URL: <http://www.embl-heidelberg.de/predictprotein/>).
- Benner, S. A., & Gerloff, D. (1991) *Adv. Enz. Regul.* **31**, 121–181.
- Benner, S. A., Gerloff, D. L., & Chelvanayagam, C. (1995) *Proteins: Struct. Funct. Genet.* **23**, 446–453.
- Murzin, A. G. (1993) *EMBO J.* **12**, 861–867.
- Bycroft, M., Hubbard, T. J. P., Proctor, M., Freund, S. M. V., & Murzin, A. G. (1997) *Cell* **88**, 235–242.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
- Jaishree, T. N., Ramakrishnan, V., & White, S. W. (1996) *Biochemistry* **35**, 2845–2853.
- Fischer, D., & Eisenberg, D. (1996) *Protein Sci.* **5**, 947–955.
- Rice, D., & Eisenberg, D. (1996) *J. Mol. Biol.* **267**, 1026–1038.
- Fischer, D., Rice, D., & Eisenberg, D. (1995) WWW: UCLA-DOE Structure Prediction Server (UCLA-DOE Laboratory, Los Angeles, U. S. A., URL: <http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html>).
- Rost, B. (1996) *Methods Enzymol.* **266**, 525–539.