A Consensus Sequence in the N-terminus of Exported Proteins: Resemblance with Metal Binding Domains and Implications in Protein Translocation Across Membranes

Panagiotis A. Tsonis, Janusz M. Sowadski* and Paul F. Goetinck

La Jolla Cancer Research Foundation,
Cancer Research Center, 10901 N. Torrey Pines Road,
La Jolla, CA 92037

*Department of Biology,
University of California, San Diego, La Jolla, CA 92093

Received July 7, 1988

SUMMARY By analyzing and comparing the N-terminus of several exported proteins we identified two consensus sequences that resemble metal binding domains. The consensus sequences are part of the signal peptide and part of the adjacent sequences of the mature protein. Three-dimensional modelling of one such domain suggests a conformation with implication in signal peptide insertion.

In order for proteins to be secreted they have to be translocated across membranes. For eukaryotes the translocation is across the endoplasmic reticulum and for bacteria it is across the cell membrane. The actual process of translocation across membranes is preceeded by a recognition step which, in eukaryotes, involves, in addition to the protein that is to be transported, a soluble receptor known as the signal recognition particle and a membrane receptor known as the docking protein (1,2). The signal recognition particle is made up of a 7S RNA and six different polypeptides (3). A third component, the signal sequence receptor, may also play a role in this process (4). The

0006-291X/88 $1.50

nascent protein that is to be secreted contains information that is instrumental for either its targeting to or its translocoation across the ER membrane. There is a great deal of evidence that indicates that this information is contained in the signal peptide, the N-terminus of the primary translation product of secretory proteins that is proteolytically removed to form the mature protein (1,2). Signal peptides range in size from 15 to 30 amino acids and although they share as a common characteristic a centrally located stretch of hydrophobic amino acids, their primary sequences show no homologies that might suggest how they are recognized in the targeting or the translocoation process. Whereas the signal peptide is essential for export of proteins, it is not known if any additional information for this process resides in the mature protein.

In an analysis of the sequence of the primary translation product of cartilage link protein (5) we identified a sequence that resembles the consensus sequence for the metal binding domains of the transcription factor IIIA of *Xenopus laevis* (6). This sequence of link protein encompasses the signal peptide and the N-terminus of the mature protein. A search of the protein sequence database of the National Biomedical Research Foundation (Georgetown University Medical Center, Washington D.C., 4028 sequences, 963031 residues) indicated that such a sequence involving the signal peptide is not restricted to link protein. Indeed we found such a sequence to be present in the N-termini of the primary translation product of 53 proteins.

Two very similar consensus sequences could be identified. These are: $\alpha\text{-}X_{0\text{-}4}\text{-}\alpha\text{-}X_{5\text{-}17}\text{-}\alpha\text{-}X_{1\text{-}5}\text{-}\alpha$ and $\alpha\text{-}X_{5\text{-}21}\text{-}\alpha\text{-}X_{1\text{-}5}\text{-}\alpha$, where α is a cysteine or a histidine and X is any amino acid. These sequences resemble the consensus sequence $C\text{-}X_4\text{-}C\text{-}X_{12}\text{-}H\text{-}X_3\text{-}H$ which makes up the metal binding domains of the *Xenopus* transcription factor IIIA (TFIIIA) (6-10). The alignment of these three consensus sequences is shown in Figure 1. TFIIIA which binds to 5SRNA genes and their transcripts contains 9 homologous repeats of this domain and the binding of Zn^{2+} to these domains has been demonstrated. In those instances where the Zn^{2+} binding activity of such domains has been demonstrated it has

Consensus	(TFIIIA)	$C - X_4 - C - X_{12} - H - X_3 - H$
Consensus I		$\alpha - X_{0-4} - \alpha - X_{5-17} - \alpha - X_{1-5} - \alpha$
Consensus II		$\alpha - X_{5-21} - \alpha - X_{1-5} - \alpha$

Fig. 1. Alignment of the two consensus sequences, I and II, found in the N-terminus, showing a resemblance with the metal binding domain of the TFIIIA. αl5, a Cys or His and X any amino acid.

been shown to be determined by the arrangement of the Cys and His ligands rather than by the primary sequence of the protein (6). In Figure 2 we present the sequences of the proteins according to the pattern of alignment presented in Figure 1. For each of the two patterns we list first the proteins whose signal peptide cleavage site is known, followed by those whose cleavage site is not known. Since we have no evidence that the sequences we describe bind metals we will, based on their structural motif, refer to them as putative metal binding domains. Of the 53 N-terminal sequences listed, 27 and 26 follow the type I and the type II consensus sequences respectively. Forty-nine of the sequences are from eukaryotes and 4 from prokaryotes. Of the eukaryotic sequences, 40 are from vertebrates, 4 from invertebrates and 5 from plants. The signal peptide cleavage site is known for ten of the sequences that follow the type I consensus sequence. Seven of these cleavage sites are situated on the N-terminal side of the first α position and 3 are situated between the second and the third α positions. Of the 15 known signal peptide cleavage sites in the sequences that follow the type II consensus sequence, three are situated on the N-terminal side of the first α position, 11 are located between the first and the second α positions and one is located on the C-terminal side of the third α position. Finally, the two most C-terminally located α positions for the type I consensus sequences are most often Cys and Cys (16 out of 27) whereas those for the type II consensus sequences are most often His and His (23 out of 26). The significance of the above observations is not known at the present time.

					α	X_{0-4}	α	X_{5-17}	α	X_{1-5}	α	
consensus I	link protein-chicken	9		LISV	C	W▼AEP	H	PDNSSLE	H	ERII	H	IQE
	vasotocin-amphibian	-11			C	FL	C	LLALSSA▼	C	YIQN	C	
	mesotonin-amphibian	1			▼C	YIQN	C	PIGGKRSVIDFMDVRK	C	IP	C	
	vasopresin-amphibian	1			▼C	YIQN	C	PRGGKRATSDMLRQ	C	LP	C	
	oxytosin-amphibian	1			▼C	YIQN	C	PLGGKRAALDLDMRK	C	LP	C	
	fibronectin receptor β-subunit-human	16			C		C	VF▼AQTDENPCLKANAKS	C	GE	C	
	prolactin-human	29		▼LPI	C	GGAAR	C	QVTLRDLFDRAVVLS	H	YI	H	
	purothionin A-I wheat	1		▼KS	C		C	KSTLGRN	C	YNL	C	
	purothionin II-precursor-barley	19		▼KS	C		C	RSTLGRN	C	YNL	C	
	vitronectin precursor-human	20		▼DQGS	C	KGR	C	TEGFNVDKKCQ	C	DEL	C	
	superoxide dismutase-photobacterium	44	F	HI	H	QNGS	C	ASSEKDGKVVLGGAAGG	H	YDPE	H	
	bromelain inhibitor-pineapple	14	Y	K	C	Y	C	ADTYSD	C	PGF	C	
	follitropin β-human	17			C	RF	C	LTINTTW	C	AGY	C	
	trypsin inhibitor-mongolian snake	15			C	W	C	ISRGYLCG	C	MP	C	
	keratin feather I-chicken	3	F	DL	C	RP	C	GPTPLANS	C	EP	C	
	somatomedin B-human	6		DKK	C	Q	C	DLCSYYQSN	C	T	C	
	large γ-3 heavy chain disease protein											
	-human	24			C	PR	C	PEPKSCDTPPP	C	PR	C	
	agglutinin isolectin-2 wheat	17			C		C	SQYGYCGMGGDY	C	GKG	C	
	vasopresin-meurophysin precursor											
	-bovine	17	F	TSA	C	YFQN	C	PRGGKRAMSDLELRQ	C	LP	C	
	creatin kinase M chain chicken	3	F	SST	H	NK	H	KLKFSAEEEFFPDLSK	H	NN	H	
	phopholipase A2 eastern cotton moth	24	Y	GFCN	C	GWG	H	RGQPKDATDRC	C	FV	H	
	urokinase-type plasminogen activator											
	precursor-pig	7			C	LSL	C	VLVVSDSKGS	H	EL	H	
	glutelin 5-maize endosperm	2			H	TSGG	C	GSQPPPPVHLPPPV	H	LPPPV	H	
	transaminase β-E. coli	37	F	EGIR	C	YDS	H	KGPVVFRHRE	H	QRL	H	
	acrosin inhibitor A_1-(pig bovine)	32	Y	ANP	C	IF	C	SEKGLRNQKFDFG	H	WG	H	
	acrosin-inhibitor I-bovine	36	Y	SNE	C	TF	C	NEKMNNDADI	H	FN	H	
	cysteine rich intestinal protein-rat	28			C	LK	C	EKCGKTLTSGG	H	AE	H	
					α	X_{5-21}			α	X_{1-5}	α	
consensus II	L-CAM-chicken	-6	F	PK	H	DPGFLRRQKR▼DWVIPPIS			C	LEN	H	
	chorioamammotropin precursor-human	13	F	ALL	C	LPWLQEAGA▼VQTVPLSRLFD			H	AMLQA	H	
	α-1 antitrypsin precursor-human	15		LC	C	LVPVSLAE▼DPQGDAAQKTDTS			H	HDQD	H	
	parathyroid hormmone precursor											
	-bovine	18			C	FLARSDGKSVKKR▼AVSEIQFM			H	NLGK	H	
	growth hormone-rat	-14	F	SLL	C	LLWPQEAGA▼LPLMPLSSLFANAVLRAQ			H	L	H	
	colony stimmulating factor-mouse	-13		LLLV	C	LLMSRSIAK▼GVSG			H	CS	H	
	asl casein-bovine	-8			C	LVAVALA▼RPK			H	PIK	H	
	aminodiphosphoribosyl-tranferase-											
	bacilus subtilis	8		LNEE	▼C	GVFGIWGHEEAPQITYYGL			H	SLQ	H	
	haptoglobin-1-human	5		▼NDVTDIADDG	C	PKPPEIA			H	GYVE	H	
	prothrombin precursor human	1		QLPG	C	LALAALCSLV			H	SQ	H▼36	
	nerve growth factor receptor-rat	1		▼KET	C	STGLYTHSGEC			C	KA	C	
	angiotensinogen precursor-human	23			C	LLAWAGLAA▼GDRVYI			H	PF	H	
	angiotensinogen precursor-rat	13	F		C	ILTWVSLTA▼GDRVYI			H	PF	H	
	urinary protein-human	-13		LLLV	C	LLASRIT▼EEVSEYCS			H	MIGSG	H	
	parathyroid hormone related											
	protein-human	-14			C	GRSVEGLSRRLKR▼AVSE			H	QLL	H	
	azurin - A. denitrificans	26			C	KQFTV			H	LK	H	
	factor XII-human	28			C	HFPFQY			H	RQLY	H	
	large heavy chain V-III human dob.	22			C	AASGFNF			H	EYNM	H	
	large chain V region mouse S107A				C	TASESLYSSK			H	KV	H	
	hemoglobin α-chain nile crocodile	34			C	AYPQTKIYFP			H	FDLS	H	
	cygnin-swan	5	Y		C	PKVGY			C	SSK	C	
	specific body pattern-fruit fly	3		LEDR	C	SPQSAPSPITLQMQ			H	LH	H	
	myohemerythrin -T zostericola	33	F		C	DIRDNSAPNLATLVKVTTN			H	FT	H	
	nerve growth factor β-chain											
	precursor-mouse	1		ML	C	LKPVKLGSLEVG			H	GQ	H	
	superoxide dismutase-horse	6			C	VLKGDGPV			H	GVI	H	
	chymotrypsin II-European hornet	25			C	GGSISKRYVLTAAHCLVGKSK			H	QVTV	H	

The finding of the presence of putative metal binding domains near or involving the signal peptide, suggests a mechanism that might provide a unique conformation for the N-termini of proteins that could be important in signal recognition and in the translocation process across membranes. An examination of the hydropathy profiles of the sequences in the N-termini of the proteins listed in Table 1 indicated a hydrophobic-hydrophilic-hydrophobic pattern. The hydropathy profiles and secondary structure prediction of link protein, urinary protein and angiotensinogen are presented as examples in Figure 3. Similar structural features have been described in zinc binding enzymes such as Cu-Zn superoxide dismutase and alkaline phosphatase whose three dimensional structure is known. In these enzymes a β-strand motif involving hydrophobic regions is the common structural element that provides the ligand for the zinc atom. Histidines are the invariant ligands although the number of histidines binding the zinc atoms may vary (11-13). Based on the stuctural features of these enzymes we aligned and compared the N-terminal signal sequence of link protein with the sequences within the region of the hydrophobic β-strand-hydrophilic loop-- hydrophobic β-strand motif that provides the histidine ligands to the zinc atoms in the three dimensional structure of alkaline phosphatases of E. coli (Fig. 4). It is apparent from Figure 3 that the two highly homologous β-strand regions of the alkaline phosphatases can be aligned with the two hydrophobic portions of the N-terminal region of link protein. Conserved ligands His-372 and His-412 of the E. coli alkaline phosphatase and corresponding ligands of the other alkaline phosphatases can be aligned with His-18 and His-31 of link protein. Using the coordinates of alkaline phosphatase and the above alignment, we built a three dimensional model of the N-terminal sequence of link protein. The resulting model, presented in Figure 5, has some important features. First, the antiparallel β-strands form a hydrophobic twisted surface.

Fig. 2. N-terminus of proteins containing putative metal-binding domains. The sequences of the proteins are arranged according to the two consensus sequences presented in Fig. 1. Numbers correspond to the first amino acid of the sequence. The number preceded with the minus sign represents cases where numbering of the reported sequences starts after the peptidase cleavage. The arrows indicate the cleavage site of the signal peptide. Standard single-letter amino acide abbreviations are used.

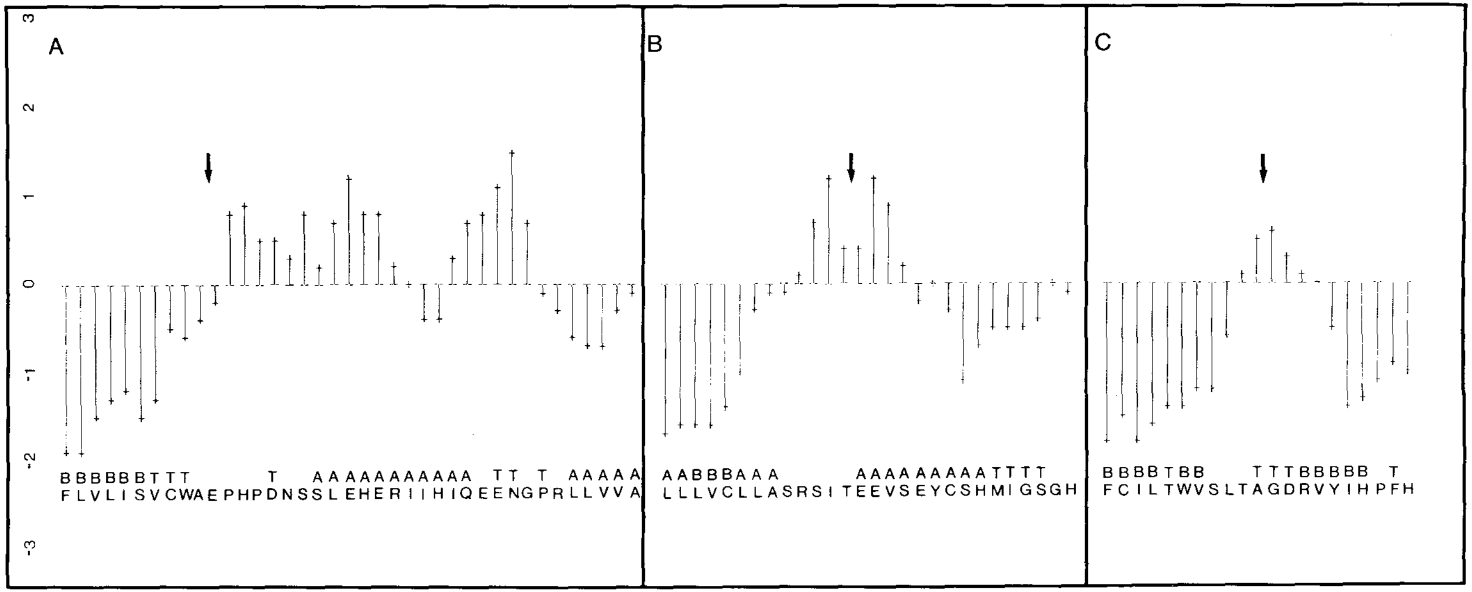

Fig. 3. Hydropathy profiles determined according to the algorithm of Hopp and Woods (14), of the putative metal binding domains found in the N-terminus of link protein (A), urinary protein (B), and rat angiotensinogen (C). The sequences and the secondary structure prediction using the algorithm of Garnier *et al.* (15) are shown. The arrows indicate the cleavage sites by signal peptidase.

		ß strand	loop	ß strand
E.	362	T L V I V T A D - - H A H		412 H T G S Q - - L R - I A A Y
P.	350	T L S L V T A D - - H S H		432 H A G E D - - V A - V F A R
L.	354	T L T V V T A D - - H S H		437 H G G E D - V A - V F S K
Link.	6	F L V L I S V C W A●E P H P D N S S L E H E R I I H I Q E E N G P R L L V V A		

Fig. 4. Sequence alignment of alkaline phosphatases and link protein. Alkaline phosphatase from E. coli whose structure has been solved at 2.8Å resolution contains two closely spaced zinc atoms. The zinc metal doublet is located between N- and C-termini of two antiparallel β-strands. These two antiparallel strands provide three out of four histidine ligands. Alignment of the sequence of the β-strand-loop-β-strand motif is presented for the alkaline phosphatases of E. coli (E) (16), kidney, liver and bone (L) (17) and placental (P) (18). Below is the putative metal binding domain of link protein. The alignment is good with the first β-strand, but less pronounced with the second. Position of cleavage site by signal peptidease is identified with a black dot.

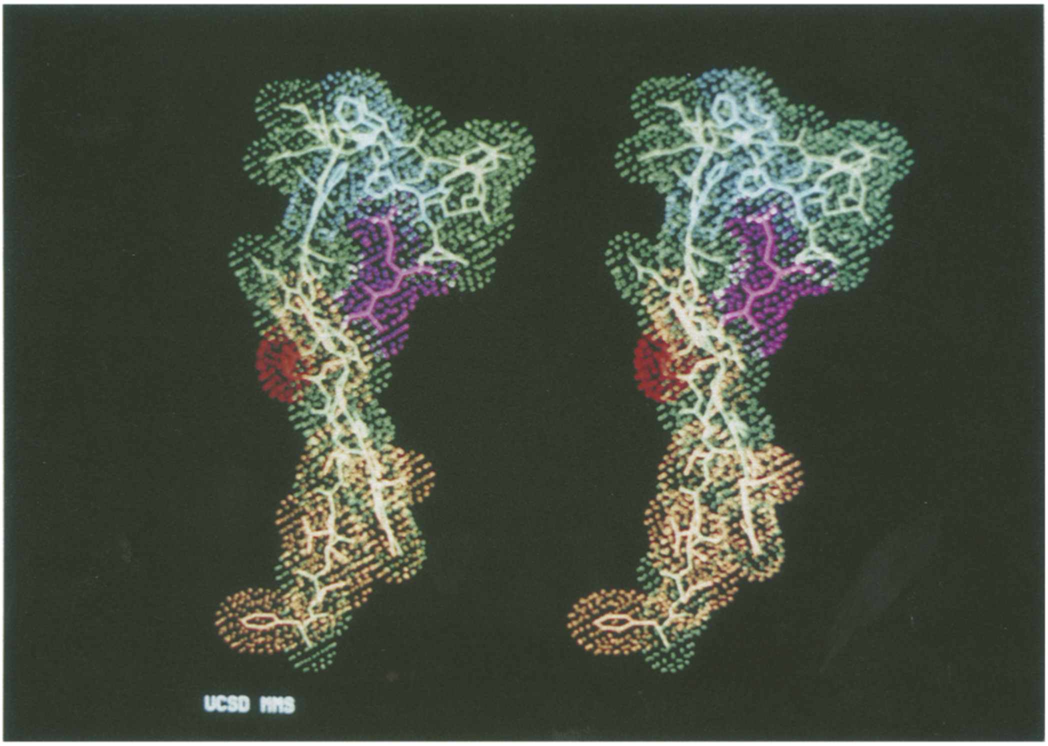

Fig. 5. The three dimensional model of the putative metal binding domain of link protein (color code: blue-histidines, yellow-hydrophobic residues, purple-signal peptidase cleavage site, red-cysteine). The model is based on the coordinates of alkaline phosphatase β-strand-loop-β strand motif and on the sequence alignment with link protein as presented in Fig. 4. His 372 and His 412 of E. coli alkaline phosphatase has been aligned in this model with His 18 and His 31 of link protein. The third histidine ligand, His 26, is proposed to be located in the loop. The link protein loop is shorter than the loop of the alkaline phosphatase family. In the model, the signal peptidase cleavage site is located in the loop close to one of the histidine ligands.

Second, a zinc atom, if present, could lie in the loop formed by the two antiparallel β-strands. Third, the sequence known to be cleaved by the signal peptidase lies in the upper portion within the loop. Fourth, the length of the signal peptide with the proposed conformation might be sufficient to span the membrane (30A) with the N- and C-termini facing the cytoplasm. Finally, we would like to suggest that the putative metal binding domains involving the signal peptides may have nucleic acid binding properties that could allow them to be recognized by the 7S RNA component of the signal recognition particle. The above suggestion is plausible since TFIIIA, which contains homologous zinc binding domains, is known to bind to transcripts of the 5S RNA gene (6-10). The identification of a putative metal binding domain with its possible resulting tertiary structure in the N-termini of several exported proteins described in the present communication may be useful in setting up new experimental approaches to elucidate the mechanism of protein translocation across cell membranes.

ACKNOWLEDGEMENTS

We are grateful to Dr. Van Beveren for advice with the searches, to Jim Cairns for assistance in the modelling studies and to Cindy Clatterbuck for manuscript preparation. This work was supported by grant HD22016 from the National Institute of Child Health and Human Development. P.A.T. is a recipient of an Arthritis Investigator Award.

REFERENCES

1. Walter, P. and Lingappa, V.R. (1986). Ann. Rev. Cell. Biol. 2, 499.
2. Wickner, W.T. and Lodish, H.F. (1985) Science 400-407.
3. Walter, P. and Blöbel, G. (1982). Nature 299, 691.
4. Wiedmann, M., Kurzchalia, T.V., Hartmann, E. and Rapoport, T.A. (1987) Nature 328:830-833.
5. Deák, F. et al. (1986). Proc. Natl. Acad. Sci. USA 83, 3766.
6. Diakum, G.D., Fairall, L. and Klug, A. (1986). Nature 324, 698.
7. Miller, J., McLachland, A.D. and Klug, A. (1985). EMBO J. 4, 1609.
8. Brown, R.S., Sander, C. and Argos, P. (1985). FEBS Lett. 186, 271.
9. Fairall, L., Rhodes, D. and Klug, A. (1986). J. Mol. Biol. 192, 577.
10. Hanas, J.S., Hazuda, D.J., Bogenhagen, D.F., Wu, F.Y.-H. and Wu,. C.-W. (1983). J. Biol. Chem. 258, 1420.
11. Argos, P., Garavito, R.M., Erentohoff, W., Rossmann, M.G., and Bränden. (1978). J. Mol. Biol. 126, 444.

12. Tainer, J.A., Getzaff, E.D., Beern, K.M., Richardson, J.S. and Richardson, D.C. (1982) J. Mol. Biol. 160, 181.
13. Sowadski, J.M., Handschumacher, M.D., Kristina-Murthy, H.M., Foster, B.A. and Wyckoff, H.W. (1985). J. Mol. Biol. 186, 417.
14. Hopp, T.P. and Woods, K.R. (1981). Proc. Natl. Acad. Sci. USA 78, 3824.
15. Garnier, J., Osguthorpoe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97.
16. Bradshaw, R.A. et al. (1981). Proc. Natl. Acad. Sci. USA 78, 3473.
17. Weiss, M.J., et al. (1986). Proc. Natl. Acad. Sci. USA 83, 7182.
18. Millán, J.L. (1986). J. Biol. Chem. 261, 3112.
19. Lipscomb, W.N. (1983) Ann. Rev. Biochem. 52:17-34.