Protein Seq Data Anal (1991) 4:105–110

Protein Sequences & Data Analysis

Analysis of inverted repeats in primary structure of proteins

A.S. Kolaskar and S.L. Samuel

Bioinformatics, Distributed Information Centre, Zoology Department, University of Poona, Pune 411 007, India

Abstract. A computer program has been developed to locate exact inverted repeating subsequences present anywhere in the given primary structure of proteins or nucleic acids. The output is amenable to protein sequence/nucleic acid query (PSQ/NAQ) packages. Our analysis has shown that there is a large number of proteins which have inverted repeats of more than four amino acid residues in length. However, the number is small when conditions such as the existence of more than 20 inverted repeats in given sequence or the existence of inverted repeats having more than five different types of amino acids are applied.

Introduction

Analysis of protein and nucleic acid sequences provides valuable information and many scientists now make use of computerized protein and nucleic acids sequence databanks [1–4]. Most of the sequence analysis studies can be characterized as pattern search studies. In one category of studies, the occurrence of a given oligopeptide/oligonucleotide is searched for throughout the databank. These oligopeptides/oligonucleotides are frequently markers and are associated with certain biological properties of the macromolecule [5]. In another type of studies the repeated occurrence of patterns in a particular macromolecule is studied. Biological significance is attached to those patterns having an unusual frequency of occurrence [6]. In rare cases sequences are also investigated by studying the properties of a test subsequence when an exact match is not possible. This approach is used to identify signal peptides, antigenic determinants and helper T cell-binding regions [7, 8]. In yet another study homologous sequences of macromolecules are optimally aligned and this information is used to generate phylogenetic trees [9, 10].

The pattern that is commonly sought in a particular macromolecule, if it is protein, is that of repeating subsequences. Even though the presence of inverted repeats (IR) in nucleic acids is well documented, the biological function of these IR and the mechanism by which such IR are formed is not well understood. Recently it has been claimed by Ford et al. [11, 12] that novel joints, including inverted joints, are important and have a specific role in gene amplification. Although IR in DNA have been assigned some biological significance as mentioned above, IR in proteins have not, to the best of our knowledge, been studied. We have, therefore, undertaken such a study and have developed a computer program which detects exact IR in the primary structure of a given protein/nucleic acid molecule. Application of this program to the PIR protein databank has shown that IR of four or more residues in length are not uncommon in proteins. However, longer IR only occur in a few proteins. IR of intermediate length, more than six residues, and those comprising at least five different amino acids are also rare and occur only singly in a protein. We have studied the tertiary structure of some IR and the particular codons used in these repeats.

Methods

The analysis of a protein sequence was performed using the following steps:

1. The sequence of given protein is read into an array (array B) and each amino acid type is identified.

2. The array A is created in which each row has a total number of particular type of amino acid residues in the first column and the remaining columns have numbers which specify its occurrence in the primary structure of the given protein.

3. The first residue of the reference subsequence and the last residue of its IR must be identical. Therefore, consecutive two positions i and j, where i>j, in the same row are considered. The first position, i^{th}, is treated as belonging to reference subsequence and the other, j^{th}, to its IR.

4. The residue type for the (i+1)th position is picked from array B and the occurrence of the second residue in potential IR subsequence is confirmed by checking the occurrence of (j−1) in the same row in which (i+1) occurs. This process is continued.

Offprint requests to: A.S. Kolaskar

Table 1. Proteins having more than 20 inverted repeats (IR)

Protein name/ accession no.[a]	Total no. of IR	Details about longest IR		
		Length	Fragment	Start positions: Reference, mirror
Circumsporozoite protein (CS) precursor. *Plasmodium falciparum* (isolate wellcome) /CSP$PLAFW*	163	62	$(\text{PNAN})_{15}$PN	149, 216
Circumsporozoite protein (CS) precursor. *Plasmodium falciparum* (isolate LE5) /CSP$PLAFL*	162	46	$(\text{PNAN})_{11}$PN	131, 182
Circumsporozoite protein (CS) precursor. *Plasmodium falciparum* /A03388	160	62	$(\text{PNAN})_{15}$)PN	147, 214
Procollagen alpha 1 (I) chain precursor (human)/A02852	115	14	PGPPGPPGPPGPPG	141, 1178
Collagen alpha 1 (III) chain (bovine) /A02862	96	12	PGPAGPPGPPGP	20, 1028
Circumsporozoite protein (CS) precursor. *Plasmodium cynomolgi* (strain GOMBAK) /E26255	94	7	GGAAAAG	219, 231
Procollagen alpha 1 (I) chain, skin fragment (Chick)/A02857	90	10	GPAGPPGPPG GPPGAPGAPG	761, 956 881, 908
Collagen alpha 1 (III) chain (human) /A02861	82	12	PGPAGPPGPPGP	6, 1014
Collagen alpha 1 (I) chain, skin fragment. (bovine)/A02853	70	12	PGPAGPAGPPGP	403, 607
Gamma precursor /GDB2$WHEAT*	37	13	QPFPQQPQQPFPQ	82, 100
Circumsporozoite protein (CS) precursor. *Plasmodium cynomolgi* (strain Berok) /D26255	26	42	GDGAPAAPAGDGAPAAPAGDGAPAAPAGDGAPAAPA	105, 147
Keratin type II cytoskeletal fragment /A02952	22	21	GMGMGGGMGMGGGMGMGGGMG	318, 342
Probable nuclear antigen/A03773	603	31	GGAGAGGAGGAGAGGAGGAGAGGAGGAGAGG	243, 285[b]
Histidine-rich protein precursor /A25942	138	9	AHHAAAHHA	294, 305[b]

[a] Each fragment is a palindrome

[b] Swissprot codes are differentiated from PIR accession numbers by marking them with an asterisk

5. If IR of given minimum length (L_{min}) exist then the reference subsequence, its IR and their respective position numbers are written in the output file.

6. To avoid fragments of IR occurring in the output file the start and end positions of the reference and its IR are checked and embedded regions are avoided. Following conditions are checked for embeddedness of the regions: if i and i′ are the start positions of reference sequences of length n and m for which IR exist at positions j and j′, and if $n \geq m$ and $i < i' < i+n$ and $j < j' < j+n$, then the subsequences i′ to $(i'+m)$ and j′ to $(j'+m)$ are embedded fragments of sequences $(i+n)$ and $(j+n)$, respectively. It should be mentioned here that $m >$ 'L_{min}' set by the user to find IR in a protein.

7. Steps 3 to 6 are repeated for all elements in the array A except those in the first column.

Results and discussion

To analyze proteins and nucleic acid sequences a large number of program packages are available for various computer systems. Some of these are available commercially and others through exchange. Protein sequence query (PSQ) and nucleic acid query (NAQ) are programs developed at Georgetown University and are available in the public domain. These programs are widely used by many researchers. We have, therefore, developed our program to detect IR to be compatible with the PSQ and NAQ programs. The output format of our program is such that the NAQ and PSQ programs can be used for further analysis. Our method allows the detection of palindromic subsequences as well as IR in a given sequence and their positions. We have applied our program to all the sequences in the PIR, PIR New, DIFBASE and Swissprot databanks. Information regarding occurrence of inverted subsequences obtained from our method can be incorporated in the feature table by giving the start and end positions of a subsequence and its IR. It should be mentioned here that our program detects exact IR and no mismatches are allowed at this stage; however, this facility can be included with minor changes.

The search of the protein databank for the occurrence of IR subsequences of four or more amino acids in length has shown that several proteins in the databank have one or more IR. In 2506 proteins 7300 IR of four or more amino acid residues were found to occur. We have excluded those proteins which have more than twenty IR per protein. There are about 140 proteins which have a large number of IR per protein. To compare the occurrence of IR with that of repeating subsequences, the protein databank was searched for the existence of repeating subsequences. In 2674 proteins 8032 repeating subsequences (exact repeats) of four or more amino acid residues in length were found to occur. Proteins which have more than 20 repeating subsequences per protein were also excluded in this set. We found about 240 proteins satisfying the above-mentioned conditions of a large number of repeating subsequences per protein. It should be mentioned here that repeating palindromic subsequences are included in both sets, exact repeats and IR. However, repeating subsequences such as GGGG- (homooligomers) are included only in the set of exact repeats and not as IR. Thus, it can be seen from above-mentioned data that the occurrence of IR is almost as common as repeating subsequences. Here we give the data for three categories.

Category 1: proteins having a large number (>20) IR of six or more residues in length

There are only 14 proteins as can be seen from Table 1 in which IR occur a large number of times. A glance at this table shows that all these proteins are very large molecular weight proteins and are structural proteins. In Table 1 we have given the longest oligopeptide for which an IR exists in the primary structure of the protein. The length of the longest IR and the number of times IR occur in these proteins are much larger than those that can be expected to occur from a random sequence of same length and composition, indicating a certain biological advantage for such patterns. At present we do not have any concrete proposal to explain this observation. However, we feel that the ancestor gene in these proteins was not only duplicated but was also read in the reverse direction giving it much higher amplification. Thus, each chain of the present day collagen, or circumsporozoite protein are the result of ancestor gene duplication as well as ligation of transcription product formed by reverse reading of the gene, and, at the next step, duplication of such a complex. Only such a mechanism can help one understand the occurrence of so many IR. Although we do not have direct proof for such a mechanism, in transposons there are several regions, particularly 5′- and 3′-regions, which are mirror images of each other and also have palindrome sequences. It can be seen from Table 1 that, in the probable nuclear antigen or histidine-rich protein precursor, the reference subsequence is itself a palindrome and, therefore, its IR can also be considered as its repeating subsequence. We have included these observations in our table only to show that such cases are difficult to distinguish as IR or repeating subsequences and will depend mainly on the mechanism through which they have arisen.

Category 2: proteins having IR in which each repeat consists of at least five different amino acids

We observed that there are IR in which more than one or two types of amino acids occur. In Table 2 we have given only those cases in which at least five different types of amino acids exist in the reference subsequence and in its IR. It can also be seen from the data in Table 2 that such IR occur in various types of proteins and enzymes. In most cases, there are at least six amino acids residues in the IR and they are not necessarily restricted to the N-terminal, middle or C-terminal region, but can occur anywhere in the protein. Furthermore, the distance between the reference subsequence and its IR also varies considerably and does not have a fixed pattern. The

Table 2. Proteins with IR in which each repeat consists of at least five different amino acids

Protein name/ accession no.[a]	Fragment	Start positions: Reference, mirror
Beta-lactamase (EC 3.5.2.6) precursor (cephalosporinase)/A24869	ALAVKS	180, 309
Arginase (EC 3.5.3.1) (gene name: ARG1)/S02132	IGLRDV	177, 203
Potassium-transporting ATPase (EC 3.6.1.36)/A29576	IAMGIT	14, 488
ATP synthase beta chain, mitochondrial (EC 3.6.1.34)/A01029	ESGVIN	202, 361
ATP synthase beta chain, Mitochondrial precursor (EC 3.6.1.3)/JS0002	ESGVIN	262, 421
Kit proto-oncogene tyrosine kinase precursor (EC 2.7.1.112)/S00474	PETSHLL	218, 705
Muscarinic acetylcholine receptor M2/A27386	NILVMV	41, 405
Muscarinic acetylcholine receptor M2 (cardiac)/A25656	NILVMV	41, 405
Apolipoprotein B-100 precursor/A25266	SYLQGT	1525, 1924
Apolipoprotein E precursor/A28189	LRDRAQ	213, 234
Collagen alpha 2(I) chain, skin (fragment)/A02866	GRVGAPG	142, 346
Chromogranin A precursor/S00291	GAKERA	86, 253
Outer cell wall protein precursor/B25039	VVTFDK	148, 305
Insulin receptor (fragment)/A26378	LQHREK	1014, 1022
Fibronectin receptor alpha subunit precursor/A27079	VELQLD	544, 654
NIF-specific regulatory protein (gene name: NIFA) /S01066	ERAPPG	41, 239
Regulatory protein NIFB (gene name: NIFB)/A24495	PEIGAKI	175, 444
Cartilage-specific proteoglycan core protein precursor/A28452	DSSGEP	830, 1726
Structural polyprotein/A27871	DGKCTV	597, 1186
Photosystem I P700 chlorophyll A apoprotein A1 /S01604	TAIGGL	159, 732
Spike precursor proteins /SPIK$IBV6*	KNFSAA	73, 798
Trans-acting activator of HO endonuclease gene /SWI6$YEAST*	KQLKDE	610, 677
Urotensin I precursor/A25966	RNLGAQ	63, 132
Genome polyprotein M/A04210	GSNPGI	817, 958
Virb10 protein/A26217	ASPSTL	351, 550
Virb10 protein/S00786	ASPSTL	92, 292
Probable membrane antigen P140/A03740	PETVLR	969, 1029
Hypothetical protein C-403/A04486	EYGKEN	114, 171

[a] Swissprot codes are differentiated from PIR accession numbers by marking them with an asterisk

Table 3. IR of four or more residues in length in proteins from the PDB data bank, their particular codon and main chain dihedral angles

PDB code	IR no.	AA details			Codons used		Dihedral angles			
		Name	Positions		Reference	Mirror	Reference		Mirror	
			Reference	Mirror						
2apr	1	VAL	21	199	GUC	GUC	-135	137	-100	126
		THR	22	198	ACU	ACU	-101	127	-113	130
		ILE	23	197	AUU	AUC	-124	136	-136	168
		GLY	24	196	GGU	GGU	95	178	-87	138
2pfk	1	ARG	25	266	CGT	CGT	-66	-39	-62	-49
		SER	26	265	TCT	TCC	-66	-54	-68	-50
		ALA	27	264	GCG	GCT	-49	-46	-63	-36
		LEU	28	263	CTG	CTG	-66	-45	-55	-51
3wrp	1	VAL	58	103	GTC	GTG	-64	-42	-80	-54
		GLU	59	102	GAA	GAG	-59	-44	-59	-54
		GLU	60	101	GAG	GAA	-76	-32	-61	-43
		LEU	61	100	CTG	CTG	-70	-34	-60	-37
2cpp	1	GLU	94	156	GAA	GAA	-69	-43	-80	-59
		ALA	95	155	GCC	GCC	-70	-26	-59	-33
		TYR	96	154	TAC	TAC	-65	119	-141	-53
		ASP	97	153	GAC	GAC	-137	21	-81	-23
2sec	1	PRO	5	168	CCT	CCT	-61	148	-87	1
		TYR	6	167	TAC	TAT	-58	-29	111	141
		GLY	7	166	GGC	GGC	-60	-32	-100	-167
		ILE	8	165	ATT	ATC	-60	-59	-77	135
	2	VAL	26	95	GTA	GTA	-91	135	-123	-9
		LYS	27	94	AAA	AAA	-100	117	-83	112
		VAL	28	93	GTA	GTT	-115	128	-113	105
		ALA	29	92	GCC	GCG	-107	127	-103	117
	3	ALA	151	232	GCG	GCT	-145	151	-66	-36
		ALA	152	231	GCT	GCA	-68	137	-57	-41
		ALA	153	230	GCT	GCA	-69	-35	-65	-43
		GLY	154	229	GGG	GGA	130	160	-68	-35
4atc	1	ASN	21	305	AAT	AAT	-51	-43	-104	158
		LEU	22	304	CTG	CTG	-52	-61	-87	-39
		VAL	23	303	GTG	GTT	-57	-57	-57	-32
		LEU	24	302	CTG	CTG	-61	-18	-53	-62
		ALA	25	301	GCG	GCA	-87	-43	-54	-69
2sbt	1	PRO	5	168	CCT	CCT	-55	121	-56	112
		TYR	6	167	TAC	TAC	-60	-12	-120	-130
		GLY	7	166	GGC	GGC	-29	-68	-78	179
		VAL	8	165	GTA	CTG	-76	-66	-137	141
	2	VAL	26	95	GTT	GTT	-83	142	-110	17
		LYS	27	94	AAA	AAA	-101	136	-113	108
		VAL	28	93	GTA	GTA	-119	127	-100	125
		ALA	29	92	GCG	GCT	-91	107	-100	106
	3	ALA	151	232	GCG	GCT	-135	147	-49	-50
		ALA	152	231	GCA	GCT	-49	-151	-56	-58
		ALA	153	230	GCC	GCG	113	138	-60	-32
		GLY	154	229	GGT	GGA	126	118	-83	-58
	4	SER	191	204	ACG	TCT	-44	139	56	60
		VAL	192	203	GTA	GTA	-126	168	172	111
		GLY	193	202	GGA	GGC	147	148	154	91
		PRO	194	201	CCT	CCT	-40	-23	-55	173

2apr Acid proteinase (rhizopuspepsin) (EC 3.4.23.6) Bread mold (*Rhizopus chinensis*); 2pfk Phosphofructokinase (EC 2.7.1.11) (*Escherichia coli*); 3wrp TRP aporepressor, (*Escherichia coli*); 2cpp Cytochrome p450cam (camphor monooxygenase) (EC 1.14.15.1) with bound camphor (*Pseudomonas putida*); 2sec Subtilisin carlsberg (EC 3.4.21.14) complex with genetically engineered *N*-acetyl eglin-C (*Bacillus subtilis*) and leech (*hirudo medicinalis*); 4atc Aspartate carbamoyltransferase (aspartate transcarbamylase) (EC 2.1.3.2) (*Escherichia coli*); 2sbt Subtilisin novo (EC 3.4.21.14) probably *bacillus amyloliquefaciens conclusions* pattern of amino acids in the reference subsequence is neither unique in terms of sequence nor properties such as hydrophobicity, bulkiness of side chain, secondary structure preference, etc. These regions are also not known to be part of the active sites in enzymes of functional groups in non-enzymatic proteins.

Category 3: IR of four or more residues in proteins from the PDB databank, their codon usage and main chain dihedral angles

To get an insight in the occurrence of IR we decided to study conformation of a reference subsequence and its IR. For this purpose we developed a software which extracts a protein sequence from PDB files and gives the output in such a format that PSQ can be used. Thus, for all the proteins existing in the PDB we have checked for the existence of IR with four or more amino acid residues. Once we had obtained the results for the IR, the ϕ, ψ values were calculated for the reference subsequence and its IR. These are given in Table 3. We have also extracted from the EMBL nucleic acids databank the codons corresponding to the reference subsequence and its IR (see Table 3). It can be seen from Table 3 that ϕ, ψ values for the reference as well as its IR are similar for residues occurring in the protein 2apr, 2pfk, 3wrp, 2cpp and some peptides in 2scc, with the exception of one residue in 4act. These observations suggest that the conformation of residues in the reference and in its IR does not depend on which residue is found in the immediate neighborhood on either side. We are aware that the protein sequences is given from the N to C terminus and, thus, there is an effect due to the directionality. However, in this region, this effect does not seems to be pronounced.

The mechanisms of the occurrence of IR seems to be very complex. This is mainly because occurrence of IR at the DNA level does not necessarily imply the occurrence of IR at the protein level. Thus, the suggestions of Ford et al. [11, 12] of novel inverted joints does not help to explain our observation at the protein level. We do not have any specific model to explain these observations. However, the use of particular codons in Table 3 points out that same codons are used in reference as well as in IR. This observation requires further studies.

Conclusions

In summary the results we have presented point out that there are many ways of looking at the patterns existing in protein molecules. Some of these patterns, such as those presented here, seem to have a biological significance and more experimental studies are necessary to understand this significance. However, analysis of sequence and structural databanks enables patterns in proteins and their probable structural and functional role to be delineated.

Acknowledgement. This work is supported financially by a grant from the Department of Biotechnology, Government of India.

References

1. Doolittle RF (1990) (ed) Molecular evolution, computational analysis of proteins and nucleic acids sequences. Methods of enzymology. Academic Press, New York
2. Lesk MA (1988) (ed) Computational molecular biology: sources and methods for sequence analysis. Oxford University Press, Oxford
3. Claverie JM (1986) In: Wakeford R (ed) Biotechnology information. IRL Press, Oxford Washington DC, pp 49–58
4. Keil B (1987) Protein Seq Data Anal 1:13–20
5. Michaelis S, Beckwith J (1982) Ann Rev Microbiol 36:435–465
6. Shin MS, Bargiello TA, Clark BT, Jackson FR, Young MW (1985) Nature 317:445–448
7. Kutubuddin M, Kolaskar AS, Galande S, Gore MM, Ghosh SN, Banerjee K (1991) J Mol Immunol 28:149–154
8. Parker JMR, Guo D, Hodges RS (1986) Biochemistry 25:5425–5432
9. Dayhoff MO, Barker WC, Hunt LC (1983) Methods Enzymol 91:524–545
10. Moore GW, Barnabas J, Goodman M (1973) J Theor Biol 38:459–485
11. Ford M, Davies B, Griffiths M, Wilson J, Fried M (1985) Proc Natl Acad Sci USA 82:3370–3374
12. Ford M, Fried M (1986) Cell 45:425–430